

UNIVERSIDAD NACIONAL DE COLOMBIA

Facultad de ciencias

Departamento de estadística

Segundo caso de estudio Estadística Bayesiana
2023-2

Autor

Jherson Guzman R., Sebastián A. Lozano R.

Docente

Juan Camilo Sosa Martínez

Octubre 2023



Ejercicios

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

Para el panel de la izquierda se utilizó la escala de color azul degradado para el puntaje medio en las pruebas Saber Pro del 2022-II departamentales, donde se muestran las etiquetas únicamente para aquellos departamentos con un puntaje medio inferior al percentil 10 (P_{10}). En el panel de la derecha tenemos el Índice de pobreza por departamento del año 2018 utilizando un degradado de color amarillo, en este panel solo se muestran las etiquetas de los departamentos que superan el percentil 90 (P_{90}), es decir los departamentos con mayor índice de Pobreza.

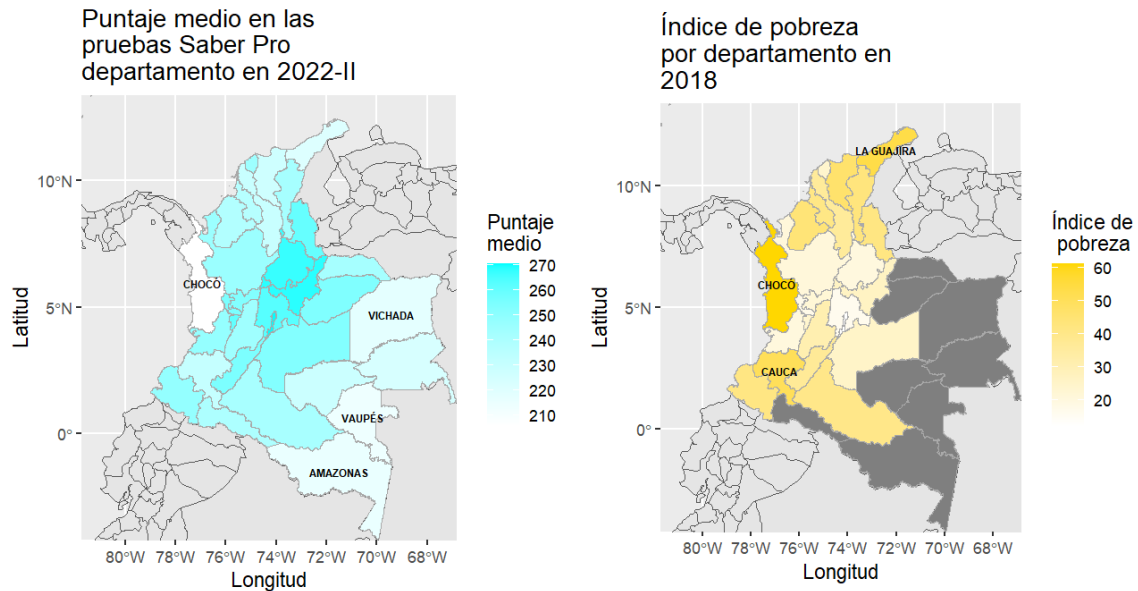


Figura 1: Mapas de puntaje medio en pruebas departamental en 2022-II y de Índice de Pobreza departamental en 2018

El gráfico anterior sugiere que existe una relación inversamente proporcional entre el índice de pobreza y los puntajes medios de la prueba saber, es decir a mayor índice de pobreza el puntaje medio será menor. Este comportamiento se evidencia claramente en el departamento del Chocó, que tiene uno de los puntajes medios más bajos y a su vez uno de los índices de pobreza más altos.

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

En el panel izquierdo se utiliza la escala de color azul degradado para el puntaje medio en las pruebas Saber Pro del 2022-II municipales. En el panel de la derecha tenemos la cobertura neta secundaria por municipio para el año 2022, se utiliza una escala de color verde degradado.

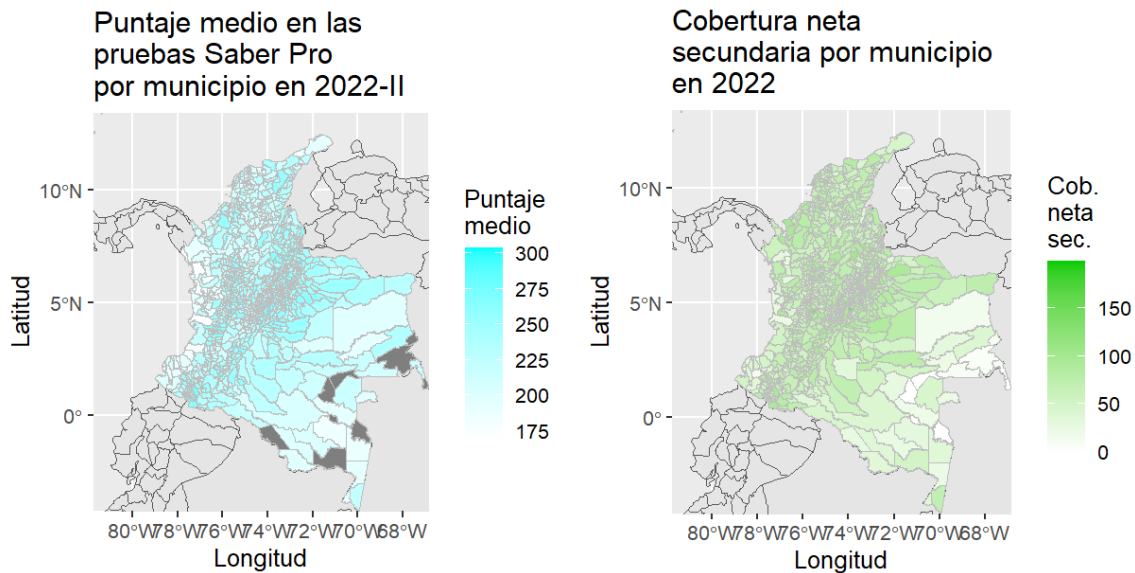


Figura 2: Mapas de puntaje medio en pruebas Saber Pro municipal en 2022-II y de Cobertura neta secundaria por municipio en 2022

El gráfico anterior sugiere que en los municipios en los cuales hay menor cobertura neta secundaria se observa un puntaje inferior en el valor promedio de las pruebas, como se ve en la región de Orinoquía y Amazonas. Así, en los municipios con mayor cobertura se pueden ver resultados altos de puntaje medio, como se puede visualizar en la región Central y Andina del país.

3. En un gráfico con cuatro paneles (2×2), hacer el DAG de M_2 (panel 1, esquina superior izquierda), M_3 (panel 2, esquina superior derecha), M_4 (panel 3, esquina inferior izquierda) y M_5 (panel 1, esquina inferior derecha).

Solución

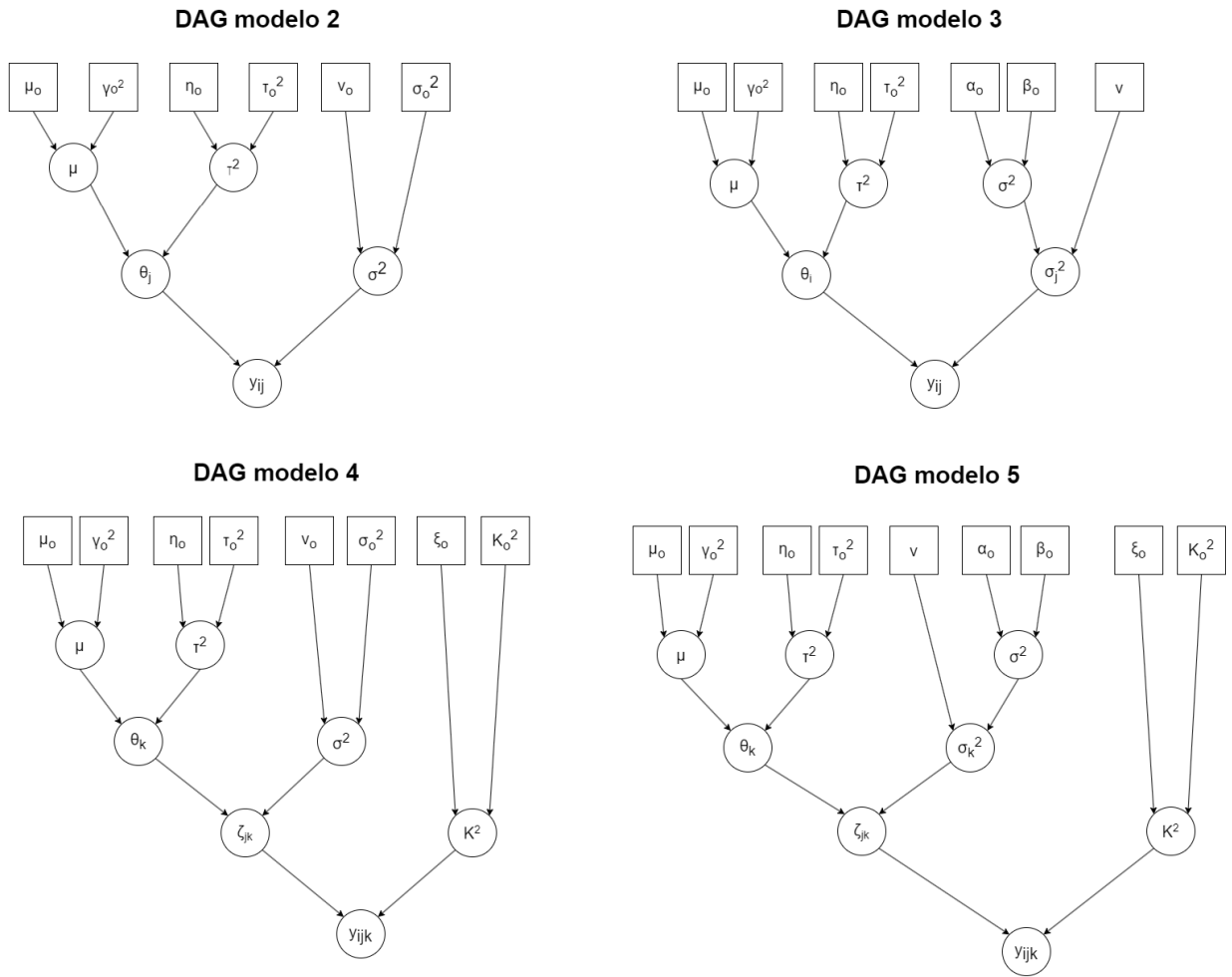


Figura 3: DAG modelos

4. En un gráfico con cuatro paneles (2×2), dibujar la cadena de la log-verosimilitud de M_2 (panel 1, esquina superior izquierda), M_3 (panel 2, esquina superior derecha), M_4 (panel 3, esquina inferior izquierda) y M_5 (panel 4, esquina inferior derecha). Los gráficos deben tener la misma escala para facilitar la comparación. Interpretar los resultados obtenidos (máximo 100 palabras).

Nota: Incluir un apéndice al final del informe con las distribuciones condicionales completas (no incluir la demostración, solo cada distribución con sus respectivos parámetros) y un resumen de los coeficientes de variación de Monte Carlo de cada parámetro de cada modelo. Tomar como ejemplo la Sección 4.3 del artículo

Solución

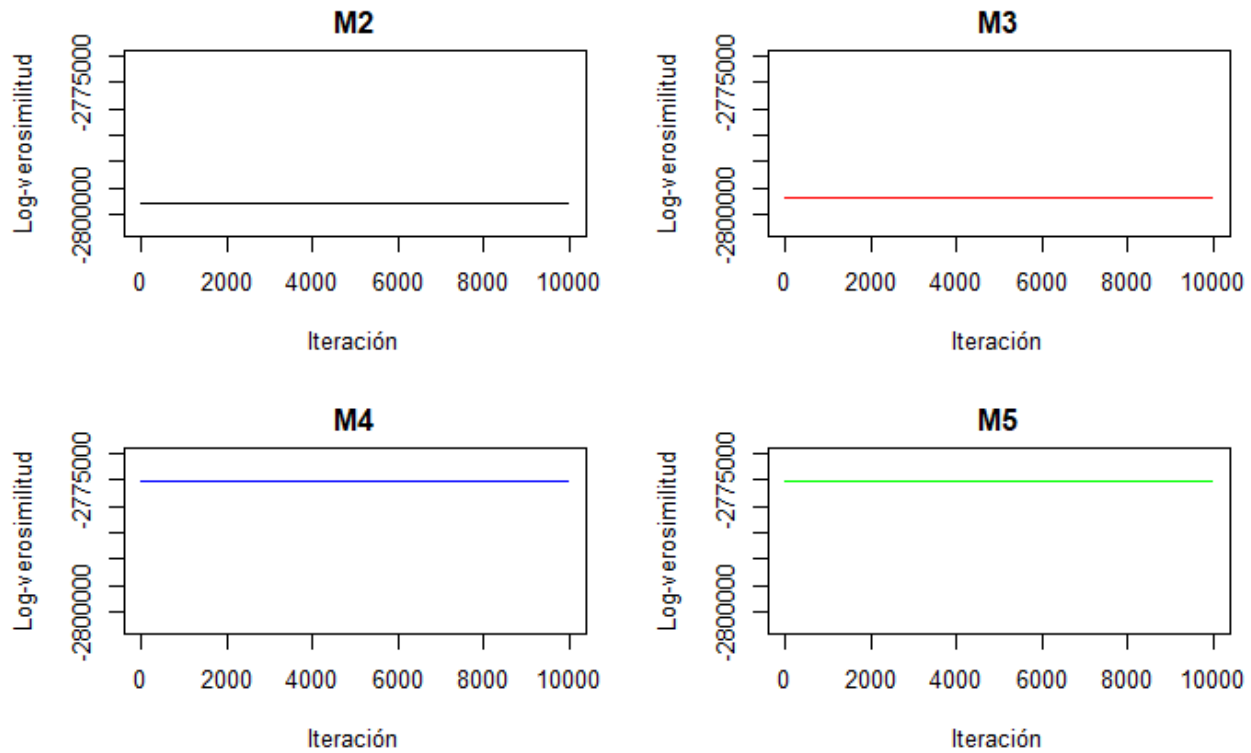


Figura 4: Panel con cadenas de log-verosimilitud de los modelos

Se puede observar como el comportamiento de la log-verosimilitud de los modelos indica que las cadenas convergen, en cuanto a los modelos M_2 y M_3 rondan un valor similar entre ellos, al igual que los modelos M_4 y M_5 .

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

Solución

Modelo	DIC	WAIC
M_1	5636785	5636785
M_2	5595932	5595930
M_3	5594240	5594236
M_4	5327057	5551603
M_5	5327223	5551593

Tabla 1: Tabla de comparación de modelos: DIC y WAIC

Según la anterior tabla el M_4 es el modelo con menor DIC, por otro lado el M_5 es el modelo con menor WAIC. Lo que nos quiere decir que los modelos que mejor precisión predictiva tienen, son los modelos M_4 y M_5 , de vernos obligados a elegir uno, deberíamos elegir a M_4 pues es el modelo más parsimonioso entre los dos. Sin embargo, es importante resaltar que la diferencia en el WAIC y DIC para estos modelos, no es mucha, lo que podría indicar que de usar otra semilla podríamos obtener un resultado que favoreciera más al M_5 .

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

	M_1	M_2	M_3	M_4	M_5
Media posterior	250.274	242.273	242.313	232.495	232.490
Lim Inf ($IC_{95\%}$)	250.131	235.470	235.507	225.917	225.753
Lim Sup ($IC_{95\%}$)	250.413	248.952	249.186	239.168	239.056

Tabla 2: Media posterior e intervalos de credibilidad al 95 % para μ de los modelos M_k

El parámetro μ se interpreta como la media general de los resultados de las pruebas Saber a nivel nacional. Es importante destacar que estas pruebas fueron diseñadas con la

intención de que su valor medio fuese de 250. Sin embargo, al analizar las estimaciones puntuales de los distintos modelos, exceptuando el modelo $M1$, observamos que todas ellas son inferiores a 250. Este hallazgo sugiere que el desempeño general en las pruebas Saber a nivel nacional es, en promedio, inferior a lo inicialmente esperado. Este argumento se ve respaldado aún más por el hecho de que ninguno de los intervalos de credibilidad, calculados al 95 %, contiene el valor de 250.

- Usando M_5 , 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

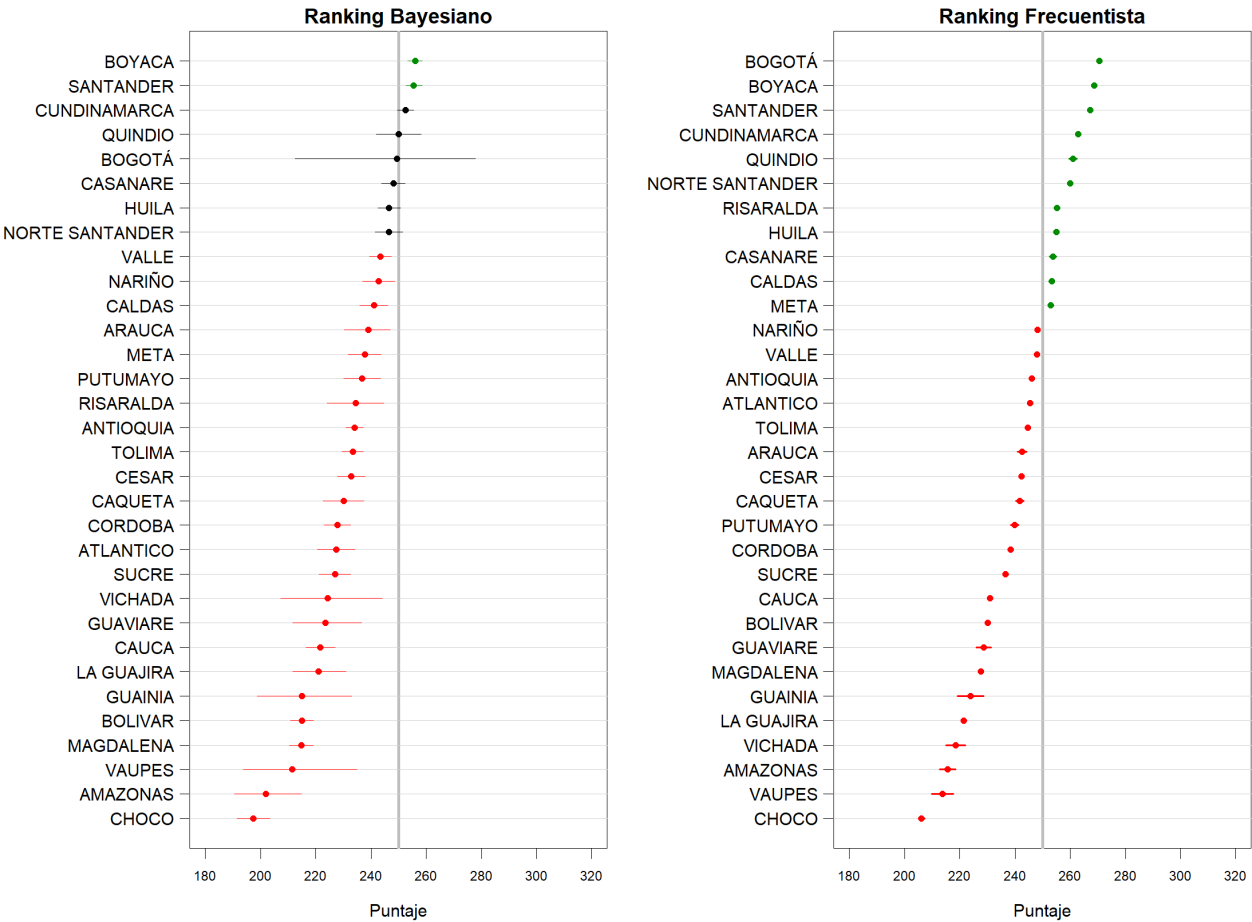


Figura 5: Panel con ranking bayesiano y frecuentista departamental

Lo primero que destaca es que los dos tipos de rankings no muestran los departamentos en el mismo orden. Además se puede observar que los intervalos de credibilidad en el ranking bayesiano son mucho más amplios, como es el caso de Bogotá, comparados con los intervalos de confianza del ranking frecuentista. Por último es importante resaltar que en ambos rankings la mayoría de departamentos evidencian un puntaje significativamente menor a 250, aunque en el ranking bayesiano si muestra departamentos que no difieren significativamente de dicho valor, cosa que no sucede para el otro tipo de ranking.

8. Usando M_5 , 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

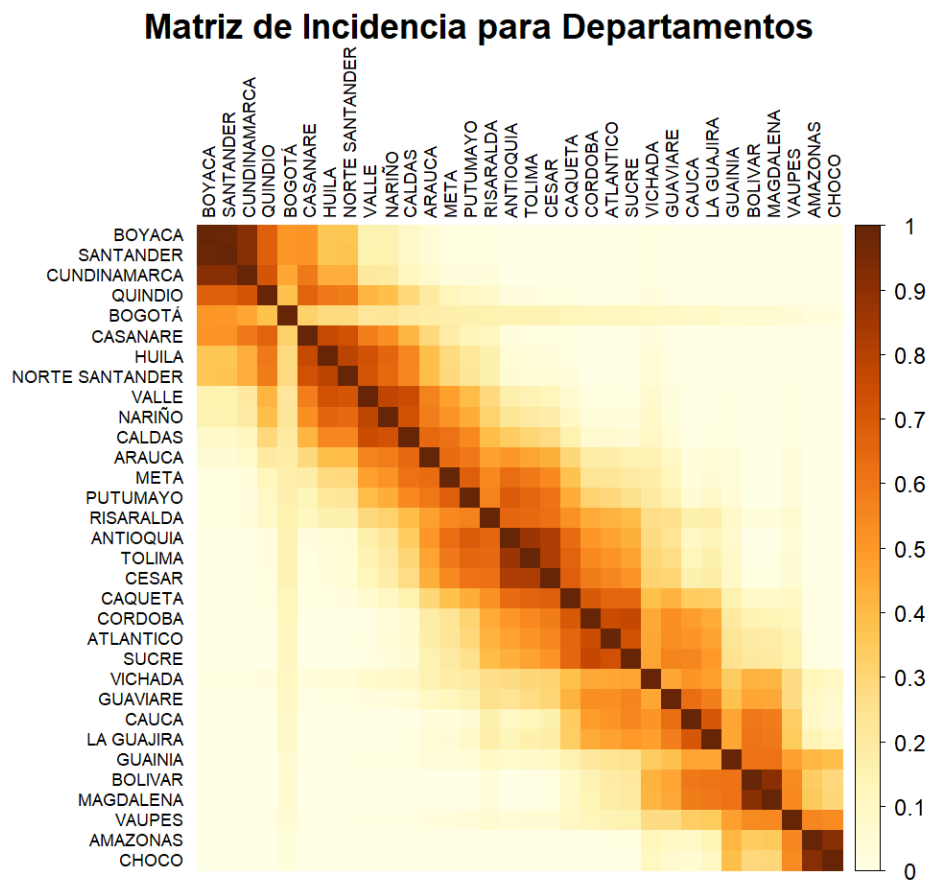


Figura 6: Matriz de incidencia para departamentos ordenada según el ranking Bayesiano

De la figura 6, observamos que los departamentos se ven altamente correlacionados con aquellos otros que están “cerca” de acuerdo al ranking bayesiano dado su puntaje medio en las Pruebas Saber, en sentido contrario, aquellos departamentos que están “lejos” se ven bajamente correlacionados. Adicionalmente, al darle este orden los clusters no se diferencian claramente.

**Clustering por departamentos en Colombia
segun el puntaje medio en la prueba SABER 2022-II**

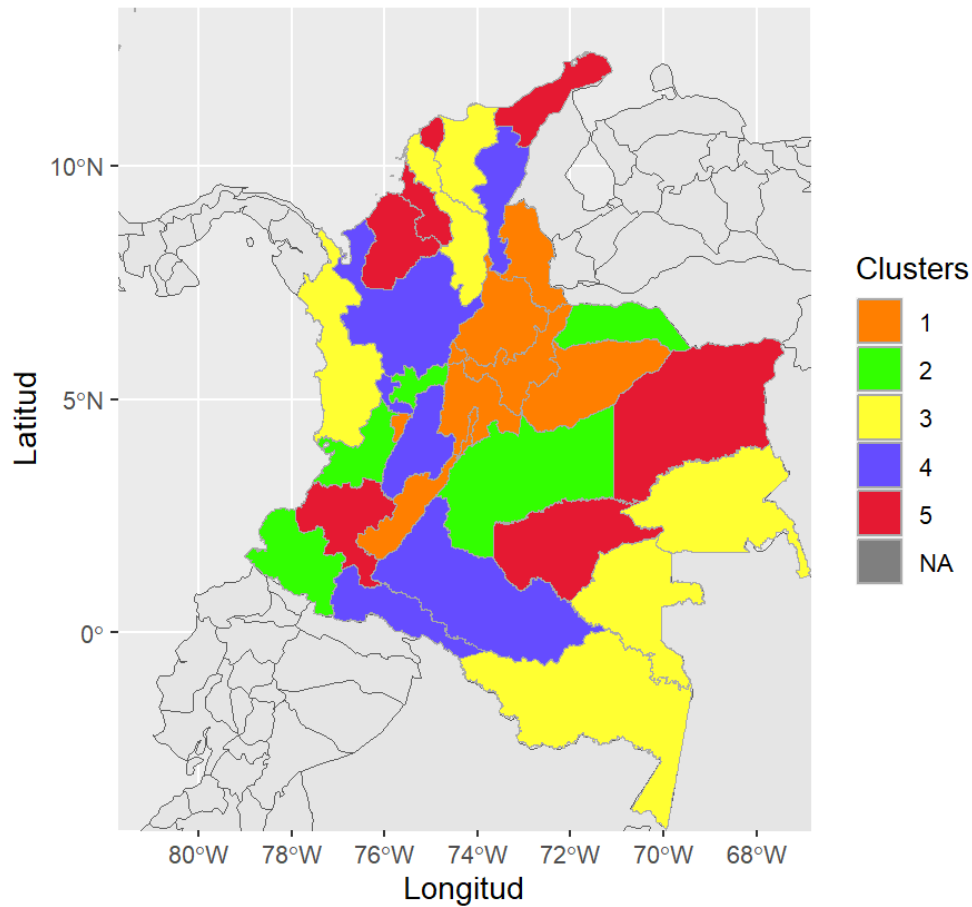


Figura 7: Mapa con clusters por departamentos

De la figura 7 observamos que puede existir una característica geográfica que influye en el agrupamiento de los departamentos de acuerdo al puntaje medio en las Pruebas Saber, esto se nota más claramente en los cluster 1 (color naranja) y el 4 (color azul) pues dichos clusters parecen concentrados en una misma zona geográfica.

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 M_5 . Presentar los resultados tabularmente (organizados descendente de acuerdo con la media posterior) y visualmente (por medio de un mapa usando la media posterior).

Solución

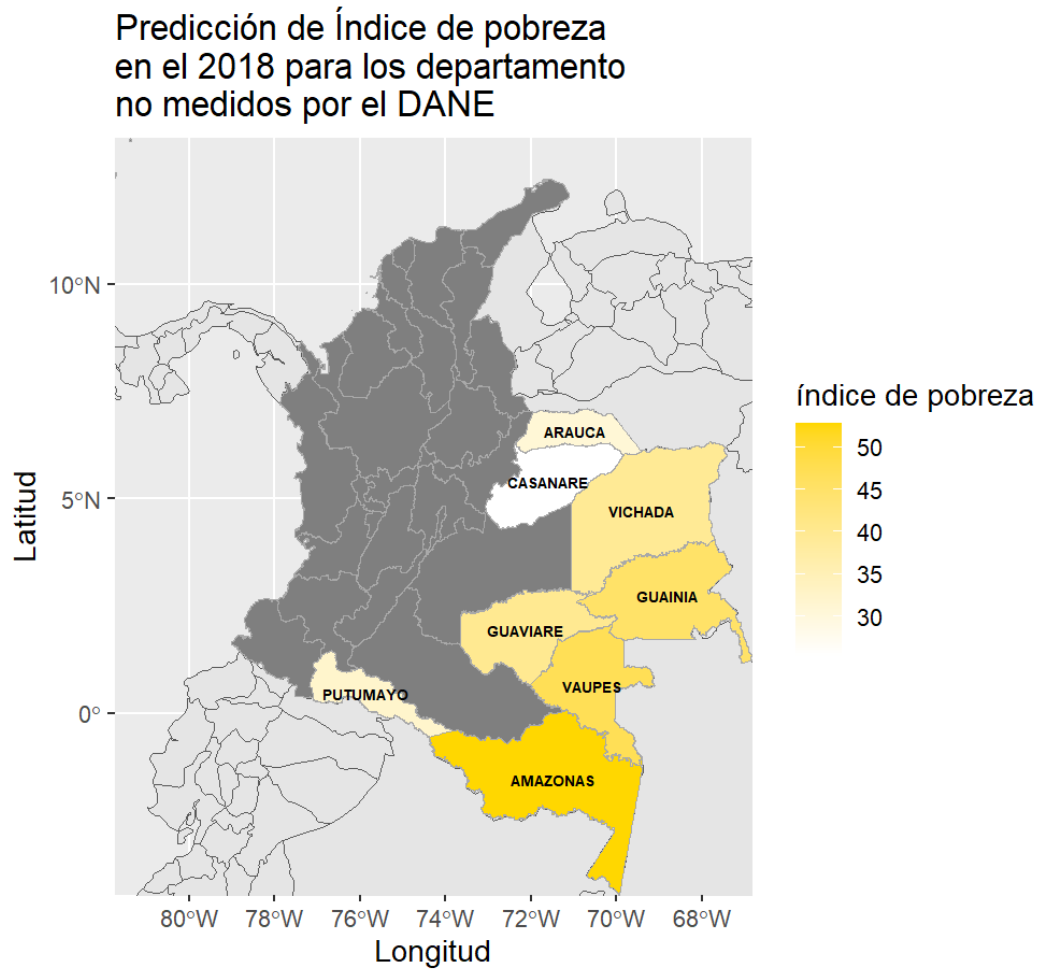


Figura 8: Predicción del Índice de Pobreza para los departamentos no medidos

Departamento	Predicción	Límite inferior al 95 %	Límite superior al 95 %
Amazonas	52.80	44.16	60.72
Vaupés	47.11	33.20	58.35
Guainía	45.06	34.41	55.42
Guaviare	40.04	32.33	47.51
Vichada	39.54	27.92	50.02
Putumayo	32.18	28.18	36.15
Arauca	30.84	26.08	36.06
Casanare	25.46	22.56	28.35

Tabla 3: Predicción e intervalos de predicción del Índice de Pobreza para los departamentos no medidos

Según las predicciones se puede observar que los departamentos que tendrían mayor Índice de Pobreza para el año 2018, están ubicados Amazonas, Vaupés y Guainía, por otro lado el departamento de Casanare tiene un Índice de pobreza relativamente bajo. Además utilizando la información del ranking bayesiano, se fortalece la hipótesis propuesta en el punto 1: evidenciamos que aquellos departamentos un bajo Puntaje Medio en la prueba Saber suelen tener un índice de pobreza más alto, así es el caso de la predicción de los departamentos del Amazonas y Vaupés, los cuales obtuvieron las predcciones más altas y a su vez ocupan los puestos 31 y 30 en el ranking bayesiano respectivamente.

- Usando M_5 , hacer el *ranking* de los municipios basados 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 pertenecen al mismo grupo. Interpretar los resultados obtenidos (máximo 100 palabras).

Solución

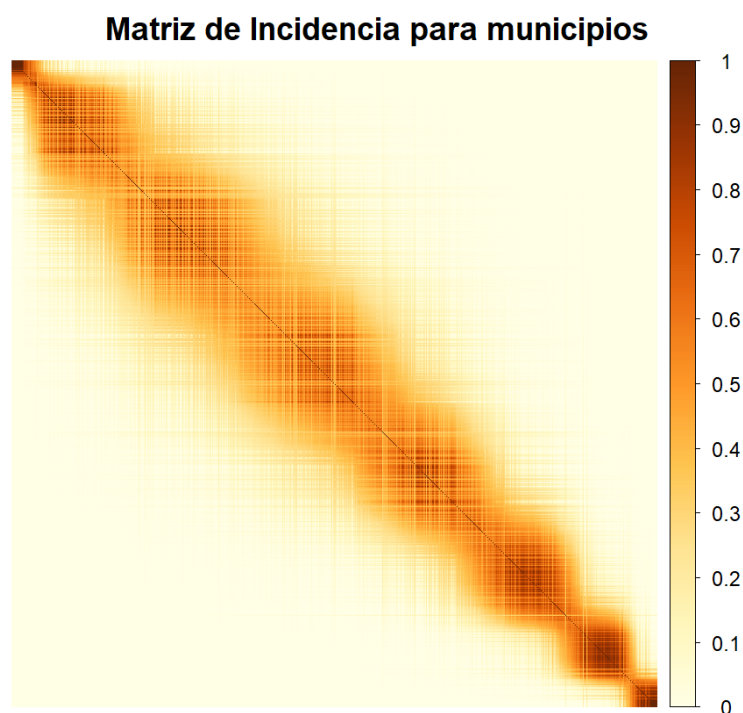


Figura 9: Matriz de incidencia para municipios ordenada según el ranking Bayesiano

Respecto a la matriz de incidencia obtuvimos resultados similares al punto 8, donde municipios ubicados en la parte alta del ranking están fuertemente correlacionados con otros municipios en posiciones cercanas del ranking, y a su vez no se ve una clara correlación con aquellos municipios distantes en el ranking.

El mapa muestra que varios municipios de un mismo cluster tienden a agruparse en una misma zona geográfica, el caso que mejor que lo representa es el cluster 4 (morado) en la región pácifica. Nuevamente, esto podría ser un indicio de que la posición geográfica del municipio afecta de algún modo en desempeño medio de la prueba Saber.

Clustering por puntaje medio en la prueba SABER por municipios en Colombia

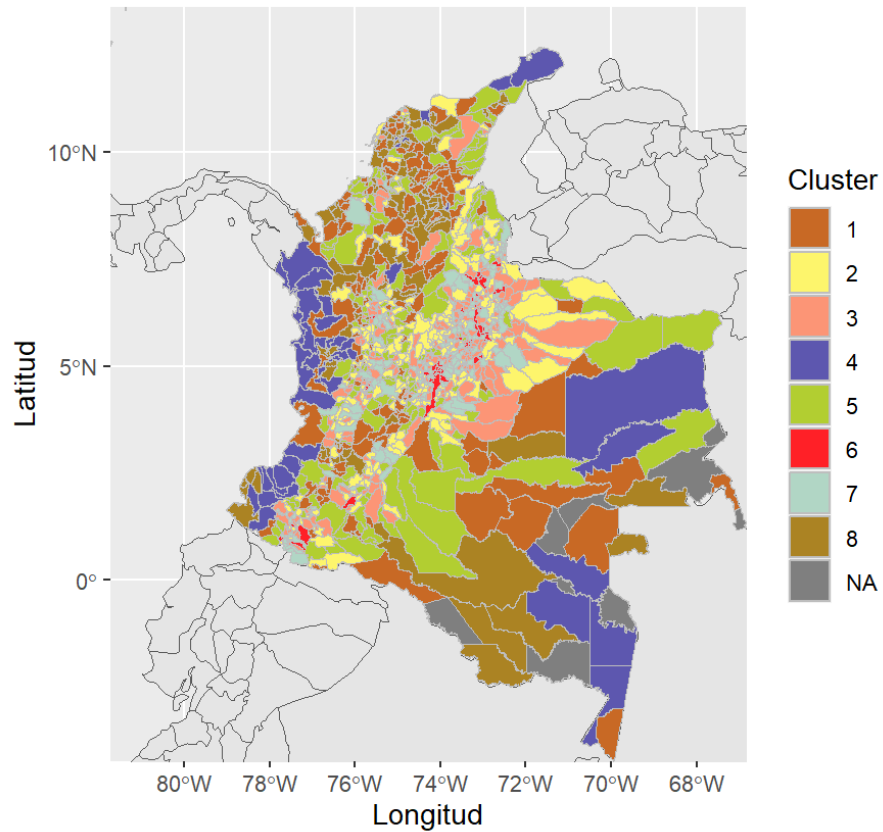


Figura 10: Mapa con clusters por municipios

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 M_5 . Presentar los resultados tabularmente (organizados descendente de acuerdo con la media posterior) y visualmente (por medio de un mapa usando la media posterior).

Solución

Nombre municipio	Departamento	Código municipio	Predicción	Límite inferior al 95 %	Límite superior al 95 %
Belén de Bajirá	Chocó	27086	57.249	54.097	60.390
Mapiripana	Guainía	94663	55.264	46.484	63.588

Tabla 4: Predicción e intervalos de credibilidad de la cobertura neta secundaria para los municipios no medidos

Si analizamos la tabla anterior acompañada del ranking bayesiano, vemos que estos municipios tienen una predicción puntual baja para la cobertura neta secundaria, y que hacen parte de departamentos con puntajes medios bajos para la prueba saber, como lo son **Chocó** y **Guainía**, lo anterior fortalece la hipótesis planteada en el punto 2, donde a una menor cobertura neta secundaria, menor será el puntaje medio del municipio.

Nota: No fue posible realizar el mapa, pues el municipio de Maripina esta catalogado como un centro poblado en el geoportal del DANE mientras que el municipio de Belén de Bajirá fue creado en el año 2022, por lo que en ninguno de los dos casos se tiene dentro del shapefile para la realización del mapa.

12. Validar la bondad ajuste de M_5 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

Vemos que en los estadísticos de prueba de la media, el máximo o la mediana, los valores ppp aparentemente se distribuyen uniformemente por todo el posible rango de valores, lo que indica que algunos municipios se adecuan correctamente, mientras que otros subestiman o sobrestiman el valor real. Por otro lado, con la desviación estándar o el IQR se ven mejores resultados, pues la mayoría de valores ppp son valores cercanos a 0.5, o sea el modelo se ajusta correctamente. Por último, con el mínimo, parece ser que el modelo no capta adecuadamente esta información, pues la mayoría de puntos sobrepasan el 0.5 o sea se subestiman el valor real.

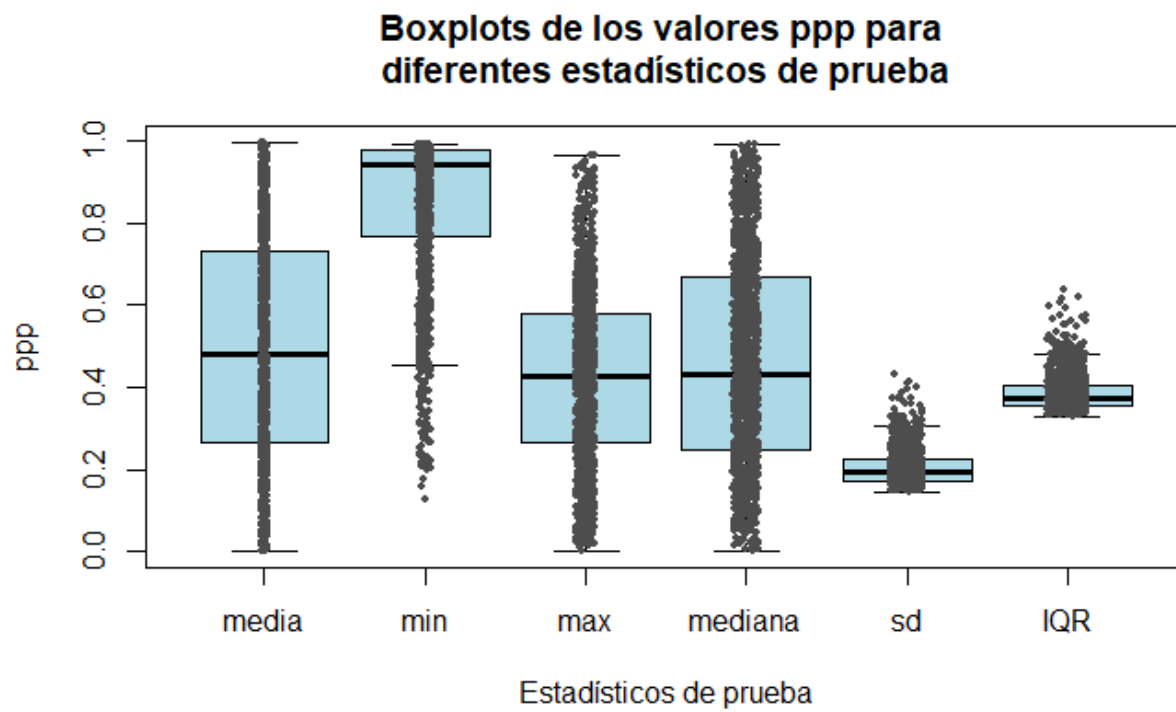


Figura 11: Bondad de ajuste

Anexo

A continuación se van a mostrar las distribuciones condicionales completas para los modelos trabajados en el caso.

Modelo 1:

$$\begin{aligned}(\theta|-) &\sim N\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\gamma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\gamma_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\gamma_0^2}}\right) \\(\sigma^2|-) &\sim GI\left(\frac{n + \nu_0}{2}, \frac{\nu_0\sigma_0^2 + \sum_{i=1}^n (y_i - \theta)^2}{2}\right)\end{aligned}$$

Modelo 2:

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

Modelo 3:

$$\begin{aligned}(\theta_j|-) &\sim N\left(\frac{n_j\bar{y}_j/\sigma_j^2 + \mu/\tau^2}{n_j/\sigma_j^2 + 1/\tau^2}, \frac{1}{n_j/\sigma_j^2 + 1/\tau^2}\right) \\(\sigma_j^2|-) &\sim GI\left(\frac{n_j + \nu}{2}, \frac{\nu\sigma^2 + \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2}{2}\right)\end{aligned}$$

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

Modelo 4:

$$\begin{aligned}
(\zeta_{jk}| -) &\sim N\left(\frac{n_{jk}\bar{y}_i/\kappa^2 + \theta_k/\sigma^2}{n_{jk}/\kappa^2 + 1/\sigma^2}, \frac{1}{n_{jk}/\kappa^2 + 1/\sigma^2}\right) \\
(\kappa^2| -) &\sim GI\left(\frac{\sum_{k=1}^m \sum_{j=1}^{n_k} n_{jk} + \xi_0}{2}, \frac{\xi_0\kappa_0^2 + \sum_{k=1}^m \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} (y_{ijk} - \zeta_{jk})^2}{2}\right) \\
(\theta_k| -) &\sim N\left(\frac{n_k\bar{\zeta}_{jk}/\sigma^2 + \mu/\tau^2}{n_k/\sigma^2 + 1/\tau^2}, \frac{1}{n_k/\sigma^2 + 1/\tau^2}\right) \\
(\mu| -) &\sim N\left(\frac{m\bar{\theta}_k/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2}\right) \\
(\tau^2| -) &\sim GI\left(\frac{m + \eta_0}{2}, \frac{\eta_0\tau_0^2 + \sum_{k=1}^m (\theta_k - \mu)^2}{2}\right) \\
(\sigma^2| -) &\sim GI\left(\frac{\sum_{k=1}^m n_k + \nu_0}{2}, \frac{\nu_0\sigma_0^2 + \sum_{k=1}^m \sum_{j=1}^{n_k} (\zeta_{jk} - \theta_k)^2}{2}\right)
\end{aligned}$$

Modelo 5:

$$\begin{aligned}
(\zeta_{jk}| -) &\sim N\left(\frac{n_{jk}\bar{y}_{ijk}/\kappa^2 + \theta_k/\sigma_k^2}{n_{jk}/\kappa^2 + 1/\sigma_k^2}, \frac{1}{n_{jk}/\kappa^2 + 1/\sigma_k^2}\right) \\
(\kappa^2| -) &\sim GI\left(\frac{\sum_{k=1}^m \sum_{j=1}^{n_k} n_{jk} + \xi_0}{2}, \frac{\xi_0\kappa_0^2 + \sum_{k=1}^m \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} (y_{ijk} - \zeta_{jk})^2}{2}\right) \\
(\theta_k| -) &\sim N\left(\frac{n_k\bar{\zeta}_{jk}/\sigma_k^2 + \mu/\tau^2}{n_k/\sigma_k^2 + 1/\tau^2}, \frac{1}{n_k/\sigma_k^2 + 1/\tau^2}\right) \\
(\mu| -) &\sim N\left(\frac{m\bar{\theta}_k/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2}\right)
\end{aligned}$$

$$\begin{aligned}
(\tau^2|-) &\sim GI\left(\frac{m + \eta_0}{2}, \frac{\eta_0\tau_0^2 + \sum_{k=1}^m(\theta_k - \mu)^2}{2}\right) \\
(\sigma_k^2|-) &\sim GI\left(\frac{n_k + \nu}{2}, \frac{\nu\sigma^2 + \sum_{j=1}^{n_k}(\zeta_{jk} - \theta_k)^2}{2}\right) \\
(\sigma^2|-) &\sim G\left(\frac{\nu + \alpha_0}{2}, \frac{\beta_0 + \nu \sum_{k=1}^m 1/\sigma_k^2}{2}\right)
\end{aligned}$$

Coeficientes de variación de Montecarlo

Todos los coeficientes de Variación de Montecarlo fueron multiplicados por 100 %

Modelo 1:

Parámetro	CVMC (%)
θ	0.003
σ^2	0.0019

Tabla 5: Tabla de CVMC para θ, σ^2

Modelo 2:

Parámetro	CVMC (%)					
	Mínimo	Q ₁	Mediana	Media	Q ₃	Máximo
θ_j	0.0007	0.0015	0.0018	0.0031	0.0033	0.0112

Tabla 6: Tabla de resumen de CVMC para θ_m

Parámetro	CVMC (%)
μ	0.0141
τ^2	0.2662
σ^2	0.0019

Tabla 7: Tabla de CVMC para μ, τ^2, σ^2

Modelo 3:

	CVMC (%)					
Parámetro	Mínimo	Q ₁	Mediana	Media	Q ₃	Máximo
θ_j	0.0007	0.0015	0.0018	0.0029	0.0029	0.0110
σ_j^2	0.0052	0.0107	0.0128	0.0202	0.0210	0.0726

Tabla 8: Tabla de resumen de CVMC para θ_j y σ_j^2

Parámetro	CVMC (%)
μ	0.0148
τ^2	0.2663
σ^2	0.2460

Tabla 9: Tabla de CVMC para μ, τ^2, σ^2

Modelo 4:

	CVMC (%)					
Parámetro	Mínimo	Q ₁	Mediana	Media	Q ₃	Máximo
ζ_{jk}	0.0007	0.0117	0.0169	0.0182	0.0237	0.0590
θ_k	0.0058	0.0104	0.0131	0.0169	0.0179	0.0472

Tabla 10: Tabla de resumen de CVMC para ζ_{jk} y θ_k

Parámetro	CVMC (%)
κ^2	0.0020
μ	0.0146
τ^2	0.2839
σ^2	0.0466

Tabla 11: Tabla de CVMC para $\kappa^2, \mu, \tau^2, \sigma^2$

Modelo 5:

	CVMC (%)					
Parámetro	Mínimo	Q ₁	Mediana	Media	Q ₃	Máximo
ζ_{jk}	0.0006	0.0116	0.0169	0.0180	0.0234	0.0650
θ_k	0.0052	0.0094	0.0124	0.0176	0.0184	0.0693
σ_k^2	0.1337	0.2456	0.3186	1.4895	0.5414	29.8878

Tabla 12: Tabla de resumen de CVMC para ζ_{jk} y θ_k y σ_k^2

Parámetro	CVMC (%)
κ^2	0.0020
μ	0.0145
τ^2	0.2792
σ^2	0.2773

Tabla 13: Tabla de CVMC para $\kappa^2, \mu, \tau^2, \sigma^2$

Referencias

- RPubs - Estadística bayesiana - Modelo jerárquico Normal con medias específicas. (s. f.). <https://rpubs.com/jstats1702/950834>
- RPubs - Estadística bayesiana - Modelo jerárquico Normal con medias y varianzas específicas. (s. f.). <https://rpubs.com/jstats1702/954522>
- Introducción a R - Luis Hernando Vanegas - <https://r2022.netlify.app/>
- Geoportal DANE - <https://www.dane.gov.co/files/geoportal-provisional/index.html>
- Guía orientación examen Saber 11 - Icfes - <https://www2.icfes.gov.co/gu%C3%ADa-de-orientaci%C3%B3n-examen-saber-11>