Universidad Nacional de Colombia

Facultad de Ciencias

Estadística Bayesiana

Parcial 2

Valeria Hernández González
Nicolás Alejandro Díaz Ubaque

Noviembre 2023

1. Introducción

La prueba Saber 11 es una evaluación estandarizada realizada por el Icfes que pretende recopilar datos y proporcionar información sobre las competencias básicas que debe desarrollar un estudiante tras culminar su educación media.

El objeto de este caso es ajustar modelos multinivel bayesianos tomando los datos del puntaje global de los estudiantes de Colombia y modelar los resultados a nivel nacional por municipio y departamento, y determinar la capacidad predictiva de los modelos que se mostraran a continuación y determinar cuál es el mejor en función de su capacidad predictiva.

2. Preguntas

1. En un gráfico con dos paneles (1 × 2), hacer un mapa de Colombia por departamentos, donde se desplieguen los valores de la media muestral del puntaje global (panel 1, izquierda) y la incidencia de la pobreza monetaria en 2018 (panel 2, derecha). Interpretar los resultados obtenidos (máximo 100 palabras)

   Solución

   El siguiente mapa permite visualizar la variación en los puntajes globales de la prueba Saber 11 durante el segundo semestre de 2022 en los diferentes departamentos, así como la incidencia de la pobreza monetaria en 23 departamentos y Bogotá para el año 2018.

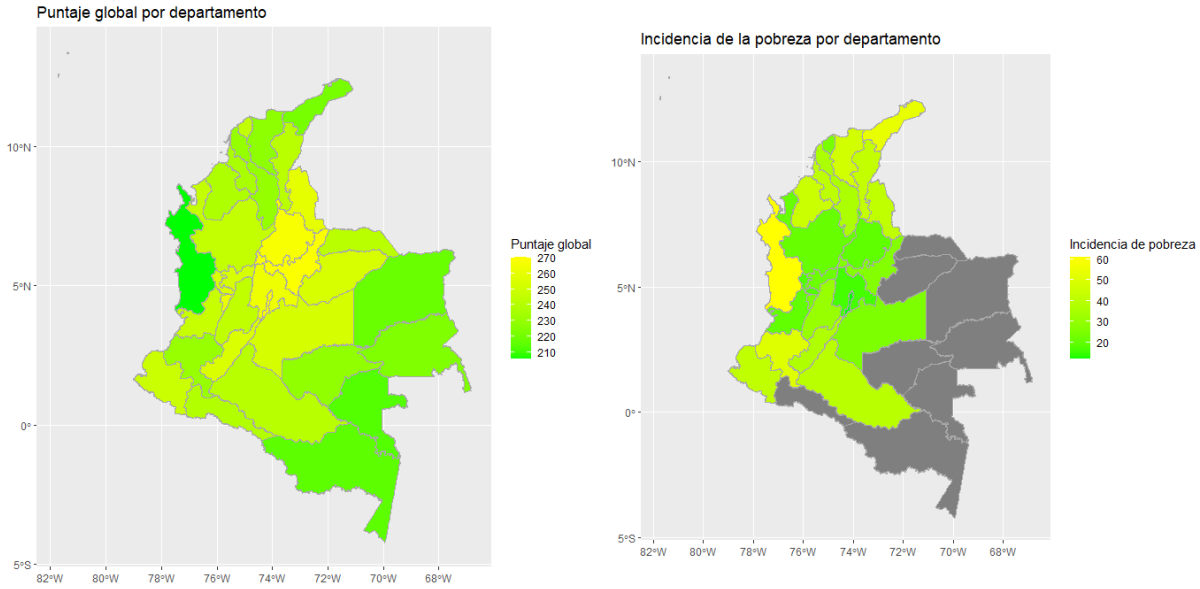


Figura 1: Puntaje global e incidencia de pobreza por departamento.

   Se puede observar que a un nivel departamental, las zonas del país que mayor incidencia de pobreza tuvieron en 2018, son también aquellas que tienen los puntajes más bajos en la prueba Saber 11 para 2022. Además, para las zonas que no fueron medidas en el estudio de pobreza monetaria, se pueden evidenciar puntajes promedio bastante bajos, por lo que gracias a la asociación ya descrita, podriamos hipotetizar que tendrán una alta incidencia de pobreza.

2. En un gráfico con dos paneles (1×2), hacer un mapa de Colombia por municipios, donde se desplieguen los valores de la media muestral del puntaje global (panel 1, izquierda) y la cobertura neta secundaria en 2022 (panel 2, derecha). Interpretar los resultados obtenidos (máximo 100 palabras).

   Solución

El siguiente mapa permite visualizar la variación los valores de la media muestral del puntaje global de la prueba Saber 11 en los diferentes municipios, así como la cobertura neta secundaria en 2022, entendiendo la cobertura neta secundaria como el "porcentaje de estudiantes matriculados en el sistema educativo; sin contar los que están en extra edad (por encima de la edad correspondiente para cada grado" (MEN).

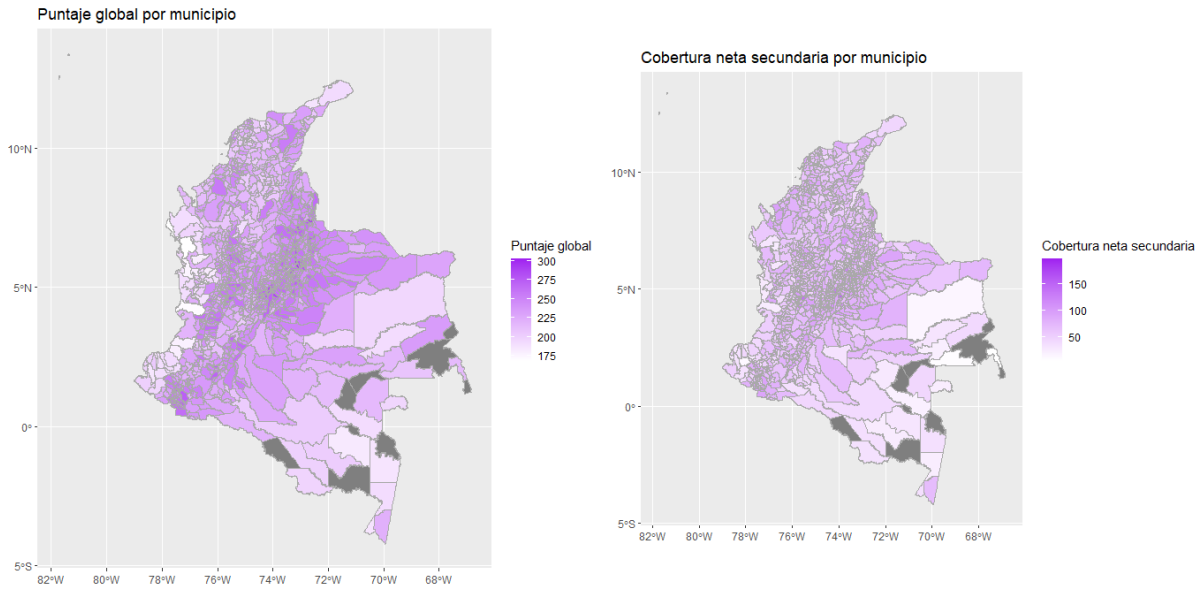


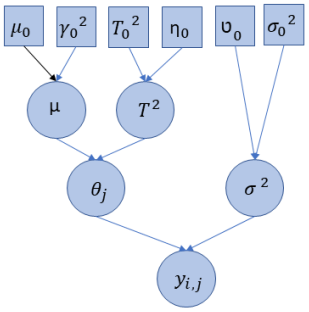
Figura 2: Puntaje global y cobertura neta secundaria por municipio.

Los municipios donde se reportó menor cobertura neta de secundaria corresponden con los municipios que tuvieron un menor puntaje global en las pruebas del ICFES del 2022. Las deficiencias en los puntajes globales se concentran en las regiones Andina y Caribe, así como en zonas remotas como Puerto Arica, Puerto Alegria (Amazonas), Guarurú (Vaupes) y Puerto Colombia (Guainia). Donde se puede hipotetizar en las regiones del Amazonas y Orinoquia, la educación podría verse afectada por su ubicación rural y alejada, limitando el acceso a recursos y docentes cualificados.

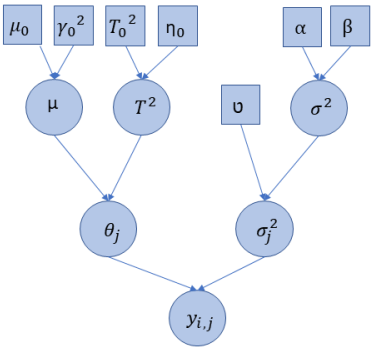
3. En un gráfico con cuatro paneles (2 × 2), hacer el DAG de M2 (panel 1, esquina superior izquierda), M3 (panel 2, esquina superior derecha), M4 (panel 3, esquina inferior izquierda) y M5 (panel 1, esquina inferior derecha).

Solución

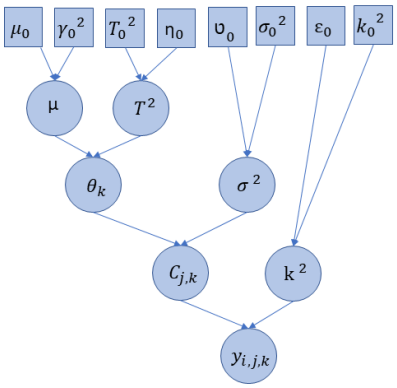
A continuación se presenta el DAG para los 4 modelos que se implementaron a lo largo del presente trabajo.



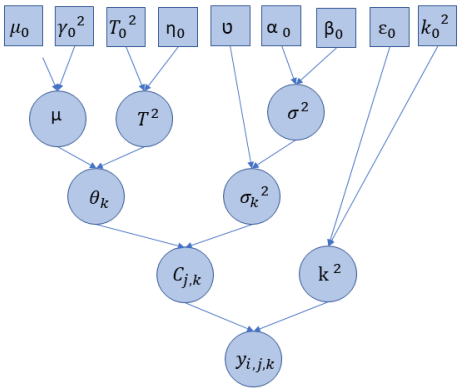
(a) DAG Modelo 2



(b) DAG Modelo 3



(c) DAG Modelo 4



(d) DAG Modelo 5

Figura 3: DAG modelos M2, M3, M4 y M5

4. En un gráfico con cuatro paneles (2 × 2), dibujar la cadena de la log-verosimilitud de M2 (panel 1, esquina superior izquierda), M3 (panel 2, esquina superior derecha), M4 (panel 3, esquina inferior izquierda) y M5 (panel 1, esquina inferior derecha). Los gráficos deben tener la misma escala para facilitar la comparación. Interpretar los resultados obtenidos (máximo 100 palabras).

Solución

Se ajustan cuatro modelos utilizando muestreadores de Gibbs con 101,000 iteraciones. Las primeras 1,000 iteraciones se consideran un periodo de calentamiento del algoritmo. Luego, se realiza un muestreo sistemático de amplitud 10. En el apéndice, se encuentran las distribuciones condicionales completas para cada modelo, así como un resumen de los coeficientes de variación de Monte Carlo de cada parámetro de cada modelo. A continuación se presenta la cadena de la log-verosimilitud para cada uno de los modelos.

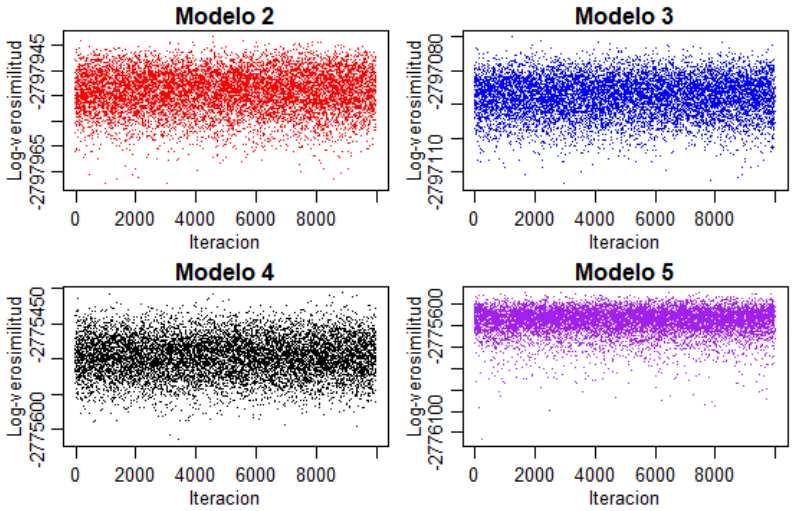


Figura 4: Cadenas de log-verosimilitud para los 4 modelos.

Debido a las grandes diferencias en los valores de la log-verosimilitud, no es posible comparar gráficamente los 4 modelos, pero podemos observar que el modelo 4 es el que alcanza mayores valores en la misma, lo que se puede deber a que el modelo 5 pudo tener problemas producto de la varianza para el departamento de Bogotá, cuestión que no representa un problema tan grande para el modelo sin varianzas específicas. Además, los modelos 4 y 5, presentan convergencia hacia un valor mayor que los modelos 2 y 3.

5. Calcular el DIC y el WAIC de cada Mk, para k = 1, . . . , 5. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras)

   Solución

   A continuación, se presentan los criterios de información DIC y WAIC para los 5 modelos.

Modelo	DIC	WAIC
1	5636785	15810021
2	5595932	5595930
3	5594240	5594236
4	5552007	5551797
5	5552225	5552005

Tabla 1: Criterios de información

   Se puede evidenciar una gran mejora en las medidas DIC y WAIC para los modelos 4 y 5, respecto a los primeros 3 modelos, pues estos tienen información específica por municipios, lo cual genera una mejor caracterización de los puntajes globales. Además, el modelo que mejor citerios de información presenta es el modelo 4, probablemente producto de la incertidumbre que introduce la varianza del departamento Bogotá al modelo 5.

6. Calcular la media posterior y el intervalo de credibilidad al 95 % basado en percentiles de μ de cada M_k, para $k = 1, ..., 5$. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

   Solución

   A continuación, se presenta el valor de la media posterior y el intervalo de credibilidad al 95 % basado en percentiles de μ de cada modelo.

Modelo	Media Posterior	Q 2.25 %	Q 97.5 %
1	250.27	250.13	250.41
2	242.29	235.59	249.18
3	244.36	239.35	249.61
4	232.82	226.71	238.98
5	232.68	226.74	238.76

Tabla 2: Media posterior e intervalos de credibilidad para μ

Podemos observar que las estimaciones para μ son parecidas para los modelos que no consideran información sobre Municipios, dado que una vez se considera esta información para los modelos 4 y 5, las estimaciones se reducen considerablemente. Adicional a esto, se evidencia que el modelo 1 presenta una media posterior y límites de credibilidad muy cercanos, lo que se debe a la naturaleza casi determinística del modelo.

7. Usando M5, hacer el ranking de los departamentos basado las medias específicas de los departamentos. Comparar los resultados con un ranking frecuentista basado en la media muestral. En un gráfico con dos paneles (1×2), hacer la visualización del ranking Bayesiano (panel 1, izquierda) y el ranking frecuentista. Las visualizaciones deben incluir simultáneamente las estimaciones puntuales y los intervalos de credibilidad/confianza al 95 %. Interpretar los resultados obtenidos (máximo 100 palabras).

Solución

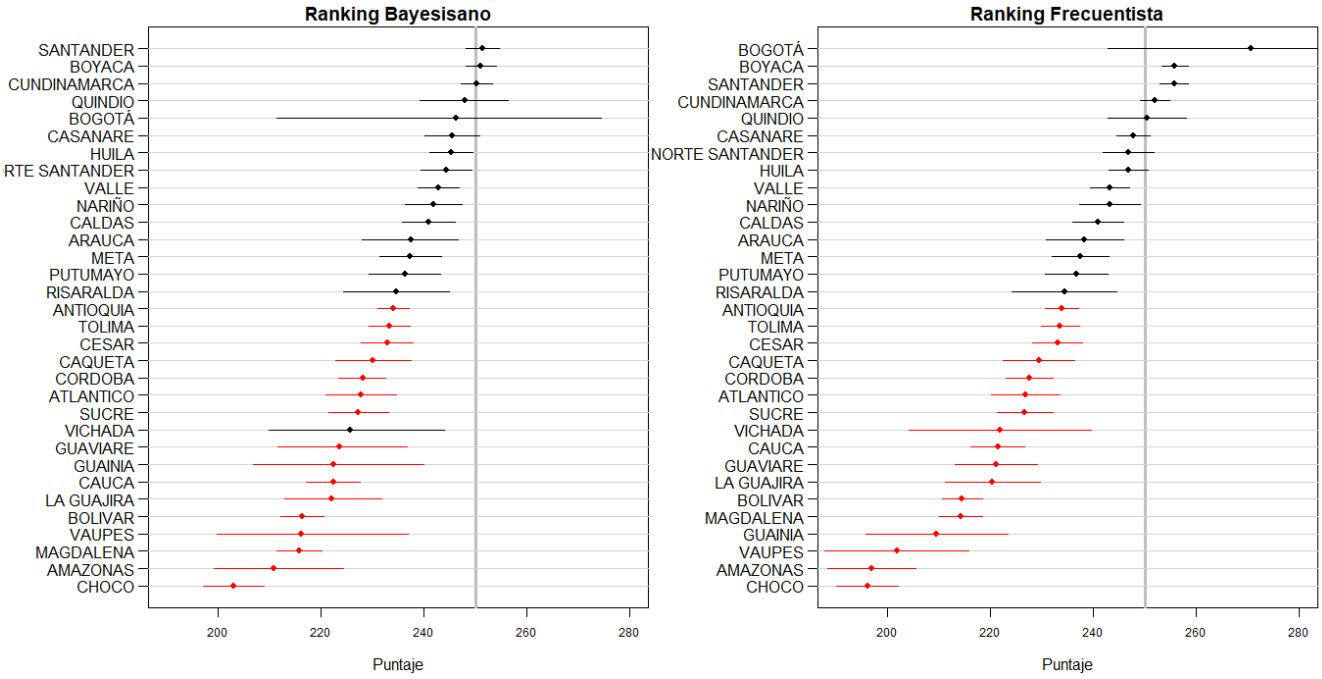


Figura 5: Ranking bayesiano y frecuentista basado en la media muestral

Se observan clasificaciones similares en el ranking bayesiano y frecuentista, sin embargo, destacan diferencias notables en los primeros 5 departamentos, con Santander liderando el ranking bayesiano y Bogotá el frecuentista. Estas diferencias sugieren que el enfoque bayesiano proporciona una evaluación más completa, al tomar en cuenta la incertidumbre específica de cada muncipio. Además, es importante notar que las medias de los puntajes globales de la mayoría de los departamentos se sitúan por debajo de los 250 puntos, en contraste al ranking bayesiano, lo cual podría deberse a una subestimación de valores por parte del modelo.

8. Usando M5, hacer una segmentación de los departamentos usando las medias específicas de los departamentos, por medio del método de agrupamiento de K-medias con cinco grupos. Presentar los resultados obtenidos

visualmente a través de una matriz de incidencia organizada a partir del ranking Bayesiano del numeral anterior y de un mapa que señale los departamentos que pertenecen al mismo grupo. Interpretar los resultados obtenidos (máximo 100 palabras).

Solución

El siguiente gráfico permite visualizar la segmentación departamental por puntaje y la matriz de incidencia departamentos.

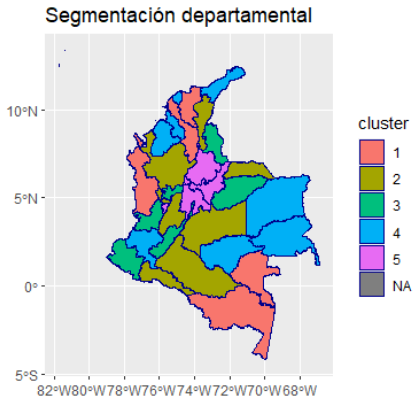


Figura 6: Segmentación departamental por puntaje.

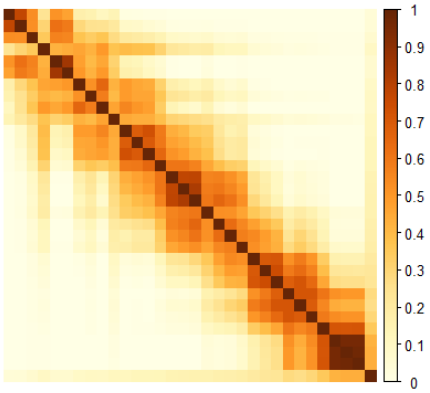


Figura 7: Matriz de incidencia departamentos.

Se puede observar en la matriz de incidencia que los departamentos que presentan medias similares, también suelen quedar en los mismos clusters iteración a iteración, como se evidencia al momento de observar la diagonal de la matriz, pues la misma está organizada por ranking de puntajes. Adicionalmente, se observa que los departamentos con los peores puntajes, estuvieron casi siempre en los mismos clusters, y el mapa de segmentación posterior permite identificar patrones regionales, donde el clúster 5 corresponde a la región media del país con puntajes altos, y el clúster 1 a regiones marginales con puntajes bajos.

9. Calcular la media posterior y un intervalo de credibilidad al 95 % de la incidencia de la pobreza monetaria en 2018 (IPM) para todos los departamentos que no fueron medidos por el DANE, por medio de una regresión lineal simple de la IPM frente a las medias específicas de los departamentos de M5. Presentar los resultados tabularmente (organizados descendente de acuerdo con la media posterior) y visualmente (por medio de un mapa usando la media posterior).

Solución

A continuación, se presenta la incidencia de pobreza con estimaciones para los departamentos no medidos y sus respectivos intervalos de credibilidad.

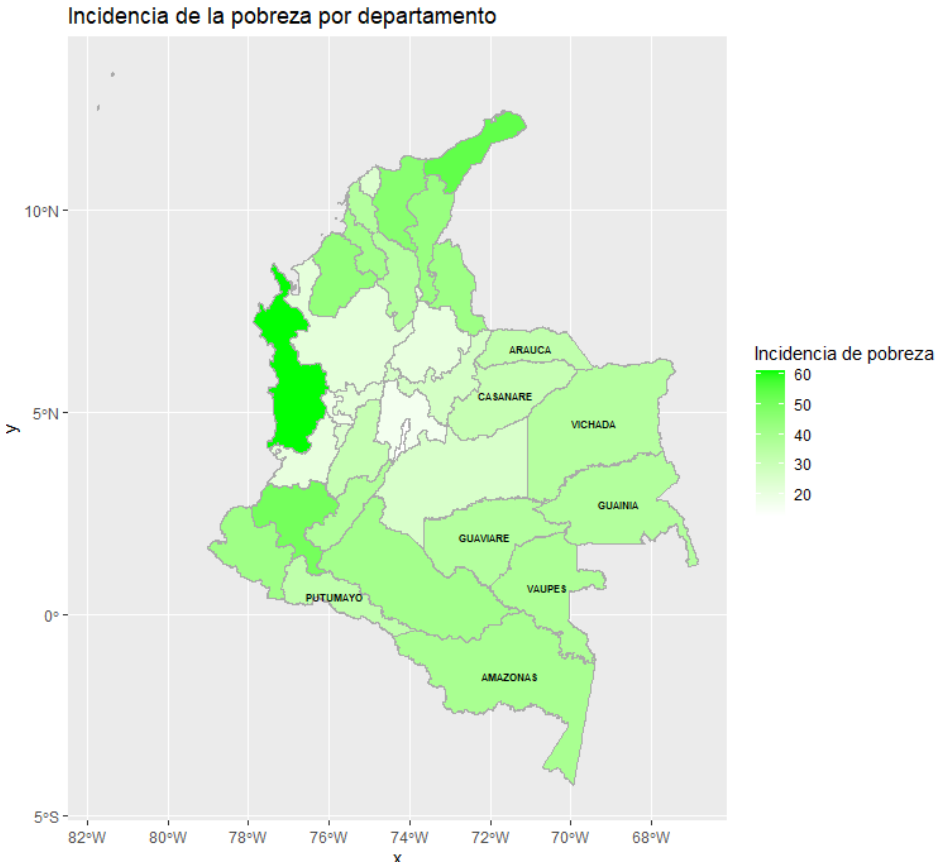


Figura 8: Incidencia de pobreza con estimaciones para los departamentos no medidos

Departamento	Incidencia de la pobreza
Arauca	32.55
Casanare	30.67
Putumayo	32.84
Amazonas	38.88
Guainía	36.17
Guaviare	35.92
Vaupés	37.70
Vichada	35.38

Tabla 3: Estimaciones puntuales de incidencia de la pobreza

Departamento	Límite inferior	Límite superior
Arauca	30.09	34.93
Casanare	28.55	32.81
Putumayo	30.97	34.68
Amazonas	33.88	43.69
Guainía	32.05	41.29
Guaviare	32.68	39.94
Vaupés	32.64	43.60
Vichada	30.90	40.30

Tabla 4: Intervalos de credibilidad incidencia de la pobreza

Las estimaciones de incidencia de pobreza para los departamentos no medidos ubicados en las regiones de Orinoquia y Amazonia, son más bajas que en la región del Caribe y Pacífico, lo cual se relaciona a que dichos departamentos tienen en promedio puntajes bajos. Las estimaciones muestran una variabilidad significativa en la incidencia de pobreza en estas áreas, lo cual resalta la necesidad de políticas adaptados a condiciones departamentales, sin embargo, esta misma incertidumbre resalta la necesidad de recopilar datos más precisos para una toma de decisiones informada.

10. Usando M5, hacer el ranking de los municipios basado las medias específicas de los municipios (no es preciso visualizar el ranking debido a la gran cantidad de municipios). Luego, hacer una segmentación de los municipios usando las medias específicas de los municipios, por medio del método de agrupamiento de K-medias con ocho grupos. Presentar los resultados obtenidos visualmente a través de una matriz de incidencia organizada a partir del ranking Bayesiano de los municipios obtenido inicialmente y de un mapa que señale los municipios que perteneces al mismo grupo. Interpretar los resultados obtenidos (máximo 100 palabras).

Solución

El siguiente gráfico permite visualizar la segmentación municipal por puntaje y la matriz de incidencia de municipios.

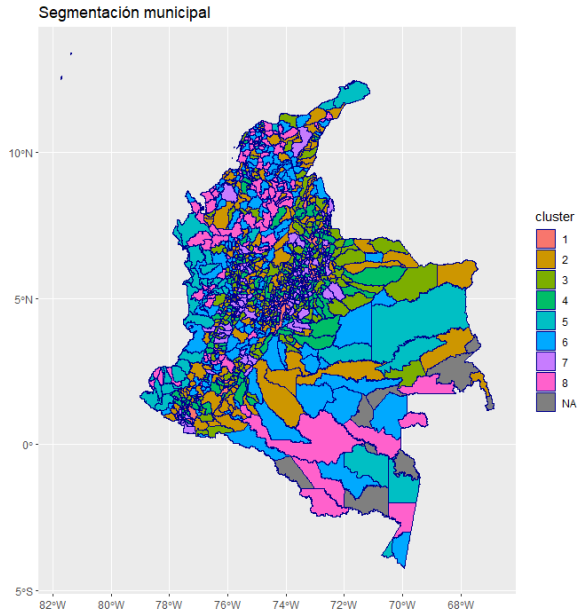


Figura 9: Segmentación municipal por puntaje.

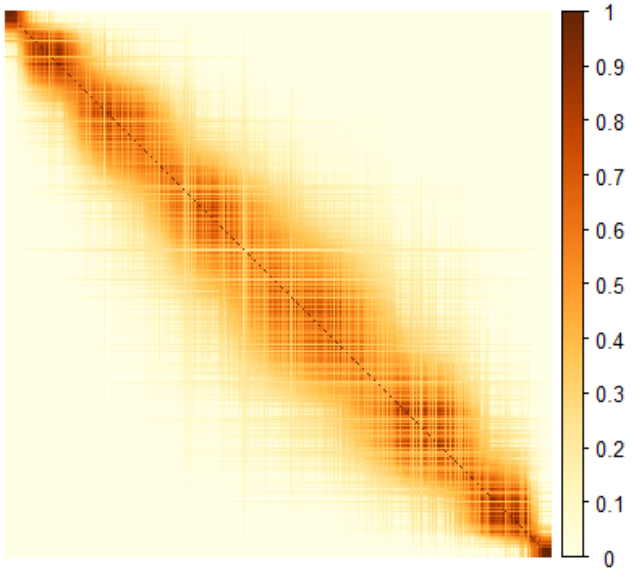


Figura 10: Matriz de incidencia Municipios.

Se observa un comportamiento similar para la matriz de incidencia respecto a la de departamentos, es decir, los municipios que suelen estar cercanos en el ranking, también suelen estar en los mismos clusters iteración a iteración. Además, los municipios con puntajes altos y cercanos y con puntajes bajos y cercanos, suelen agruparse más que los que los que se ubican en el centro del ranking. Por último, se puede observar en el mapa que la mayoria de municipios suelen agruparse en los clusters 5 y 6, pues seguramente corresponderán a puntajes cercanos al promedio general.

11. Calcular la media posterior y un intervalo de credibilidad al 95 % de la cobertura neta secundaria en 2022 (CNS) para todos los municipios que no fueron medidos por el MEN, por medio de una regresión lineal simple de la CNS frente a las medias específicas de los municipios de M5. Presentar los resultados tabularmente (organizados descendente de acuerdo con la media posterior) y visualmente (por medio de un mapa usando la media posterior).

Solución

A continuación, se presenta la cobertura neta con estimaciones para los municipios no medidos y sus respectivos intervalos de credibilidad.

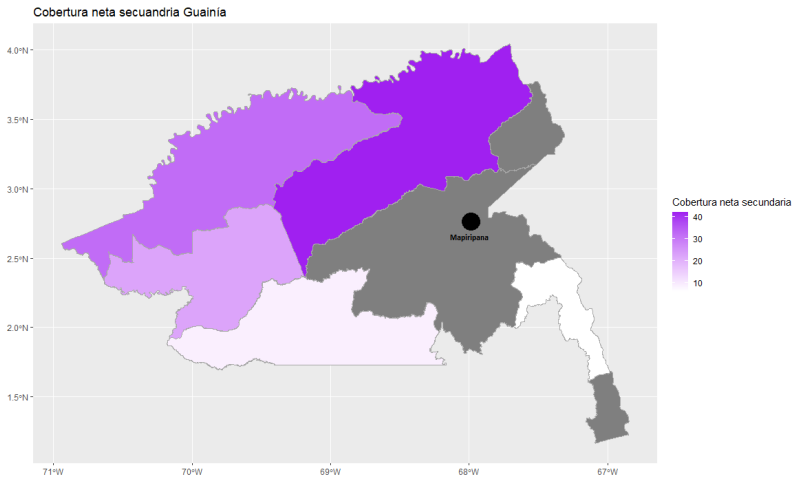


Figura 11: Cobertura neta secundaria Guainía

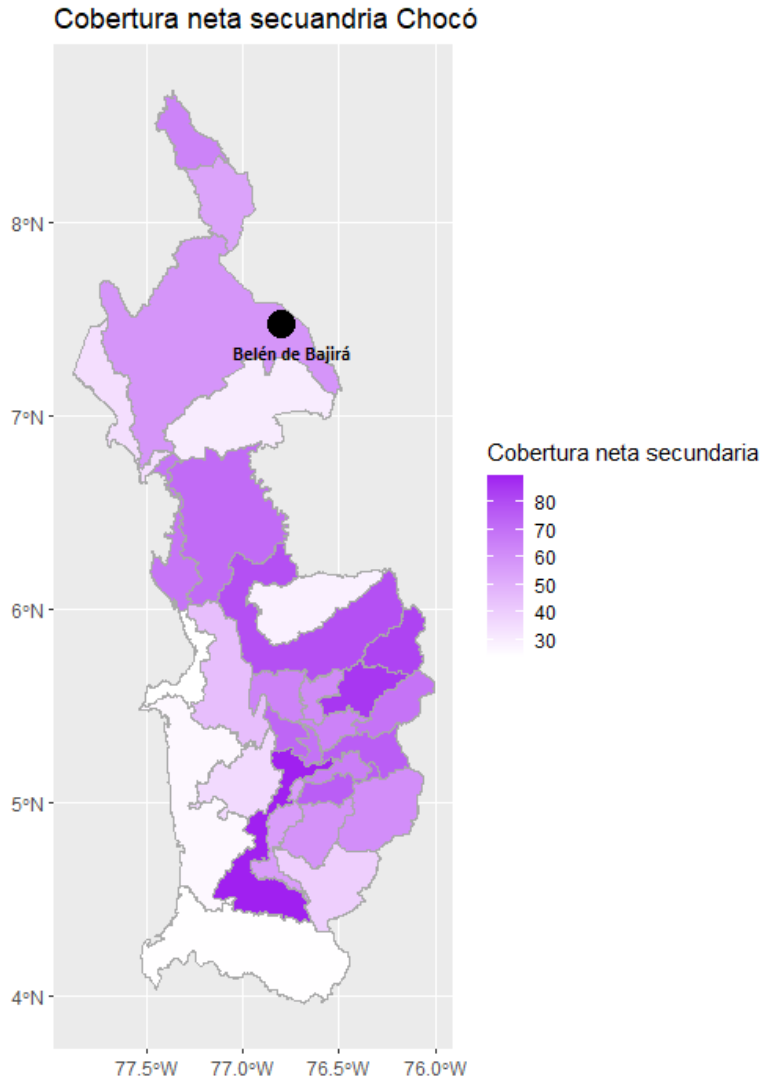


Figura 12: Cobertura neta secundaria Chocó

Municipio	Cobertura neta secundaria
Belén de Bajirá	65.14
Mapiripana	65.55

Tabla 5: Estimaciones puntuales cobertura neta secundaria

Municipio	Límite inferior	Límite superior
Belén de Bajirá	63.23	66.99
Mapiripana	61.26	69.63

Tabla 6: Intervalos de credibilidad cobertura neta secundaria

Aunque no es posible observar la comparación de los dos municipios respecto a los demás dentro del mapa debido a que no se encuentran en el registro cartográfico de 2022, las estimaciones obtenidas dejan ver índices bajos de cobertura neta secundaria, lo cual está relacionado a que se encuentran ubicados en departamentos caracterizados por puntajes bajos, siendo Chocó y Guainía respectivamente.

12. Validar la bondad ajuste de M5 por medio de la distribución predictiva posterior en cada municipio, utilizando como estadísticos de prueba el mínimo, el máximo, el rango intercuartílico, la media, la mediana, y la desviación estándar. Presentar los resultados visualmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Solución

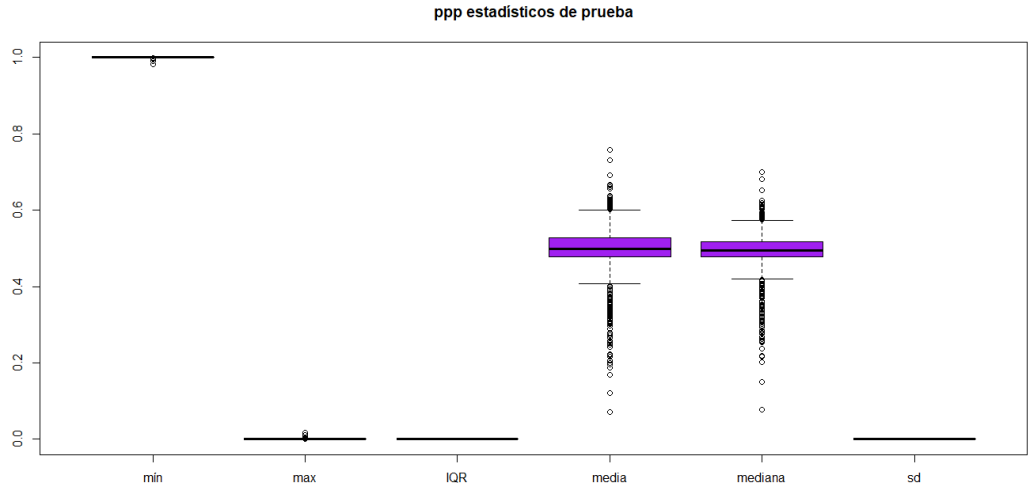


Figura 13: ppp por municipio para mínimo, máximo, media, mediana, IQR y desvaición estándar

Se evidencia que el modelo 5 no logra capturar de buena manera el mínimo, máximo e IQR, lo cual puede deberse a la alta variabilidad que introduce al modelo la caracterización por municipios. La media y la mediana se capturan de manera correcta, indicando que el modelo caracteriza bien estas estadísticas. Por último, la desviación estándar presenta una tasa de captura de 0, producto de la alta incertidumbre que introduce al modelo el departamento de Bogotá.

3. Apéndice

Resúmen coeficientes de variación de Monte Carlo

σ^2	μ	τ^2
0.048	0.034	1.00

(a) CV Monte Carlo para σ^2, μ y τ^2 Modelo 2

μ	τ^2	ν	υ^2
0.026	0.815	0.035	1.549

(b) CV Monte Carlo para μ, τ^2, ν y υ^2 Modelo 3

κ^2	σ^2	μ	τ^2
0.045	0.107	0.032	0.806

(a) CV Monte Carlo para κ^2, σ^2, μ y τ^2 Modelo 4

κ^2	μ	τ^2	σ^2
0.044	0.030	0.778	0.511

(b) CV Monte Carlo para κ^2, μ, τ^2 y σ^2 Modelo 5

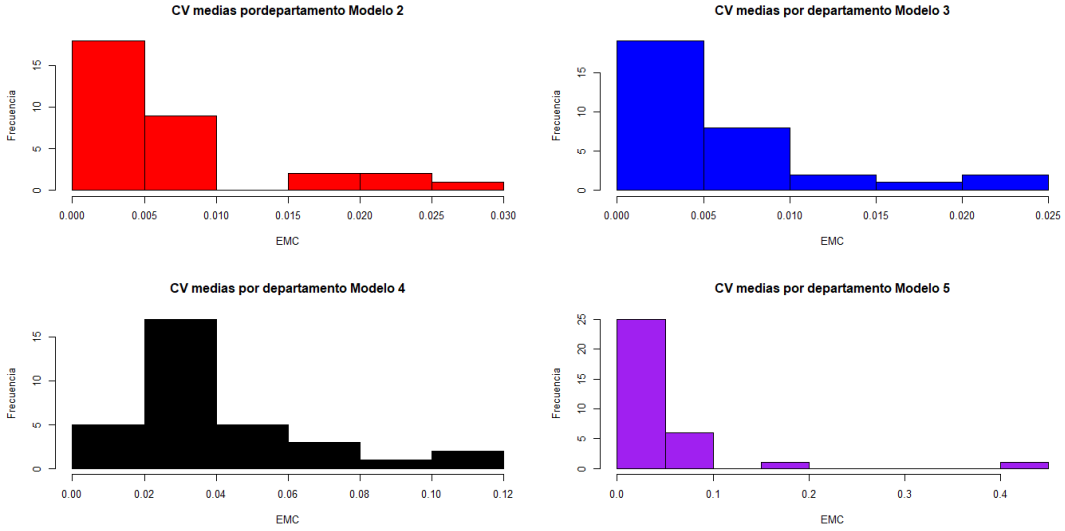


Figura 14: CV de Monte Carlo medias por departamento

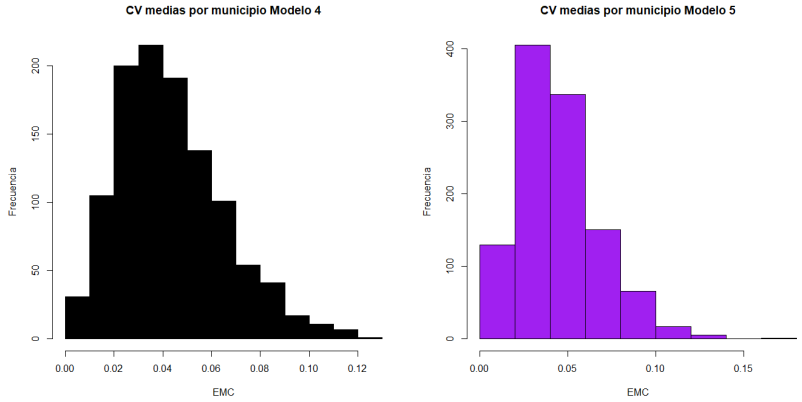


Figura 15: CV de Monte Carlo medias por municipio

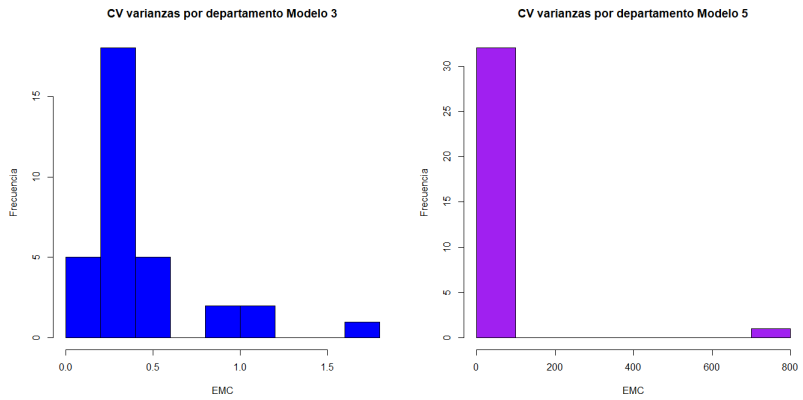


Figura 16: CV de Monte Carlo varianzas por departamento

M1: Modelo Normal
Distribución muestral:

$$y_i \mid \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

Distribución previa:

$$\theta \sim N(\mu_0, \gamma_0^2)$$
$$\sigma^2 \sim \text{GI}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

Distribución posterior:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \quad = \prod_{i=1}^{n} \text{N}\left(y_i \mid \theta, \sigma^2\right) \times \text{N}\left(\theta \mid \mu_0, \gamma_0^2\right) \times \text{GI}\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

Distribuciones condicionales completas:

$$\theta \mid \text{resto} \sim \text{N}\left(\frac{\frac{\mu_0}{\gamma_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\gamma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\gamma_0^2} + \frac{n}{\sigma^2}}\right)$$
$$\sigma^2 \mid \text{resto} \sim \text{GI}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + n s_{\theta^2}}{2}\right)$$

M2: Modelo Normal con medias específicas por departamento
Distribución muestral:

$$y_{i,j} \mid \theta_j, \sigma^2 \overset{iid}{\sim} N(\theta_j, \sigma^2)$$

Distribución previa:

$$
\begin{aligned}
\theta_j \mid \mu, \tau^2 &\overset{iid}{\sim} N(\mu, \tau^2) \\
\sigma^2 &\sim \text{GI}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \\
\mu &\sim \text{N}\left(\mu_0, \gamma_0^2\right) \\
\tau^2 &\sim \text{GI}\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right)
\end{aligned}
$$

Distribución posterior:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto & p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
= & \prod_{j=1}^{m} \prod_{i=1}^{n_j} \text{N}\left(y_{i,j} \mid \theta_j, \sigma^2\right) \times \prod_{j=1}^{m} \text{N}\left(\theta_j \mid \mu, \tau^2\right) \\
& \times \text{GI}\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \times \text{N}\left(\mu \mid \mu_0, \gamma_0^2\right) \times \text{GI}\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right)
\end{aligned}
$$

Distribuciones condicionales completas:

$$
\begin{aligned}
\theta_j \mid \text{ resto } &\sim \text{N}\left(\frac{\mu/\tau^2 + n_j\bar{y}_j/\sigma^2}{1/\tau^2 + n_j/\sigma^2}, \frac{1}{1/\tau^2 + n_j/\sigma^2}\right) \\
\sigma^2 \mid \text{ resto } &\sim \text{GI}\left(\frac{\nu_0 + \sum_{j=1}^{m} n_j}{2}, \frac{\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(y_{i,j} - \theta_j\right)^2}{2}\right) \\
\mu \mid \text{ resto } &\sim \text{N}\left(\frac{\mu_0/\gamma_0^2 + m\bar{\theta}/\tau^2}{1/\gamma_0^2 + m/\tau^2}, \frac{1}{1/\gamma_0^2 + m/\tau^2}\right) \\
\tau^2 \mid \text{ resto } &\sim \text{GI}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{j=1}^{m}\left(\theta_j - \mu\right)^2}{2}\right)
\end{aligned}
$$

M3: Modelo Normal con medias y varianzas específicas por departamento
Distribución muestral:

$$
y_{i,j} \mid \theta_j, \sigma^2 \overset{\text{iid}}{\sim} \text{N}\left(\theta_j, \sigma_j^2\right)
$$

Distribución previa:

$$
\begin{aligned}
\theta_j \mid \mu, \tau^2 &\overset{\text{iid}}{\sim} \text{N}\left(\mu, \tau^2\right) \\
\sigma_j^2 \mid \nu, \sigma^2 &\overset{\text{iid}}{\sim} \text{GI}\left(\frac{\nu}{2}, \frac{\nu\sigma^2}{2}\right) \\
\mu &\sim \text{N}\left(\mu_0, \gamma_0^2\right) \\
\tau^2 &\sim \text{GI}\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right) \\
\sigma^2 &\sim \text{Gamma}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right)
\end{aligned}
$$

Distribución posterior:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = & \prod_{j=1}^{m} \prod_{i=1}^{n_j} \mathrm{N}\left(y_{i,j} \mid \theta_j, \sigma^2\right) \\
& \times \prod_{j=1}^{m} \mathrm{N}\left(\theta_j \mid \mu, \tau^2\right) \times \prod_{j=1}^{m} \mathrm{GI}\left(\sigma_j^2 \mid \frac{\nu}{2}, \frac{\nu\sigma^2}{2}\right) \\
& \times \mathrm{N}\left(\mu \mid \mu_0, \gamma_0^2\right) \times \mathrm{GI}\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right) \\
& \times e^{-\lambda_0 \nu} \times \mathrm{G}\left(\sigma^2 \mid \alpha_0, \beta_0\right)
\end{aligned}
$$

Distribuciones condicionales completas:

$$
\begin{aligned}
\theta_j \mid \text{ resto } &\sim \mathrm{N}\left(\frac{\mu/\tau^2 + n_j \bar{y}_j/\sigma_j^2}{1/\tau^2 + n_j/\sigma_j^2}, \frac{1}{1/\tau^2 + n_j/\sigma_j^2}\right) \\
\sigma_j^2 \mid \text{ resto } &\sim \mathrm{GI}\left(\frac{\nu + n_j}{2}, \frac{\nu\sigma^2 + \sum_{i=1}^{n_j} \left(y_{i,j} - \theta_j\right)^2}{2}\right) \\
\mu \mid \text{ resto } &\sim \mathrm{N}\left(\frac{\mu_0/\gamma_0^2 + m\bar{\theta}/\tau^2}{1/\gamma_0^2 + m/\tau^2}, \frac{1}{1/\gamma_0^2 + m/\tau^2}\right) \\
\tau^2 \mid \text{ resto } &\sim \mathrm{GI}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0 \tau_0^2 + \sum_{j=1}^{m} \left(\theta_j - \mu\right)^2}{2}\right) \\
\sigma^2 \mid \text{ resto } &\sim \mathrm{G}\left(\alpha_0 + \frac{m\nu}{2}, \beta_0 + \frac{\nu}{2} \sum_{j=1}^{m} \frac{1}{\sigma_j^2}\right)
\end{aligned}
$$

M4: Modelo Normal con medias específicas por municipio y departamento)
Distribución muestral:

$$
y_{ijk} | \zeta_{j,k}, k^2 \overset{ind}{\sim} N(\zeta_{j,k}, k^2)
$$

para $i = 1, ..., n_{j,k}$, $j =, ..., n_k$ y $k = 1, ..., m$ donde $y_{i,j,k}$ es el puntaje global del estudiante i en el municipio j del departamento k.

Distribución previa:

$$
\begin{gathered}
\zeta_{j,k} | \theta_k, \sigma^2 \overset{iid}{\sim} N(\theta_k, \sigma^2), \\
\theta_k | \mu, \tau^2 \overset{ind}{\sim} N(\mu, \tau^2), \\
\mu \sim N(\mu_0, \gamma_0^2) \\
k^2 \sim GI\left(\tfrac{\xi_0}{2}, \tfrac{\xi_0 k_0^2}{2}\right) \\
\tau^2 \sim GI\left(\tfrac{\eta_0}{2}, \tfrac{\eta_0 \tau_0^2}{2}\right) \\
\sigma^2 \sim GI\left(\tfrac{\nu_0}{2}, \tfrac{\nu_0 \sigma_0^2}{2}\right)
\end{gathered}
$$

Distribución posterior:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = & \prod_{k=1}^{m} \prod_{j=1}^{n_k} \prod_{i=1}^{n_{j,k}} \mathrm{N}\left(y_{i,j,k} \mid \zeta_{j,k}, k^2\right) \\
& \times \prod_{k=1}^{m} \prod_{j=1}^{n_k} \mathrm{N}\left(\zeta_{j,k} \mid \theta_k, \sigma^2\right) \\
& \times \prod_{k=1}^{m} \mathrm{N}\left(\theta_k \mid \mu, \tau^2\right) \times \mathrm{GI}\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \\
& \times \mathrm{N}\left(\mu_{|} \mu_0, \gamma_0^2\right) \times \mathrm{GI}\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right) \\
& \times \mathrm{GI}\left(\ k^2 \mid \frac{\xi_0}{2}, \frac{\xi k_0^2}{2}\right)
\end{aligned}
$$

Distribuciones condicionales completas:

$$
\begin{aligned}
\zeta_{j,k} \mid \text{ resto } &\sim \mathrm{N}\left(\frac{n_{jk}\bar{y}_{jk} + n_k\theta_k}{n_k/k^2 + n_{jk}/\sigma^2}, \frac{1}{n_k/k^2 + n_{jk}/\sigma^2}\right) \\
\theta_j \mid \text{ resto } &\sim \mathrm{N}\left(\frac{\mu/\tau^2 + n_k\bar{\zeta}_k/\sigma^2}{1/\tau^2 + n_k/\sigma^2}, \frac{1}{1/\tau^2 + n_k/\sigma^2}\right) \\
\sigma^2 \mid \text{ resto } &\sim \mathrm{GI}\left(\frac{n + \eta_0}{2}, \frac{\nu_0\sigma_0^2 + \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left(\zeta_{j,k} - \theta_k\right)^2}{2}\right) \\
\mu \mid \text{ resto } &\sim \mathrm{N}\left(\frac{\mu_0/\gamma_0^2 + m\bar{\theta}/\tau^2}{1/\gamma_0^2 + m/\tau^2}, \frac{1}{1/\gamma_0^2 + m/\tau^2}\right) \\
\tau^2 \mid \text{ resto } &\sim \mathrm{GI}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{k=1}^{m}\left(\theta_k - \mu\right)^2}{2}\right) \\
k^2 \mid \text{ resto } &\sim \mathrm{GI}\left(\frac{\sum_{k=1}^{m}\sum_{j=1}^{n_k} n_{j,k} + \xi_0}{2}, \frac{\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{j,k}}\left(y_{i,j,k} - \zeta_{j,k}\right)^2 + \xi_0 k_0^2}{2}\right)
\end{aligned}
$$

M5:Modelo Normal con medias específicas por municipio y departamento
Distribución muestral:

$$
y_{ijk} | \zeta_{j,k}, k^2 \overset{ind}{\sim} N(\zeta_{j,k}, k^2)
$$

para $i = 1, ..., n_{j,k}$, $j =, ..., n_k$ y $k = 1, ..., m$ donde $y_{i,j,k}$ es el puntaje global del estudiante i en el municipio j del departamento k.
Distribución previa:

$$
\begin{gathered}
\zeta_{j,k} | \theta_k, \sigma^2 \overset{iid}{\sim} N(\theta_k, \sigma_k^2), \\
\theta_k | \mu, \tau^2 \overset{ind}{\sim} N(\mu, \tau^2), \\
\sigma_k^2 | \nu, \sigma^2 \sim GI(\frac{\nu}{2}, \frac{\nu\sigma^2}{2}) \\
\mu \sim N(\mu_0, \gamma_0^2) \\
k^2 \sim GI\left(\tfrac{\xi_0}{2}, \tfrac{\xi_0 k_0^2}{2}\right) \\
\tau^2 \sim GI\left(\tfrac{\eta_0}{2}, \tfrac{\eta_0\tau_0^2}{2}\right) \\
\sigma^2 \sim G\left(\tfrac{\alpha_0}{2}, \tfrac{\beta_0}{2}\right)
\end{gathered}
$$

Distribución posterior:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = & \prod_{k=1}^{m} \prod_{j=1}^{n_k} \prod_{i=1}^{n_{j,k}} \mathrm{N}\left(y_{i,j,k} \mid \zeta_{j,k}, k^2\right) \\
& \times \prod_{k=1}^{m} \prod_{j=1}^{n_k} \mathrm{N}\left(\zeta_{j,k} \mid \theta_k, \sigma_k^2\right) \\
& \times \prod_{k=1}^{m} \mathrm{N}\left(\theta_k \mid \mu, \tau^2\right) \times \prod_{k=1}^{m} \mathrm{N}\left(\sigma_k^2 \mid \frac{\nu}{2}, \frac{nu\sigma^2}{2}\right) \\
& \times \mathrm{G}\left(\sigma^2 \mid \frac{\alpha_0}{2}, \frac{\beta_0}{2}\right) \\
& \times \mathrm{N}\left(\mu_{\mid} \mu_0, \gamma_0^2\right) \times \mathrm{GI}\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right) \\
& \times \mathrm{GI}\left(k^2 \mid \frac{\xi_0}{2}, \frac{\xi k_0^2}{2}\right)
\end{aligned}
$$

Distribuciones condicionales completas:

$$
\begin{aligned}
\zeta_{j,k} \mid \text{ resto } & \sim \mathrm{N}\left(\frac{n_{jk}\bar{y}_{jk} + n_k\theta_k}{n_k/k^2 + n_{jk}/\sigma_K^2}, \frac{1}{n_k/k^2 + n_{jk}/\sigma_k^2}\right) \\
\theta_j \mid \text{ resto } & \sim \mathrm{N}\left(\frac{\mu/\tau^2 + n_k\bar{\zeta}_k/\sigma_k^2}{1/\tau^2 + n_k/\sigma_k^2}, \frac{1}{1/\tau^2 + n_k/\sigma_k^2}\right) \\
\sigma_K^2 \mid \text{ resto } & \sim \mathrm{G}\left(\frac{n_k + \nu}{2}, \frac{\nu\sigma^2 + \sum_{j=1}^{n_k}\left(\zeta_{jk} - \theta_k\right)^2)}{2}\right) \\
\sigma^2 \mid \text{ resto } & \sim \mathrm{G}\left(\frac{m\nu + \alpha_0}{2}, \frac{\beta_0 + \nu\sum_{k=1}^{m}\left(\frac{1}{\sigma_k^2}\right)}{2}\right) \\
\mu \mid \text{ resto } & \sim \mathrm{N}\left(\frac{\mu_0/\gamma_0^2 + m\bar{\theta}/\tau^2}{1/\gamma_0^2 + m/\tau^2}, \frac{1}{1/\gamma_0^2 + m/\tau^2}\right) \\
\tau^2 \mid \text{ resto } & \sim \mathrm{GI}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{k=1}^{m}\left(\theta_k - \mu\right)^2}{2}\right) \\
k^2 \mid \text{ resto } & \sim \mathrm{GI}\left(\frac{\sum_{k=1}^{m}\sum_{j=1}^{n_k} n_{j,k} + \xi_0}{2}, \frac{\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{j,k}}\left(y_{i,j,k} - \zeta_{j,k}\right)^2 + \xi_0 k_0^2}{2}\right)
\end{aligned}
$$

Se emplean las distribuciones previas difusas definidas por los siguientes hiperparámetros a partir de la información de la prueba:

- M_1: $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\nu_0 = 1$, $\sigma_0^2 = 50^2$.
- M_2: $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\eta_0 = 1$, $\tau_0^2 = 50^2$, $\nu_0 = 1$, $\sigma_0^2 = 50^2$.
- M_3: $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\eta_0 = 1$, $\tau_0^2 = 50^2$, $\nu = 1$, $\alpha_0 = 1, \beta_0 = \frac{1}{50^2}$.
- M_4: $\xi_0 = 1$, $k_0^2 = 50^2$, $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\eta_0 = 1$, $\tau_0^2 = 50^2$, $\nu_0 = 1$, $\sigma_0^2 = 50^2$.
- M_5: $\xi_0 = 1$, $k_0^2 = 50^2$, $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\eta_0 = 1$, $\tau_0^2 = 50^2$, $\nu 0 = 1$, $\alpha_0 = 1, \beta_0 = \frac{1}{50^2}$.

Referencias

[1] DANE (2018). Pobreza Monetaria y Multidimensional en Colombia 2018. Recuperado de https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-y-desigualdad/pobreza-monetaria-y-multidimensional-en-colombia-2018#pobreza-por-departamentos-2018

[2] Datos abiertos (2023). MEN estadísticas en educación en preescolar, básica y media por municipio. Recuperado de https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/nudc-7mev

[3] MEN. Definición Cobertura Neta Recuperado de https://www.mineducacion.gov.co/1621/article-82702.html

[4] Sosa, J., Aristizabal, J. (2022). Some Developments in Bayesian Hierarchical Linear Regression Modeling. Facultad de Administración de Empresas, Universidad Externado de Colombia, Bogotá, Colombia. Recuperado de https://revistas.unal.edu.co/index.php/estad/article/view/98988/83944

[5] Sosa, J. (2023). Comparación de modelos (Notas de clase). Recuperado de https://rpubs.com/jstats1702/967868

[6] Sosa, J. (2023). Modelos de mezcla finitos (Notas de clase). Recuperado de https://rpubs.com/jstats1702/961380

[7] Sosa, J. (2023). Modelo jerárquico Normal con medias específicas (Notas de clase). Recuperado de https://rpubs.com/jstats1702/950834

[8] Sosa, J. (2023). Modelo jerárquico Normal con medias y varianzas específicas (Notas de clase). Recuperado de https://rpubs.com/jstats1702/954522