UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Proyecto de trabajo de grado
Marzo 2021

Desarrollo de modelo Bayesiano jerárquico de mezcla con número de componentes aleatorio para cuantificar el riesgo de suscripción en seguros de no vida

Development of Bayesian mixture model with number of random components in order to measure the risk subscription in nonlife insurance

Yesid Alberto Capera Martínez.[a]
yesidcapera@usantotomas.edu.co

Juan Camilo Sosa Martínez.[b]
jcsosam@unal.edu.co

Resumen

Se propone desarrollar un modelo jerárquico de mezcla completamente Bayesiano con número de componentes aleatorio en la severidad de los siniestros, para cuantificar el riesgo de suscripción en seguros de no vida. El riesgo de suscripción desde el punto de vista de la tarifación se asocia con la probabilidad de insolvencia que tiene una compañía aseguradora. Esta probabilidad depende de un proceso estocástico de ruina que pone en consideración: el capital inicial de la compañía, el monto total de primas recibidas y el monto de pérdida agregada en siniestros (Rincón, [12]). En consecuencia, modelar las distribuciones posteriores para la frecuencia y severidad de los siniestros en este proceso estocástico, resulta ser apropiado para cuantificar el riesgo de suscripción. Con respecto a la incertidumbre que existe en el número de clústeres asociados a riesgos presentes en la distribución de severidad de los siniestros, se hace uso de la propuesta de Gelman, A. et al. (2014) vía cadenas de Markov de Monte Carlo para el número de componentes de la mezcla. Una vez se verifica tanto la bondad de ajuste del modelo como los criterios Bayesianos WAIC y DIC, se establece una comparación vía validación cruzada $k-$fold entre los resultados de frecuencia y severidad de los siniestros con los modelos de enfoque clásico para el riesgo de suscripción (e.g, Barañano, De La Peña y Garayeta, 2015). El objetivo es mostrar si estos resultados son más eficientes con el modelo Bayesiano propuesto al evaluar medidas de calidad del pronóstico tales como el error cuadrático medio y el error porcentual absoluto medio. Posteriormente, los mejores modelos para frecuencia y severidad obtenidos en la validación cruzada se utilizan como una importante aplicación que permite medir el riesgo de suscripción en seguros de no vida o Seguros Generales mediante un proceso de estocástico de ruina y los cálculos de requerimiento de capital para volumen de primas y reservas tomados de la Directiva de Solvencia II. La metodología propuesta se ilustra por medio del conjunto de datos de siniestros dataCar que se encuentra en la librería insuranceData del software estadístico R.

***Palabras clave*:** (Frecuencia, Severidad, Pérdida Agregada, Modelo de mezcla completamente Bayesiano, Riesgo de suscripción, Probabilidad de Insolvencia, Ruina).

Resumen

The purpose of this project is developmenting a fully Bayesian hierarchical mixture model with a random number of components, in order to quantify the risk of subscription in non-life insurance. Subscription risk

[a]Estudiante
[b]Director

from the point of view of pricing is associated with an insurance company's probability of insolvency. This probability depends on a ruin stochastic process who is in function of the initial capital of the company, the total premiums received and the aggregate loss of claims (Rincón, [12]). The posterior distribution for the parameters of severity claims in the risk model is obtained with the implementation of the Gelman et. al's proposal (2014) for the number of mixture random components. The goodness of fit for model is verified with WAIC and DIC Bayesian criteria, after we will compare with cross-validation the results of frequency and severity of claims with the classic risk models (eg, Barañano, De La Peña and Garayeta, 2015). The objective is showing if these results are more efficient with the proposed Bayesian model when we forecast and we verify the mean square error and mean absolute percentage error into cross validation. The methodology is illustrated by the dataCar claims dataset in the insuranceData library of the R statistical software. –

***Palabras clave*:** (Frequency, Severity, Aggregated Loss, Fully Bayesian Mixture Model, Subscription Risk, Insolvency probability, ruin stochastic process).

1. Introducción

La medición del riesgo de suscripción en el sector asegurador surge ante el problema de obtener un margen de rentabilidad esperado cuando el producto de un ramo entra en operación. Este problema deriva en la necesidad de evaluar la probabilidad de incurrir en pérdidas cuando las primas que se cobran para amparar determinadas contingencias no son suficientes para realizar el pago de los siniestros, así como los gastos administrativos y de personal. Según el Marco Integral de Supervisión de la Superintendencia Financiera de Colombia, parte II, Titulo IV, Capítulo IV, numeral 1.3.4, las posibles razones por las cuales las primas son insuficientes para cubrir los gastos y los siniestros son: cálculos incorrectos en las tarifas , descuentos inadecuados en las primas, concentración de los riesgos asumidos en la suscripción y diferencia de condiciones aceptadas simultáneamente por el tomador de un contrato de seguros y el reasegurador.

Desde el punto de vista de la tarifación, el riesgo de suscripción se entiende como la probabilidad de entrar en pérdidas cuando las primas cobradas a los tomadores de un contrato de seguros son insuficientes para cubrir los siniestros y los gastos. Bajo esta perspectiva, interesa modelar una variable aleatoria como la del monto total de los siniestros ocurridos dentro del tiempo de vigencia aceptado en el proceso de suscripción. La variable del monto total de los siniestros se denota por S y depende de dos variables aleatorias independientes: el número $N(t)$ de siniestros ocurridos dentro de un tiempo de vigencia $[0, t]$, y la cuantía X_i del $i-$ésimo siniestro expresada en una misma unidad monetaria. La variable S se define de la siguiente manera:

$$S = \sum_{i=1}^{N(t)} X_i. \tag{1}$$

De acuerdo con Rincón(2012), la media y la varianza de la variable aleatoria S se obtienen de la siguiente manera:

$$\mathbb{E}(S) = \mu_N \cdot \mu_X \tag{2}$$

$$\mathbb{V}\text{ar}(S) = \sigma_x^2 \cdot \mu_N + \sigma_N^2 \cdot \mu_x^2 \tag{3}$$

En la literatura de la ciencia estadística y actuarial, se ha estudiado extensivamente el monto total de los siniestros por medio de modelos de distribución clásicos paramétricos (Leigh et al. 2009, Tse 2009, Rincón 2012, Madroñal 2014, Barañano et al. 2015, Sarabia et al. 2017). Bajo estos modelos, se estima el valor esperado de la variable aleatoria S por medio del ajuste de una distribución discreta para el número de siniestros ocurridos y una distribución continua de soporte positivo para el monto de la cuantía de los siniestros. Aunque este enfoque clásico resulta valioso, siempre existirán dudas frente a las distribuciones

de las variables aleatorias del número de siniestros ocurridos y su cuantía, pues se encuentra presente la incertidumbre de su eficiencia en las predicciones de la frecuencia, la severidad y el monto total de los siniestros para estimar de manera adecuada las tasas puras de riesgo y las tasas comerciales que garanticen la rentabilidad del negocio asegurador en futuros periodos de suscripción.

Barañano, De La Peña y Garayeta (2015) en su metodología para la estimación del riesgo de suscripción mediante modelos internos de solvencia II para el mercado español de seguros multirriesgo de hogar entre 2008 y 2013, utilizan las pruebas de kolmogorov-Smirnov, Anderson-Darling y Chi-Cuadrado para encontrar las distribuciones discretas y continuas que mejor ajuste presenten sobre los conjuntos de datos de número de siniestros ocurridos y de sus cuantías para posteriormente modelar con un enfoque clásico de Simulación Monte Carlo, la pérdida agregada esperada en este ramo de no vida. Sin embargo, durante el desarrollo de la metodología se ignora si la distribución de la cuantía de los siniestros es de cola ligera o cola pesada, o si esta tiene al menos dos componentes que corresponden a un modelo de mezcla. El problema fundamental de omitir diagnósticos de cola ligera o cola pesada, así como de posibles componentes de mezcla en una distribución continua de soporte positivo deriva en consecuencias como la subestimación de la probabilidad de ocurrencia de siniestros de alta cuantía o severidad, los cuales pueden generar insolvencia en una aseguradora al realizar el proceso de suscripción con primas insuficientes para cubrir los costos. De esta manera, los montos de pérdida agregada esperada no necesariamente permiten calcular la mejor estimación para las reservas de siniestros sino para los requerimientos de capital que pueden exigir las normativas vigentes frente al riesgo de suscripción del volumen de primas y reservas para las compañías aseguradoras.

En otro estudio similar, Omary, C. el. al (2018) describieron el modelamiento clásico para la frecuencia y la severidad en siniestros de daños en automóviles utilizando distribuciones estadísticas. Una vez realizado el análisis descriptivo de los datos junto con algunas visualizaciones de las distribuciones del número de siniestros y los montos de sus cuantías, la propuesta para estimar el valor total de pérdida agregada consistió en implementar la prueba Chi Cuadrado para verificar si estadísticamente distribuciones discretas tales como: Binomial, Geométrica, Binomial Negativa y Poisson se ajustan al número de siniestros en conjuntos de datos del software estadístico R como AutoCollision, dataCar y dataOhlsson. Adicionalmente, se ilustraron los resultados de los estadísticos de prueba para comprobar la bondad de ajuste de distribuciones tales como: exponencial, gamma, Weibull, Pareto y lognormal sobre los montos de las cuantías de los siniestros de las bases de datos en mención. Sin embargo, el estudio con respecto a estos estadísticos no reportó los p valores para concluir el rechazo o la aceptación de la hipótesis nula bajo la cual los datos siguen alguna de las distribuciones en mención. De esta manera, las metodologías para la modelación de la pérdida agregada de la cuantías de los siniestros están evaluando fundamentalmente la posibilidad de realizar un ajuste de una distribución univariada sobre los conjuntos de datos de número de siniestros ocurridos así como de sus respectivas cuantías pero ignoran una nueva posibilidad para hacer la estimación cuando estos ajustes no son posibles de apoyar desde las evidencias estadísticas suficientes que proporcionan la validación de los sistemas de hipótesis clásicos de la bondad de ajuste de la inferencia estadística. Desde luego, esto representa un potencial riesgo para las compañías aseguradoras porque al adoptar estas metodologías, pueden incurrir en errores al momento de cobrar la tarifa para la prima del contrato de seguros y derivar luego en estados de insuficiencia de ingresos para cubrir gastos y siniestros, o en un evento de insolvencia.

Finalmente, el estudio de García, M (2014) mostró una metodología de modelación de pérdida agregada aplicada a siniestros de cáncer de mama. Bajo este enfoque se observó que el número de siniestros ocurridos es ajustado mediante un modelo Bayesiano Beta - Binomial y la severidad de los siniestros es modelada entre 2008 y 2011 año a año, mediante modelos clásicos de una distribución lognormal. El ajuste de esta distribución lognormal se valida mediante pruebas de hipótesis de bondad utilizando las pruebas Chi cuadrado, Kolmogorov - Smirnov y Anderson Darling. Aunque este y los demás aportes resultan valiosos, queda la importante pregunta de cómo hacer el tratamiento de una distribución multimodal que sea el resultado de cambios y desviaciones inesperadas en las distribuciones de frecuencia y severidad explicados por posibles modificaciones en los patrones que originan los siniestros. Desde luego estas implicaciones generan cambios sustanciales en las tarifas, así como una toma de decisiones frente a los riesgos asegurables porque él no considerarlo puede afectar la gestión de los riesgos de suscripción en

la compañía aseguradora de manera que eventos de siniestralidad con 'cuantías' muy grandes pueden afectar el capital o el patrimonio de las compañías.

De acuerdo con lo expuesto anteriormente, proponer modelos de enfoque clásico a partir de diagnósticos de cola ligera o pesada y componentes de mezcla, debería generar una mejor estimación de la pérdida agregada de los siniestros explicada por la variable aleatoria S. Sin embargo, esto puede ser insuficiente si se considera el tamaño de muestra de los siniestros. Rincón (2012) citó en [12] las ideas de Mowbray (1914) para afirmar que es necesario garantizar un número adecuado de periodos de suscripción anuales m con $m \in \{1, 2, 3, ...\}$ de manera que al observar respectivamente el monto de los siniestros $S_1, S_2, ..., S_m$ se tenga la convergencia de $\bar{S}$ al valor de $\mathbb{E}(\mathbb{S})$ por la ley de los números grandes. De esta manera, teniendo en cuenta el tamaño de muestra de los siniestros y el número de periodos de suscripción observados, se propuso una mejor estimación de $\mathbb{E}(\mathbb{S})$ ya no dada por $\bar{S}$ sino por:

$$\widehat{\mathbb{E}}(S) = Z \cdot \bar{S} + (1 - Z) \cdot \mathbb{E}(S), \tag{4}$$

para $0 \leq Z \leq 1$. En estos términos la suficiencia de datos para hacer las estimaciones del valor esperado del monto de pérdida agregada de los siniestros depende del coeficiente o factor de credibilidad Z. En consecuencia, Z toma el valor de 1 si la cantidad de datos es suficiente para estimar con una media el monto promedio de los siniestros y en ese caso, se dice que existe credibilidad total. En caso contrario, Z toma un valor inferior a 1 si la cantidad de datos no es suficiente para estimar con una media el monto promedio de los siniestros y se dice que existe credibilidad parcial.

Tse (2009) señala que tres maneras de abordar el problema de la suficiencia de datos son: Credibilidad Clásica, Credibilidad de Buhlmann y Credibilidad Bayesiana. En este sentido, bajo la credibilidad clásica si el número de siniestros N sigue una distribución Poisson con parámetro θ y la cuantía de los siniestros X sigue una distribución continua con soporte positivo de media μ y varianza σ^2, se dice que hay suficiencia de datos para el número de siniestros si este es mayor o igual que:

$$\theta_F = \frac{z^2_{1-\frac{\alpha}{2}}}{k^2},$$

donde $z_{1-\frac{\alpha}{2}}$ es el percentil de la distribución normal estándar tal que $\Phi(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$ y k es el margen de error de la media del número de siniestros tal que:

$$P(\mu_N - k\mu_N \leq N \leq \mu_N + k\mu_N) = 1 - \alpha$$

La suficiencia de datos para $\bar{X}$ se tiene si el total de datos de los siniestros es mayor o igual que $\theta_X = \theta_F \cdot \mathrm{CV}(X)^2$ en donde $\mathrm{CV}(X) = \frac{\sigma_x}{\mu_x}$. Finalmente la suficiencia de datos para un modelo de monto de pérdida agregada de siniestros o tasa pura de riesgo es $\theta_S = \theta_F + \theta_X$. Para complementar los resultados anteriores, se establece que el valor del factor de credibilidad Z para el número de siniestros es $Z_N = \sqrt{\frac{\mu_N}{\theta_F}}$, para el monto promedio de los siniestros $\bar{X}$ es $Z = \sqrt{\frac{N}{\theta_X}}$ y para el monto total de los siniestros S o la tasa pura de riesgo es $Z = \sqrt{\frac{\mu_N}{\theta_S}}$.

Según Bulhmann y Straub (1972), la credibilidad de Buhlmann se basa en un predictor lineal de $\widehat{S}_{n+1}$ como combinación lineal de n periodos de suscripción en donde:

$$\widehat{S}_{n+1} = \beta_0 + \beta_1 S_1 + ... + \beta_n S_n,$$

de manera que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_n)$ minimiza el error cuadrático medio definido como $\mathbb{E}[(S_{n+1} - \widehat{S}_{n+1})^2]$. En consecuencia, la estimación de S_{n+1} es:

$$\widehat{S}_{n+1} = Z \cdot \bar{S} + (1 - z) \cdot \mu_S,$$

donde $Z = \frac{n}{n+k}$ con $k = \frac{\mathbb{E}(\mathbb{V}\text{ar}(S|\Theta))}{\mathbb{V}\text{ar}(\mathbb{E}(S|\Theta))}$. Una problemática de la credibilidad de Buhlmann es el hecho de asumir que la variable S se encuentra idénticamente distribuida en m periodos de suscripción observados con riesgos expuestos diferentes. Comúnmente, estos riesgos son asegurados o tomadores en pólizas de seguros. Sin embargo, dependiendo del ramo, la exposición puede ser considerada como la suma asegurada o el tiempo de exposición. Ante el inconveniente mencionado anteriormente, surge una nueva propuesta como la credibilidad de Buhlmann-Straub. La estimación basada en el predictor lineal de $\hat{S_{n+1}}$ dado por $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_n)$ para minimizar el error cuadrático medio definido como $\mathbb{E}[(S_{n+1} - \widehat{S}_{n+1})^2]$ está definida de la siguiente manera:

$$\widehat{S}_{n+1} = Z \cdot \bar{S} + (1 - z) \cdot \mu_S,$$

con $Z = \frac{\sum_{i=1}^{m} m_i}{\sum_{i=1}^{m} m_i + k}$ donde m_i, es una medida de exposición del riesgo en el i-ésimo periodo de suscripción y $k = \frac{\mathbb{E}(\mathbb{V}\text{ar}(S|\Theta))}{\mathbb{V}\text{ar}(\mathbb{E}(S|\Theta))}$. Por otra parte, la credibilidad Bayesiana permite estimar con la media posterior, el monto o pérdida agregada de los siniestros, así como la tasa pura de riesgo por medio de la siguiente descomposición:

$$\mathbb{E}(\theta|S) = Z \cdot \bar{S} + (1 - Z)\mathbb{E}(\theta_{prior}),$$

donde θ_{prior} es la distribución previa de θ, es decir es una distribución que genera a θ. Citando a Tse (2009), en algunos casos, los factores de credibilidad Bayesiana corresponden a los mismos factores de credibilidad de Buhlmann y Buhlmann-Straub. Sin embargo, cuando la metodología Bayesiana pasa por la elicitación de los hiperparámetros de las distribuciones previas del número de siniestros y la cuantía de estos, el valor del parámetro k de la credibilidad de Buhlmann-Straub se encuentra en función de estas distribuciones previas y los factores de credibilidad de ambas propuestas metodológicas terminarán siendo equivalentes. En tal sentido, ante la dificultad que existe en la práctica para determinar $\mathbb{E}(\mathbb{V}\text{ar}(S|\Theta))$ y $\mathbb{V}\text{ar}(\mathbb{E}(S|\Theta))$, una credibilidad Bayesiana además de generalizar los resultados de la credibilidad de Buhlmann-Struab, resulta ser más apropiada.

Ante limitaciones que tienen los modelos clásicos como: la incertidumbre en el número de componentes que puede tener la variable aleatoria del monto de los siniestros y la suficiencia de datos entendida como el mínimo tamaño de muestra para la obtención de 'buenas' estimaciones; se propone el desarrollo de un modelo Bayesiano de mezcla con un número aleatorio de componentes que permita estimar la pérdida agregada del monto total de los siniestros así como cuantificar el riesgo de suscripción en seguros de no vida. El riesgo de suscripción es un proceso estocástico de ruina C_t definido en tiempo continuo con $t \geq 0$ de la siguiente manera:

$$C_t = \begin{cases} u & \text{si } t = 0 \\ u + P(t) - \sum_{i=1}^{N(t)} X_i & \text{si } t > 0 \end{cases}$$

donde u es el capital inicial de la compañía aseguradora, $P(t)$ es una función que permite determinar el total de primas recibidas hasta el tiempo t y X_i es la cuantía del i-ésimo siniestro para $i \in \{1, \ldots, N(t)\}$. $N(t)$ es un proceso de conteo que indica el número de siniestros ocurridos en un intervalo de tiempo $[0, t]$ para $t > 0$. En la literatura de la ciencia estadística, Lundberg (1903), definió $N(t)$ como un proceso homogéneo de Poisson con parámetro $\lambda > 0$ y Cramér (1930) formalizó esta idea con algunos conceptos de la teoría de los procesos estocásticos. De esta manera $N(t)$ es una variable aleatoria tal que:

$$N(t) \overset{\text{d}}{=} \text{Poisson}(\lambda t). \tag{5}$$

Sin embargo, una desventaja que presenta la definición de C_t es que resulta apresurado considerar la hipótesis de que $N(t)$ es un proceso homogéneo de Poisson porque en la práctica un conjunto de datos de conteo en una misma unidad de tiempo no necesariamente tiene una media estadísticamente igual a la varianza.

Se espera que, con esta propuesta metodológica en comparación con la contraparte clásica, se obtengan mejores estimaciones para frecuencia, severidad y pérdida agregada de manera que al realizar los procesos de tarifación se generen menores probabilidades de insolvencia para las aseguradoras y por lo tanto sus primas sean suficientes para cubrir gastos y siniestros lo cual permite generar márgenes de rentabilidad aptos para su crecimiento económico explicado por el proceso estocástico C_t. De esta manera, el principal aporte de la presente propuesta permite a los principales actores de la estadística actuarial seleccionar modelos de frecuencia y severidad robustos al momento de definir la tarifa generando así, una importante garantía dentro del proceso de elaboración de las notas técnicas que favorezcan un precio 'justo' en el proceso de suscripción de las diferentes pólizas de productos de no vida en una compañía aseguradora. La robustez de los modelos se explica en que, al realizar estudios de simulación, sensibilidad y validación cruzada $k-$fold para predicción de datos faltantes, se superan etapas importantes y necesarias que permiten apoyar el uso de un modelo para una decisión fundamental como lo es, el precio de un amparo o cobertura para contrato de seguros. Estas metodologías como $k-$fold que suelen ser usadas en campos como Machine Learning y su combinación con modelos Bayesianos permite el fortalecimiento de la estadística actuarial hasta el punto de que la rentabilidad de las compañías de seguros se ve favorecida ante la mitigación del riesgo de insuficiencia de primas, dado que este afecta las metas con relación al capital de una compañía aseguradora. Por otra parte, este tipo de modelamiento considera un número de componentes aleatorio en la cuantía de los siniestros, lo que de alguna manera permite responder al principio de incertidumbre que genera el número de clases latentes que existen en los riesgos de las cuantías que se han materializado. Finalmente, si las normativas vigentes exigen la mejor estimación para las reservas de siniestros y el monto de capital que deben constituir las compañías aseguradoras para su funcionamiento y para el respaldo de sus obligaciones ante eventualidades inesperadas, los modelos de la presente propuesta representan un importante insumo para las organizaciones no solamente desde el punto de vista de la tarifa más óptima, sino de las proyecciones de rentabilidad que sustenten el posible éxito de un nuevo producto así como del mejor contrato de reaseguros que puede escoger la compañía para amparar de manera adecuada los siniestros de 'altas' cuantías que se presenten en el futuro. Así mismo, el modelo tiene la 'capacidad' de analizar la probabilidad de insolvencia que tiene el negocio haciendo un monitoreo periódico que permita tomar decisiones tempranas frente al producto y a los riesgos asegurados. No menos importante que lo anterior, se encuentra el capital regulatorio de las compañías aseguradoras, el cual depende en gran medida de los pronósticos de la mejor estimación del monto de los siniestros y de la expectativa de crecimiento en apetitos de riesgo que tenga la compañía; de manera que una gestión basada en modelos como este permite analizar si los requerimientos de capital superan o no el patrimonio de la compañía en cada negocio y/o producto que se suscriba en el futuro.

2. Marco teórico y revisión de literatura

2.1. Conceptos básicos

Los conceptos básicos tenidos en cuenta para el presente marco teórico consideran un marco matemático y estadístico de: probabilidad, modelos de riesgo actuarial, teoría de la credibilidad, procesos estocásticos y estadística Bayesiana. En adición a lo anterior, se construye un marco de aplicación en el contexto de seguros que considera la tarifación y las clases de reaseguro.

2.1.1. Modelos de Riesgo Actuarial

Resulta de vital importancia para las compañías aseguradoras, estimar el monto agregado de pago de los siniestros en virtud de la constitución de sus reservas para siniestros. En términos financieros, el monto esperado de pago de los siniestros es junto con su reserva constituida, el pasivo de las aseguradoras. Ante la incertidumbre en la ocurrencia y la cuantía de los siniestros, el pasivo termina siendo observado al final de las vigencias de las pólizas del portafolio de asegurados; razón por la cual, las compañías deben realizar anticipadamente la mejor estimación del pasivo para promover la oportuna toma decisiones en contratos de reaseguros o en la cancelación de pólizas con cuantías de alto riesgo, entre otros, pues no realizar ninguna acción puede causar la insolvencia de las compañías de manera que no sea posible continuar con su actividad económica.

La mejor estimación del pasivo de las compañías aseguradoras termina siendo el reto más importante para las áreas de riesgo, de manera que encontrar la metodología más adecuada para su estimación es la piedra angular que permite la oportuna toma de decisiones para no entrar en el proceso de insolvencia. Los modelos de riesgo actuarial son una propuesta para realizar la estimación del monto total del pago de los siniestros. Existen dos modelos para realizar la estimación: El modelo individual y el modelo colectivo. Sin embargo, para efectos de estimar las trayectorias del proceso estocástico de riesgo que considera la ruina en una compañía aseguradora se relacionan a continuación únicamente los detalles del modelo de riesgo colectivo.

Modelo de riesgo colectivo. Según Rincón [12], si se dispone de un número no fijo de pólizas de seguros en un tiempo de vigencia $[0, t]$, de manera que N es el número de siniestros ocurridos y $X_1, X_2, ..., X_N$ las respectivas cuantías de sus pagos, entonces la variable aleatoria S también denominada 'riesgo' es la perdida agregada colectiva del monto de los siniestros y se define de la siguiente manera:

$$S = \sum_{i=1}^{N} X_i.$$

La media y la varianza de S se obtienen de la siguiente manera:

$$\mathbb{E}(S) = \mu_N \cdot \mu_X \ ; \ \mathbb{V}\text{ar}(S) = \sigma_x^2 \cdot \mu_N + \sigma_N^2 \cdot \mu_x^2$$

La fórmula de recursión de Panjer. De acuerdo con Rincón (2012) en su obra [12], para estimar las probabilidades $g_r = P(S = r)$ del modelo de riesgo colectivo, Panjer (1981) definió un método de recursión considerando que $p_k = P(N = k)$ y $f_r = P(X = r)$:

1. $g_0 = p_0$
2. $g_r = \sum_{i=1}^{r} \left(a + \frac{bi}{r}\right) f_i g_{r-i}$, para $r \geq 1$

En donde la variable aleatoria del número de siniestros ocurridos N es un proceso de Panjer, es decir que su función de masa de probabilidad p_k puede ser expresada de la siguiente manera:

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1},$$

para todo $k \geq 1$. Los valores de a y b dependen de la variable aleatoria discreta que se considere y se encuentran en función de sus parámetros. En particular las distribuciones: binomial, Poisson y binomial negativa son procesos de Panjer.

Los anteriores resultados muestran que la estimación de las probabilidades de las cuantías del monto agregado de los siniestros $g_r = P(S = r)$ dependen del proceso adecuado de discretización de la variable aleatoria continua del monto de los siniestros X. De la misma manera, el costo computacional de este algoritmo es alto cuando se observan cuantías muy altas en los siniestros.

Modelos clásicos para pérdida agregada. Tal y como lo proponen Barañano, De La Peña y Garayeta (2015), los modelos clásicos se basan en el ajuste de una distribución discreta para el número de siniestros ocurridos y una distribución continua de soporte positivo para el monto de los siniestros si se pretende estimar la pérdida colectiva del monto de los siniestros. Las pruebas de bondad de ajuste como Kolmogorov-Smirnov, Anderson Darling, Chi Cuadrado o Cramer Von Mises entre otros, resultan de gran relevancia al momento de escoger los modelos de distribución de probabilidad para estimar la frecuencia y la severidad de los siniestros.

Una vez escogidos los modelos de distribución de probabilidad para la frecuencia y la severidad de los siniestros, se utiliza una simulación de Monte Carlo con los siguientes pasos para estimar el valor de $\mathbb{E}(S)$:

1. Se define el número de iteraciones S de la simulación.
2. En cada iteración se simula un número aleatorio para la frecuencia N. Posteriormente, se simulan N números aleatorios con la distribución continua del monto de los siniestros X.
3. Con los resultados del paso anterior se calcula:

$$S = \sum_{i=1}^{N} X_i$$

4. Una vez terminadas las iteraciones de la Simulación de Montecarlo se calcula el valor de $\bar{S}$

Las posibles desventajas de estos modelos clásicos se describen a continuación:

1. Si se ajusta un modelo de mezcla o de mixturas para el monto de los siniestros, existe incertidumbre sobre el número de componentes que tiene el modelo. Este problema puede generar una subestimación de las tasas puras de riesgo con las cuales se tarifican los seguros aumentando la probabilidad de insolvencia de una compañía aseguradora. Así mismo, una sobrestimación de las tasas puras de riesgo puede generar una caída en las ventas de pólizas de seguro reduciendo así los márgenes de rentabilidad esperados para el crecimiento de una compañía aseguradora.
2. La estimación del pasivo de una compañía aseguradora es $\widehat{\mathbb{E}}(S)$. Sin embargo, $\bar{S}$ no es la mejor estimación de $\mathbb{E}(S)$ porque se debe evaluar si el tamaño de muestra de los siniestros es suficiente o no para realizar inferencia de $\mathbb{E}(S)$ a partir de $\bar{S}$.

2.1.2. Teoría de la credibilidad

La teoría de la credibilidad tiene sus orígenes en la pregunta fundamental de si el número de siniestros N observado en m periodos de suscripción, resulta suficiente o no para realizar estimaciones de buena calidad sobre: el número de siniestros que ocurrirán, el monto promedio de los siniestros que se observará en periodos posteriores de suscripción, el monto total de pago por concepto de siniestros y la prima pura

de riesgo que se debe tasar a cada tomador de póliza de seguros para cubrir en su totalidad el pago total de los siniestros. En consecuencia, algunos enfoques intentaron dar solución a la pregunta sobre la suficiencia de siniestros y entre los más conocidos se encuentran: credibilidad clásica, credibilidad de Buhlmann, credibilidad de Buhlmann-Straub y credibilidad Bayesiana. A continuación, se presentan algunos detalles de cada enfoque:

Credibilidad clásica. De acuerdo con las ideas expuestas en [12] por Rincón (2012), durante 1914, Mowbray en compañía de algunos actuarios norteamericanos utilizaron fundamentos de la inferencia estadística para abordar el problema de la credibilidad en un enfoque conocido como clásico. En razón de lo anterior, el enfoque también es conocido como credibilidad americana.

Aunque Rincón (2012), expone las ideas de la credibilidad clásica utilizando una distribución Poisson como modelo para el número de siniestros ocurridos, las siguientes ideas pueden ser implementadas en distribuciones de conteo como la Binomial Negativa, la Binomial o sus formas truncadas o cero modificadas.

Tamaño de muestra para los siniestros bajo credibilidad clásica. Típicamente este análisis se basa en el supuesto de que la frecuencia de los siniestros sigue una distribución Poisson tal y como lo presenta Tse (2009). Sin embargo, en la práctica suele suceder que la varianza de la frecuencia de los siniestros resulta ser mayor que su media, de manera que el análisis en mención se extiende para una variable aleatoria N con distribución binomial negativa de parámetros r, β denotada por:

$$N \overset{\mathrm{d}}{=} \text{Binomial Negativa}(r, \beta),$$

con $r > 0$, $\beta > 0$ y $n \in \{0, 1, 2, 3, ...\}$. La función de masa de probabilidad para N según Blanco(2005) está dada por:

$$p_x = \mathbb{P}\mathrm{r}(N = x) = \left(\begin{array}{c} x + r - 1 \\ x \end{array} \right) \left(\frac{1}{1+\beta} \right)^r \left(\frac{\beta}{1+\beta} \right)^x$$

Las propiedades de media y varianza para la variable aleatoria N se presentan a continuación:

$$\mu_N = \mathbb{E}(N) = r\beta \ ; \ \sigma_N^2 = \mathbb{V}\mathrm{ar}(N) = r\beta(1+\beta)$$

Bajo las ideas expuestas por Migon, H. et al. (2015), el problema para determinar la suficiencia de datos de los siniestros al momento de estimar su total en el siguiente periodo de suscripción se plantea por medio de la idea de intervalo de confianza para un margen de error k para la media con $k > 0$ y un nivel de confiabilidad $1 - \alpha$:

$$\mathbb{P}\mathrm{r}(\mu_N - k\mu_N < N < \mu_N + k\mu_N) = 1 - \alpha.$$

Si se tiene una muestra aletoria de N con un tamaño suficientemente 'grande', se satisface que:

$$n \geq \theta_F(1 + \beta), \tag{6}$$

donde n es el tamaño de muestra y $\theta_F = \left(\frac{z_{1-\frac{\alpha}{2}}}{k} \right)^2$. Esta última parametrización se utilizó para hacer comparación con el tamaño de muestra requerido en la inferencia del total de siniestros bajo una distribución Poisson pues este es equivalente a θ_F. Esto evidencia que, en una distribución binomial negativa, el tamaño de muestra mínimo se incrementa en un $100\beta\,\%$ con respecto al tamaño de muestra mínimo de una distribución Poisson.

Credibilidad de Buhlmann. Según Tse(2009), Philbrick (1981) expone los supuestos de la Credibilidad de Buhlmann:

- Se dispone de información de n periodos de suscripción en los cuales se observa el vector de pérdida agregada en el modelo colectivo y denotado como $S = \{S_1, S_2, ..., S_n\}$. Estas observaciones son independientes e idénticamente distribuidas bajo un vector de parámetros θ, el cual resulta de una realización del vector aleatorio Θ.

- La variable aleatoria S_{n+1} tiene la misma distribución de S. Su predicción está dada por:

$$S_{n+1} = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + ... + \beta_n S_n,$$

  donde $\boldsymbol{\beta} = \{\beta_0, \beta_1, ..., \beta_n\}$ minimiza el error cuadrático medio MSE definido como:

$$\text{MSE} = \mathbb{E}\left[(S_{n+1} - \widehat{S}_{n+1})^2\right].$$

La media condicional y la varianza condicional dado el vector de parámetros θ, se denotan de la siguiente manera: $\mathbb{E}(S \mid \theta) = \mu_S(\theta)$ y $\mathbb{V}\text{ar}(S \mid \theta) = \sigma_S^2(\theta)$. Si se define $\mathbb{E}[\mathbb{V}\text{ar}(S \mid \Theta)] = \mu_{PV}$ (media del proceso de varianza) y $\mathbb{V}\text{ar}[\mathbb{E}(S \mid \Theta)] = \sigma_{HM}^2$ (varianza de medias teóricas), la descomposición de la varianza de S es:

$$\mathbb{V}\text{ar}(S) = \mathbb{V}\text{ar}[\mathbb{E}(S \mid \Theta)] + \mathbb{E}[\mathbb{V}\text{ar}(S \mid \Theta)] = \sigma_{HM}^2 + \mu_{PV} \tag{7}$$

Definiendo $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_n)$, en Tse (2009) se demuestra que $\beta_0 = \mu_S - \mu_S \widehat{\boldsymbol{\beta}}_S^T \mathbb{1}$ y que $\widehat{\boldsymbol{\beta}}_S = \frac{1}{n+k}\mathbb{1}$ en donde $\mathbb{1}$ es un vector de tamaño $n \cdot 1$ y corresponde a un vector de unos. Por otra parte, $k = \frac{\mu_{PV}}{\sigma_{HM}^2}$. De esta manera la predicción para S_{n+1} es:

$$S_{n+1} = \frac{n}{n+k}\bar{S} + \frac{k}{n+k}\mu_S.$$

En consecuencia el factor de credibilidad Z para $\bar{S}$ es $Z = \frac{n}{n+k}$.

Credibilidad de Buhlmann-Straub. Bulhmann y Straub (1972), Philbrick (1981), Tse (2009) y Khapeava (2014) exponen los supuestos de la Credibilidad de Buhlmann-Straub de la siguiente manera:

- Sean n el número de periodos de suscripción observados, m_i la exposición en un periodo de suscripción i y S_i la pérdida agregada por unidad de exposición. Las variables $\{S_1, S_2, ..., S_n\}$ son independientes pero no idénticamente distribuidas, y dependen del vector de parámetros $\boldsymbol{\theta}$. El vector $\boldsymbol{\theta}$ es una realización del vector aleatorio Θ.

- Dado el vector de parámetros $\boldsymbol{\theta}$, la media y varianza condicional de S son:

$$\mathbb{E}(S_i \mid \boldsymbol{\theta}) = \mu_S(\boldsymbol{\theta}) \; ; \; \mathbb{V}\text{ar}(S_i \mid \boldsymbol{\theta}) = \frac{\sigma_S^2(\boldsymbol{\theta})}{m_i}$$

- La predicción de S_{n+1} está en función de los periodos anteriores:

$$S_{n+1} = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + ... + \beta_n S_n,$$

  de manera que $\boldsymbol{\beta} = \{\beta_0, \beta_1, ..., \beta_n\}$ minimiza el error cuadrático medio MSE con $\text{MSE} = \mathbb{E}[(S_{n+1} - \widehat{S}_{n+1})^2]$.

Los anteriores supuestos permiten deducir que $\mathbb{E}[\mathbb{V}\text{ar}(S_i \mid \Theta)] = \frac{\mu_{PV}}{m_i}$ y $\mathbb{V}\text{ar}[\mathbb{E}(S_i \mid \Theta)] = \sigma_{HM}^2$. En consecuencia, la descomposición de la varianza está dada por:

$$\mathbb{V}\text{ar}(S_i) = \mathbb{V}\text{ar}[\mathbb{E}(S_i \mid \Theta)] + \mathbb{E}[\mathbb{V}\text{ar}(S_i \mid \Theta)] = \sigma^2_{HM} + \frac{\mu_{PV}}{m_i} \tag{8}$$

En Tse(2009) se deduce que:

$$\widehat{\boldsymbol{\beta}}_S = \left(\frac{\sigma^2_{HM}}{\mu_{PV} + \sigma^2_{HM} \sum_{i=1}^n m_i} \right) \boldsymbol{m},$$

donde $\boldsymbol{m} = (m_1, m_2, ..., m_n)$ y $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_n)$. Por otra parte, $\widehat{\beta}_0 = \mu_S - \mu_S \widehat{\beta}_S^T \mathbb{1}$. En consecuencia, el pronóstico de $\widehat{S}_{n+1}$ es:

$$\widehat{S}_{n+1} = \frac{\sum_{i=1}^n m_i}{k + \sum_{i=1}^n m_i} \bar{S} + \frac{k}{k + \sum_{i=1}^n m_i} \mu_S,$$

donde el factor de credibilidad Z para la media $\bar{S}$, es $Z = \frac{\sum_{i=1}^n m_i}{k + \sum_{i=1}^n m_i}$ con $k = \frac{\mu_{PV}}{\sigma^2_{HM}}$.

Credibilidad Bayesiana. De acuerdo con las ideas expuestas por Rincón (2012) su obra Introducción a la Teoría del Riesgo [12], la perspectiva Bayesiana considera que para la pérdida agregada S del modelo colectivo tiene un conjunto de parámetros $\boldsymbol{\theta}$, el cual se considera como un vector aleatorio con una distribución de probabilidad $h(\boldsymbol{\theta} \mid \boldsymbol{\Theta})$ denominada previa o a priori. De esta manera, se considera $\boldsymbol{\theta}$ como una realización de Θ. En consecuencia, esta perspectiva permite incoporar el historial de los siniestros con información subjetiva o cuantitativa para estimar el valor de μ_S, de manera que con una combinación lineal convexa tal y como lo define Tse (2009) se deduce el pronóstico del valor esperado del monto agregado de la siguiente manera:

$$\mathbb{E}(S) = Z\bar{S} + (1 - Z)\mu_S$$

Con $0 < Z \leq 1$. El cálculo de la distribución posterior en estadística Bayesiana, permite asociar $\mathbb{E}(S)$ con la media posterior y μ_S con la media de la distribución previa o a priori. En el siguiente ejemplo, se observa la aplicación de la estadística Bayesiana para determinar los factores de credibilidad, así como el pronóstico del número de siniestros en un futuro periodo de suscripción.

El número de siniestros ocurridos N por asegurado tiene una distribución Binomial tal que:

$$N \mid \theta, m \overset{\text{d}}{=} \text{Binomial}(\theta, m).$$

En total se observaron n periodos de suscripción con m_i asegurados en el i-ésimo periodo de suscripción. El parámetro θ tiene una distribución beta tal que $\theta \mid \alpha, \beta \overset{\text{d}}{=} \text{Beta}(\alpha, \beta)$. Si N_i es el total de siniestros en el periodo i, es decir, $N_i = \sum_{j=1}^{m_i} X_j$, la distribución posterior es:

$$p(\theta \mid N, \alpha, \beta) \propto \theta^{\sum_{i=1}^m N_i + \alpha - 1} (1 - \theta)^{\sum_{i=1}^m m_i - \sum_{i=1}^m N_i + \beta - 1}.$$

Por consiguiente, $\theta \mid \text{resto} \overset{\text{d}}{=} \text{Beta}\left(\sum_{i=1}^m N_i + \alpha, \sum_{i=1}^m m_i - \sum_{i=1}^m N_i + \beta\right)$, donde resto denota a (N, α, β). La media posterior es:

$$\mathbb{E}(\theta \mid N, \alpha, \beta) = \frac{\sum_{i=1}^m m_i}{\sum_{i=1}^m m_i + \alpha + \beta} \cdot \bar{X} + \frac{\alpha + \beta}{\sum_{i=1}^m N_i + \alpha + \beta} \cdot \mathbb{E}(\theta \mid \alpha, \beta).$$

De esta manera, el factor de credibilidad Bayesiana para la media es $Z = \frac{\sum_{i=1}^{m} m_i}{\sum_{i=1}^{m} m_i + \alpha + \beta}$. En este caso los factores de credibilidad Bayesiana y Buhlmann-Straub son equivalentes. El cálculo de μ_{PV} está dado por:

$$\mu_{PV} = \mathbb{E}[Var(X \mid \theta)] = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}.$$

El cálculo de σ^2_{HM} es:

$$\sigma^2_{HM} = \mathbb{V}\text{ar}[\mathbb{E}(X \mid \theta)] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

De esta manera, el coeficiente k es tal que:

$$k = \frac{\mu_{PV}}{\sigma^2_{HM}} = \alpha + \beta.$$

Finalmente, el factor de credibilidad de Buhlmann-Straub para la media es:

$$Z = \frac{\sum_{i=1}^{m} m_i}{k + \sum_{i=1}^{m} m_i} = \frac{\sum_{i=1}^{m} m_i}{\alpha + \beta + \sum_{i=1}^{m} m_i},$$

y este resultado coindice con el factor de credibilidad Bayesiana.

2.1.3. Procesos Estocásticos

El modelo de riesgo de suscripción de una compañía aseguradora es un proceso estocástico que depende del capital de la compañía, las primas recibidas y el monto total de pago de siniestros ocurridos dentro de un intervalo de tiempo $[0, t]$. Sin embargo, previo a la definición de este modelo de riesgo, es necesario definir conceptos importantes que permiten su construcción de una manera natural, así como el entendimiento de sus más importantes propiedades.

Proceso Estocástico. Sea $(\Omega_2, \mathscr{F}_1)$ un espacio medible. Utilizando las ideas de Rincón (2012) en su obra [13], un proceso estocástico $\{X_t\}$ es un conjunto de variables aleatorias $\{X_t : t \in T\}$ definidas en un mismo espacio de probabilidad $(\Omega_1, \mathscr{F}_1, P)$ de manera que $X_t : \Omega_1 \to \Omega_2$, donde T se conoce como conjunto de índices y es numerable o no numerable. El conjunto Ω_2 es el conjunto de estados del proceso estocástico $\{X_t\}$. Si T es numerable, se dice que el proceso estocástico $\{X_t\}$ es discreto y si T es no numerable, se dice que el proceso estocástico $\{X_t\}$ es continuo.

Proceso homogéneo de Poisson. Para Rincón (2012), en su obra [13], un proceso estocástico de conteo $X(t)$ que indica el número de eventos de interés ocurridos en un intervalo de tiempo $[0, t]$ con $t \in \mathbb{R}$, es un proceso homogéneo de Poisson de parámetro λ con $\lambda > 0$ si y solo si $X(t)$ es tal que:

1. $X(0) = 0$
2. $X(t)$ tiene incrementos estacionarios e independientes
3. Para cualesquiera $s, t, i, j \geq 0$ con $t \geq s$ y $j \geq i$ se satisface que:

$$\mathbb{P}\text{r}(X(t) = j \mid X(s) = i) = \frac{e^{-\lambda(t-s)}\left[\lambda(t-s)\right]^{j-i}}{(j-i)!}.$$

Note que la última propiedad implica que la variable aleatoria de incrementos $X(t) - X(s)$ tiene distribución Poisson de parámetro $\lambda(t-s)$. Adicionalmente $X(t)$ tiene distribución Poisson de parámetro λt como consecuencia de que:

$$X(t) = X(t) - X(0) = X(t) \overset{\mathrm{d}}{=} \text{Poisson}(\lambda t)$$

Filtración. Sea $\{X_t\}$ un proceso estocástico con conjunto de índices T de manera que la variable aleatoria X_t está definida sobre el espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$ para todo $t \in T$. Según las ideas expuestas por Rincón (2012) en [13], una colección $\{\mathscr{F}_t\}_{t\in T}$ de $\sigma-$ álgebras de $\mathscr{F}$, es una filtración si y solo si $\mathscr{F}_s \subseteq \mathscr{F}_t$ para todo $s, t \in T$ con $s \leq t$.

Si el conjunto T está dado por $T = \{1, 2, \dots\}$, se dice que la filtración canónica se define como la colección $\{\mathscr{F}_t\}$ donde $t \in T$ y cada $\mathscr{F}_t$ es tal que:

$$\mathscr{F}_t = \sigma\{X_1, X_2, \dots X_t\}.$$

Si el conjunto T es no numerable y está dado por $T = [0, \infty)$, la filtración canónica se define como la colección $\{\mathscr{F}_t\}_{t\geq 0}$, donde cada $\mathscr{F}_t$ es tal que:

$$\mathscr{F}_t = \sigma\{X_s : 0 \leq s \leq t\}.$$

Proceso estocástico adaptado. Sea $\{X_t\}$ un proceso estocástico con conjunto de índices $T = \{1, 2, \dots\}$. En términos de Rincón (2012) en su obra [13], se dice que el proceso $\{X_t\}$ es adaptado a una filtración $\{\mathscr{F}_t : t \in T\}$ si y solo si la variable aleatoria X_t es $\mathscr{F}_t$-medible para todo $t \in T$.

Martingala. Sea $\{X_t\}$ un proceso estocástico con conjunto de índices $T = \{1, 2, \dots\}$. De acuerdo con Rincón(2012) en [13], se dice que el proceso $\{X_t\}$ es una martingala respecto a una filtración $\{\mathscr{F}_t : t \in T\}$ si y solo si:

1. $\mathbb{E}\,(\mid X_t \mid) < \infty$ para todo $t \in T$.
2. El proceso $\{X_t\}$ es adaptado a la filtración $\{\mathscr{F}_t : t \in T\}$.
3. Para todo $s \leq t$,
$$\mathbb{E}(X_t \mid \mathscr{F}_s) = X_s \tag{9}$$

Observación: Si en la ecuación anterior se satisface que $E(X_t \mid \mathscr{F}_s) > X_s$ se dice que el proceso estocástico $\{X_t\}$ es una submartingala. Por otra parte, si $E(X_t \mid \mathscr{F}_s) < X_s$, se dice que el proceso estocástico $\{X_t\}$ es una supermartingala.

Tiempo de paro. cSea τ una variable aleatoria con valores en $\{1, 2, ...\} \cup \{\infty\}$. En términos de Rincón (2012) en [13], se dice que τ es un tiempo de paro con respecto a la filtración $\{\mathscr{F}_t\}_{t\geq 1}$ si y solo si el evento A tal que $A : \tau \leq t$ satisface que $A \in \mathscr{F}_t$.

Note que si un evento de interés ocurre antes del tiempo t, se hace un 'paro' para observar dicho evento. Algunos eventos de interés en términos de riesgo son la ruina de una compañía aseguradora, el alcance de una utilidad superior a un umbral o el momento en el que ocurre un siniestro. Por otra parte, si el evento nunca es observable para $t \geq 1$, se dice que $\tau = \infty$.

Proceso estocástico de riesgo a tiempo continuo. Rincón (2012) en [12], define $\{C_t\}_{t\geq 0}$ como un proceso estocástico de riesgo a tiempo continuo de la siguiente manera:

$$C_t = u + c \cdot t - \textstyle\sum_{i=1}^{N(t)} X_i,$$

donde u es el capital de la compañía aseguradora con $u \geq 0$, c es la razón de cambio del crecimiento de las primas con respecto al tiempo para $c \in \mathbb{R}^+$, $N(t)$ es un proceso homogéneo de Poisson con parámetro $\lambda > 0$ que indica el número de siniestros ocurridos en el intervalo $[0, t]$ y X_i es la variable aleatoria continua de soporte positivo para el monto del $i-$ésimo siniestro para todo $i \in \{1, 2, \ldots, N(t)\}$. Este modelo de riesgo es conocido como el modelo de riesgo de Cramér-Lundberg. Inicialmente el modelo propuesto por Lundberg en 1903 fue formalizado con la teoría de los procesos estocásticos hacia 1930 por Lundberg.

El proceso estocástico de riesgo a tiempo continuo presenta una particularidad en el proceso de recaudo de primas para la compañía aseguradora, y que se encuentra dado por una función lineal $f(t)$ con $f(t) = c \cdot t$. Es común que la prima de un contrato de seguros por vigencia de un año sea recibida mediante cuotas mensuales. Esta modalidad de recaudo se conoce como prima fraccionada. De esta manera, si los contratos de seguros de la compañía se realizan por medio de esta modalidad esperando recibir un valor $k > 0$ durante un año, el valor de c es $c = \frac{k}{12}$. Sin embargo, este recaudo se realiza de esta manera en seguros de autos, o en ramos de vida, pero no en seguros de cumplimiento donde se hace el pago completo de la prima de manera anticipada al inicio del contrato de una obra con una entidad pública o privada.

En términos de los siniestros, el modelo de riesgo a tiempo continuo fue propuesto bajo la hipótesis de que el número de eventos ocurridos hasta el tiempo t sigue un proceso homogéneo de Poisson. Esto significa que en el tiempo $t > 0$, el número de siniestros ocurridos tiene una distribución Poisson de parámetro $\lambda \cdot t$ para algún $\lambda > 0$. Este hecho implica que el tiempo que transcurre entre dos siniestros consecutivos tiene una distribución exponencial de parámetro λ con valor esperado $\frac{1}{\lambda}$. De esta manera, definir la variable aleatoria T_k como el tiempo que transcurre entre el siniestro $k-1$ y el siniestro k, permite concluir que $T_k \overset{\text{d}}{=} \text{Exponencial}(\lambda)$. En consecuencia, la variable aleatoria W_k que mide el tiempo en el que ocurre el $k-$ésimo siniestro tiene la siguiente distribución:

$$W_k = \textstyle\sum_{j=1}^{k} T_j \overset{\text{d}}{=} \text{Gamma}(k, \lambda),$$

donde el conjunto de variables $T_1, \ldots T_k$ son independientes e idénticamente distribuidas. En la ciencia estadística, la variable aleatoria W_k con distribución Gamma presentada anteriormente es conocida como distribución Earlang.

Sin embargo, en la práctica el modelo de riesgo presentado anteriormente solo es aplicable al caso en el cual, el número de siniestros tiene una distribución Poisson de parámetro λ por semana o por mes citando estas medidas como ejemplo de unidades de tiempo para estudiar el total de eventos ocurridos. Por lo anterior, resulta bastante complejo aceptar en la práctica que el proceso de conteo de siniestros por unidad de tiempo tendrá una distribución con una varianza estadísticamente igual a su media para implementar esto proceso estocástico de riesgo a tiempo continuo. De esta manera, adaptar el modelo presentado al comportamiento de los datos resulta vital para cuantificar el riesgo en una forma eficiente, por lo que identificar la distribución discreta que mejor ajuste tiene al conteo de los siniestros es una tarea fundamental para el modelamiento.

Función de media y covarianza para el proceso estocástico de riesgo a tiempo continuo. Sea $\{C_t\}_{t \geq 0}$ un proceso estocástico de riesgo a tiempo continuo. Su función de media $\mu_C(t)$ está dada por:

$$\mu_C(t) = u + (c - \lambda \mu_X)t.$$

Para todo $s, t \geq 0$ con $t \geq s$, la función de covarianza $\gamma(t, s)$ del proceso estocástico de riesgo a tiempo continuo está dada por:

$$\gamma(t, s) = \mathbb{E}(X^2)\lambda t + \lambda^2 t^2 \mu_X^2 - \lambda^2 t s \mu_X^2.$$

Note que si $t = s$ se tiene que la función de varianza $\sigma_C^2(t)$ del proceso estocástico de riesgo a tiempo continuo es $\sigma_C^2(t) = \mathbb{E}(X^2)\lambda t$. Con estos resultados se pureba que el proceso estocástico $\{C_t\}$ no es

estacionario en el sentido débil porque su función de media y covarianza dependen del tiempo.

Por otra parte, la media del proceso estocástico es $u + (c - \lambda\mu_X)$, lo cual indica que el crecimiento del capital de la compañía aseguradora se da en forma lineal con una razón de cambio de $c - \lambda\mu_X$ con respecto al tiempo. La condición de ganancia y generación de utilidades para la aseguradora depende de que $c - \lambda\mu_X > 0$. Lo anterior implica que, para garantizar el crecimiento de la compañía, el valor de c debe estar acotado inferiormente por $\lambda\mu_X$, y este es el valor esperado del modelo de pérdida agregada del monto total de los siniestros.

El resultado presentado anteriormente indica que el valor de prima que se cobra a los tomadores para asegurar sus riesgos debe estar por encima del valor esperado del monto total de los siniestros:

$$c > \mathbb{E}(S).$$

De esta manera, la condición observada para la generación de ganancias y utilidades que impliquen el crecimiento del capital de una compañía aseguradora permite que los conceptos de prima pura de riesgo y prima comercial surjan de manera natural. La prima pura de riesgo se denota como p y se define como el valor esperado del monto total de los siniestros:

$$p = \mathbb{E}(S).$$

La interpretación de la prima pura de riesgo indica que, si se cobra el valor esperado del monto total de los siniestros a un conjunto de tomadores de póliza de seguros, no habrá ganancia para la compañía aseguradora y su capital inicial se mantendrá constante en el tiempo. Por otro lado, la prima comercial se define como la tasa pura de riesgo incrementada en un factor de recargo que permite generar utilidades en el resultado técnico de un producto de seguros para el crecimiento del capital inicial de la compañía aseguradora. La prima comercial es el valor que se cobra a un conjunto de tomadores de póliza en un contrato de seguros y se denota como c, de manera que c es tal que:

$$c = p \cdot (1 + \theta) > p = \mathbb{E}(S),$$

donde $\theta > 0$ es el factor de recargo que se aplica a la prima pura de riesgo.

Tiempo de ruina para el proceso estocástico de riesgo a tiempo continuo. Sea $\{C_t\}_{t\geq 0}$ un proceso estocástico de riesgo a tiempo continuo. El tiempo de ruina definido por Rincón (2012) en su obra [12] para el proceso estocástico se denota por τ y está dado por:

$$\tau = \text{ínf}\{t \geq 0 : C_t \leq 0\}.$$

La intepretación de τ indica que esta variable aleatoria muestra el primer instante en el cual se presenta la ruina de una compañía aseguradora. Si el conjunto $\{t \geq 0 : C_t \leq 0\}$ es vacío se tiene que la ruina nunca se presenta y en ese caso $\tau = \infty$.

Probabilidad de ruina con horizonte finito de un proceso de riesgo a tiempo continuo. Sean $\{C_t\}_{t\geq 0}$ un proceso estocástico de riesgo a tiempo continuo, u el capital de la aseguradora con $u \geq 0$ y τ la variable aleatoria del tiempo de ruina del proceso. La probabilidad de ruina de la compañía aseguradora definida por Rincón (2012) en [12] hasta el instante t con $t > 0$ se denota como $\psi(u, t)$ y se define de la siguiente manera:

$$\psi(u, t) = \mathbb{P}\text{r}(\tau \leq t \mid C_t = u).$$

Se dice que $\psi(u, t)$ se conoce también como la probabilidad de ruina de una compañía aseguradora en un horizonte finito t para un proceso de riesgo a tiempo continuo.

Probabilidad de ruina con horizonte infinito de un proceso de riesgo a tiempo continuo. Sean $\{C_t\}_{t\geq 0}$ un proceso estocástico de riesgo a tiempo continuo y $\psi(u,t)$ la función de probabilidad de ruina con horizonte finito con $\psi(u,t)$ se define en [12] por Rincón (2012) de manera que:

$$\psi : (\mathbb{R}^+ \cup \{0\}) \times \mathbb{R}^+ \to [0,1],$$

donde u es el capital de la compañía aseguradora y t el horizonte finito de la ruina. La probabilidad de ruina de la compañía en un horizonte infinito se denota por $\psi(u)$ y se define de la siguiente manera:

$$\psi(u) = \lim_{t\to\infty} \psi(u,t).$$

Algunas propiedades de la probabilidad de ruina con horizonte infinito para el modelo de riesgo a tiempo continuo. Estas propiedades son importantes porque permiten calcular la probabilidad de ruina de una aseguradora con un proceso estocástico de riesgo a tiempo continuo por medio de una fórmula de recursión. El desarrollo de la fórmula depende fundamentalmente de dos de las hipótesis del modelo como lo son: el crecimiento lineal del recaudo de primas por parte de la compañía y un proceso homogéneo de Poisson para ajustar la distribución del total de siniestros en un intervalo de tiempo $[0,t]$ con $t > 0$. El primer resultado importante se describe en seguida:

Sea $\{C_t\}_{t\geq 0}$ un proceso de riesgo a tiempo continuo con $C_t = u + ct - \sum_{i=1}^{N(t)} X_i$, donde $u \geq 0$, $c > 0$, $N(t)$ es un proceso homogéneo de Poisson con parámetro $\lambda > 0$ y $\{X_1, \dots X_{N(t)}\}$ es un conjunto de variables aleatorias continuas independientes e idénticamente distribuidas con función de densidad $f_X(x)$ absolutamente continua. La razón de cambio de la probabilidad de ruina en un horizonte infinito con respecto al cambio del capital de acuerdo con Rincón (2012) en [12] está dada por:

$$\frac{d\bar{\psi}(u)}{du} = \frac{\lambda}{c}\left[\bar{\psi}(u) - \int_0^u \bar{\psi}(u-x)\cdot f(x)dx\right],$$

donde $\bar{\psi}(u)$ es la probabilidad de que la compañía no tenga ruina en un horizonte infinito cuando su capital inicial es u. En otras palabras $\bar{\psi}(u) = 1 - \psi(u)$. Los detalles de la demostración del resultado anterior se aprecian en Rincón (2012). Sus pasos deben seguirse para calcular de forma analítica la razón de cambio de interés cuando el modelo de riesgo a tiempo continuo presenta cambios frente a la distribución discreta que permite contar el número de siniestros ocurridos hasta el tiempo t, debido a que esto cambiaría la distribución de la variable aleatoria T_1. Por otra parte, si la modalidad de recaudo de las primas de la compañía no es lineal, también se generan cambios en el cálculo de la razón de cambio para el valor de c.

Una aplicación que tiene el resultado de la razón de cambio para la probabilidad de ruina en un horizonte infinito con respecto al cambio del capital de la compañía es la siguiente fórmula de recursión que permite obtener la probabilidad $\psi(u)$ bajo los supuestos del modelo de riesgo de Cramér-Lundberg:

$$\psi(u) = \begin{cases} \frac{\lambda}{c}\mathbb{E}(X) & \text{,si } u = 0 \\ \frac{\lambda}{c}\left[\int_u^\infty \mathbb{S}(x)dx - \int_0^u \psi(u-x)\mathbb{S}(x)dx\right] & \text{,si } u > 0, \end{cases}$$

donde $\mathbb{S}(x)$ es la función de sobrevivencia de la variable aleatoria X de la cuantía de los siniestros.

2.1.4. Estadística Bayesiana

La estadística Bayesiana juega un papel fundamental en el estudio de un proceso estocástico de riesgo a tiempo continuo para cuantificar el riesgo de suscripción en una compañía aseguradora. Una nueva adaptación del modelo de riesgo de Cramér-Lundberg, permite definir el proceso $\{C_t\}_{t\geq 0}$ de la siguiente manera:

$$C_t = \begin{cases} u & \text{si } t = 0 \\ u + P(t) - \sum_{i=1}^{N(t)} X_i & \text{si } t \geq 0, \end{cases}$$

donde $P(t)$ es una función de recaudo de primas que depende del tiempo t pero no necesariamente es lineal en la práctica. El resultado de $P(t)$ no depende únicamente de la modalidad de recaudo de la aseguradora, y es necesario considerar factores adicionales tales como: la producción o emisión de pólizas, el ramo de seguros de interés, pólizas aceptadas en coaseguro, cancelación y renovación de pólizas, la constitución de reservas para primas que exige la Superintendencia Financiera de Colombia y la probabilidad de impago de la prima por parte del tomador del contrato de seguros o del intermediario que participa en la negociación. En consecuencia, la complejidad de la función $P(t)$ requiere inclusive del estudio de modelos de series de tiempo para la producción (ventas) o de modelos predictivos para estudiar la incertidumbre en la cancelación de las pólizas o el impago de la prima en un contrato de seguros.

El entendimiento de las variables aleatorias $N(t)$ como un proceso de conteo de los siniestros ocurridos en el intervalo $[0, t]$ con $t > 0$, y X_i como la cuantía del $i-$ésimo siniestro permite realizar una óptima tarifación de seguros por medio de la técnica de credibilidad bayesiana. Esta tarifación explica en gran medida el comportamiento de la función $P(t)$ y el proceso estocástico de riesgo a tiempo continuo C_t.

Para implementar la técnica de credibilidad bayesiana en el proceso de tarifación de un contrato de seguros para el amparo o cobertura de una contingencia en particular, es necesario partir del siguiente principio: las variables aleatorias independientes $N(t)$ y X que representan el número de eventos de siniestro ocurridos para el amparo en un intervalo $[0, t]$ y la cuantía del siniestro, respectivamente, se definen como:

$$N(t) \mid \boldsymbol{\lambda} \overset{\text{d}}{=} p\left[N(t) \mid \boldsymbol{\lambda}\right] \ ; \ X \mid \boldsymbol{\theta} \overset{\text{d}}{=} p(X \mid \boldsymbol{\theta}),$$

donde la dupla de vectores de parámetros $\boldsymbol{\lambda}, \boldsymbol{\theta}$ son una realización de los vectores aleatorios $\boldsymbol{\Lambda}$ y $\boldsymbol{\Theta}$ respectivamente:

$$\boldsymbol{\lambda} \mid \boldsymbol{\Lambda} \overset{\text{d}}{=} p(\boldsymbol{\lambda} \mid \boldsymbol{\Lambda}) \ ; \ \boldsymbol{\theta} \mid \boldsymbol{\Theta} \overset{\text{d}}{=} p(\theta \mid \boldsymbol{\Theta}).$$

Los conceptos mencionados se formalizarán a continuación por medio de las definiciones de distribución previa, distribución muestral y distribución posterior.

Principios de estadística Bayesiana.

El interés de aprender sobre un vector de parámetros $\boldsymbol{\theta}$ que generan un conjunto de observaciones $\mathbb{X} = (X_1, \ldots, X_n)$, de manera que $\boldsymbol{\theta}$ es una realización del vector aleatorio Θ; permite que la estadística Bayesiana surja de manera natural:

- Distribución previa o priori: Es un conjunto de información subjetiva o cuantitativa de creencias que se tiene sobre el vector de parámetros $\boldsymbol{\theta}$. Si $\boldsymbol{\theta}$ es una realización del vector aleatorio $\boldsymbol{\Theta}$, se define la distribución previa de la siguiente manera:

  $$\boldsymbol{\theta} \mid \boldsymbol{\Theta} \overset{\text{d}}{=} p(\boldsymbol{\theta} \mid \boldsymbol{\Theta}).$$

  Por notación, se conviene que: $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\Theta})$.

- Distribución muestral: Es la función de verosimilitud para la muestra aleatoria $\mathbb{X} = (X_1, \ldots, X_n)$ dado su vector de parámetros $\boldsymbol{\theta}$:

  $$\mathbb{X} \mid \boldsymbol{\theta} \overset{\text{d}}{=} p(\mathbb{X} \mid \boldsymbol{\theta}).$$

- Distribución posterior: La aplicación del teorema de Bayes permite la obtención de la distribución posterior del vector de parámetros denotada por $p(\boldsymbol{\theta} \mid \mathbb{X})$:

$$p(\boldsymbol{\theta} \mid \mathbb{X}) \propto p(\mathbb{X} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}). \tag{10}$$

  Observe que en el resultado final se establece la proporcionalidad entre la distribución posterior y el producto entre la función de verosimilitud y la distribución previa dado que $\frac{1}{p(\mathbb{X})}$ es constante. El valor de $p(\mathbb{X})$ es conocido como constante de normalización.

  Si el vector de parámetros $\boldsymbol{\theta}$ es tal que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, se dice que $p(\theta_j \mid \text{resto})$ es la distribución condicional completa para θ_j con $j \in \{1, \ldots, m\}$, con resto $= \{\mathbb{X}, \boldsymbol{\theta}_{-i}\}$ de manera que el vector $\boldsymbol{\theta}_{-i}$ de $m-1$ componentes no contiene a θ_j. Se dice que el vector $\boldsymbol{\theta}$ es el conjunto de parámetros de interés, mientras que el vector $\boldsymbol{\Theta}$ representa el conjunto de hiperparámetros en el modelo Bayesiano.

- Distribución previa conjugada: Se dice que una distribución muestral $p(\mathbb{X} \mid \boldsymbol{\theta})$ tiene una distribución previa o a priori $p(\boldsymbol{\theta})$ conjugada si y solo si su distribución posterior $p(\boldsymbol{\theta} \mid \mathbb{X})$ tiene la misma clase de distribución que la previa.

Distribución predictiva. La distribución predictiva se denota por $p(x \mid \mathbb{X})$ y permite realizar el pronóstico de una observación futura x dado que se conoce la información histórica del conjunto de observaciones $\mathbb{X} = (X_1, \ldots, X_n)$. El resultado de $p(x \mid \mathbb{X})$ está dado por:

$$p(x \mid \mathbb{X}) = \int_{\boldsymbol{\theta}} p(x, \boldsymbol{\theta} \mid \mathbb{X}) d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(x \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbb{X}) d\boldsymbol{\theta}.$$

Observe que la integral obtenida en el paso anterior no necesariamente puede ser resulta mediante métodos analíticos en todos los casos.

Valores p predictivos y bondad de ajuste. Sea $\mathbb{X} = (X_1, \ldots, X_n)$ una muestra aleatoria con $X_i \mid \boldsymbol{\theta} \overset{\text{iid}}{=} p(X_i \mid \boldsymbol{\theta})$ para todo $i \in \{1, \ldots, n\}$, con vector de parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ y vector de hiperparámetros $\boldsymbol{\Theta}$. El valor p predictivo se denota por $p_{\mathbb{X},\boldsymbol{\theta}}$ y se define de la siguiente manera:

$$p_{\mathbb{X},\boldsymbol{\theta}} = \mathbb{P}\text{r}(T(\mathbb{X}^*, \boldsymbol{\theta}) \geq T(\mathbb{X}, \boldsymbol{\theta})),$$

donde $\mathbb{X}^*$ es un conjunto de observaciones futuras con la misma cardinalidad del conjunto de observaciones $\mathbb{X}$ y T es un estadístico que depende de una muestra aleatoria o un conjunto de observaciones y de su vector de parámetros. De esta manera, T es tal que

$$T : (\mathbb{X}, \boldsymbol{\theta}) \to \mathbb{R}.$$

Si T es un estadístico, algunos ejemplos asociados a T son: la media, la desviación estándar, la mediana. un cuantil y el coeficiente de variación, entre otros. Se dice que el modelo Bayesiano para el conjunto de observaciones $\mathbb{X}$ con vector de parámetros $\boldsymbol{\theta}$ y vector de hiperparámetros $\boldsymbol{\Theta}$ tiene un buen ajuste con respecto al estadístico T si y solo si su valor p predictivo $p_{\mathbb{X},\boldsymbol{\theta}}$ no se concentra alrededor de valores extremos como cero o uno. En términos de Gelman, A. et al. (2014), esto significa que un buen ajuste para un modelo Bayesiano alcanza un valor p predictivo óptimo si se encuentra alrededor de $0,5$.

Criterios Bayesianos DIC y WAIC. Sea $\mathbb{X} = (X_1, \ldots, X_n)$ una muestra aleatoria con $X_i \mid \boldsymbol{\theta} \overset{\text{iid}}{=} p(X_i \mid \boldsymbol{\theta})$ para todo $i \in \{1, \ldots, n\}$, con vector de parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ y vector de hiperparámetros $\boldsymbol{\Theta}$.

Para Gelman, A. et al. (2014), el DIC conocido como criterio de información de la devianza se calcula de la siguiente manera:

$$\text{DIC} = -2\log\left[p(\mathbb{X} \mid \boldsymbol{\theta} = \mathbb{E}(\boldsymbol{\theta} \mid \mathbb{X}))\right] + 2p_{\text{DIC}},$$

donde p_{DIC} es tal que:

$$p_{\text{DIC}} = 2\left[\log p(\mathbb{X} \mid \boldsymbol{\theta} = \mathbb{E}(\boldsymbol{\theta} \mid \mathbb{X}) - \mathbb{E}_{\text{posterior}}(\log p(\mathbb{X} \mid \boldsymbol{\theta}))\right].$$

Observe que $\log p(\mathbb{X} \mid \boldsymbol{\theta} = \mathbb{E}(\boldsymbol{\theta} \mid \mathbb{X}))$ es la función de log verosimilitud evaluada en el conjunto de observaciones $\mathbb{X}$ cuando cada componente del vector de parámetros es la esperanza de la distribución condicional completa de su respectiva componente:

$$\log p(\mathbb{X} \mid \boldsymbol{\theta} = \mathbb{E}(\boldsymbol{\theta} \mid \mathbb{X})) = \textstyle\sum_{i=1}^{n} \log p(x_i \mid \mathbb{E}(\boldsymbol{\theta} \mid \mathbb{X})).$$

En la práctica si se realizan K iteraciones del algoritmo MCMC con $K \in \mathbb{Z}^+$ para obtener muestras de $\boldsymbol{\theta}$ sujeto a $p(\boldsymbol{\theta} \mid \mathbb{X})$ y se combina con alguna aplicación del algoritmo de Metrópolis o Metrópolis-Hastings dado que la distribución condicional completa de algún parámetro θ_i no es conocida con $i \in \{1, \ldots, m\}$, se tiene que:

$$\mathbb{E}(\theta_j \mid \text{resto}) = \frac{1}{K} \textstyle\sum_{k=1}^{K} \theta_j^{(k)},$$

para todo $j \in \{1, \ldots, m\}$. De esta manera:

$$\mathbb{E}(\mathbb{X} \mid \boldsymbol{\theta}) = \left(\mathbb{E}(\theta_1 \mid \text{resto}), \ldots, \mathbb{E}(\theta_m \mid \text{resto})\right).$$

Por otra parte, el valor $\mathbb{E}_{\text{posterior}}(\log p(\mathbb{X} \mid \boldsymbol{\theta})$ es tal que:

$$\mathbb{E}_{\text{posterior}}(\log p(\mathbb{X} \mid \boldsymbol{\theta})) = \frac{1}{K} \textstyle\sum_{k=1}^{K} \log p\left(\mathbb{X} \mid \boldsymbol{\theta}^{(k)}\right).$$

Por otra parte, el criterio WAIC para Gelman, A. et al. (2014) conocido como criterio de información Bayesiano Watanabe - Akaike se define como:

$$\text{WAIC} = 2 \textstyle\sum_{i=1}^{n} \left[\log(\mathbb{E}_{\text{posterior}} p(y_i \mid \boldsymbol{\theta})) - \mathbb{E}_{\text{posterior}}(\log p(y_i \mid \boldsymbol{\theta}))\right],$$

donde

$$\log(\mathbb{E}_{\text{posterior}} p(y_i \mid \boldsymbol{\theta})) = \log\left[\frac{1}{K} \textstyle\sum_{k=1}^{K} p\left(y_i \mid \boldsymbol{\theta}^{(k)}\right)\right]$$
$$\mathbb{E}_{\text{posterior}}(\log p(y_i \mid \boldsymbol{\theta})) = \frac{1}{K} \textstyle\sum_{k=1}^{K} \log p\left(y_i \mid \boldsymbol{\theta}^{(k)}\right).$$

Cuando se tienen al menos dos propuestas $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_s\}$ para el vector de hiperparámetros $\boldsymbol{\Theta}$, se dice que el mejor modelo Bayesiano para el conjunto de datos $\mathbb{X}$ con vector de parámetros $\boldsymbol{\theta}$ y vector de hiperparámetros $\boldsymbol{\Theta}_B$ con respecto al criterio DIC es aquel que satisface que:

$$\text{DIC}_{\boldsymbol{\Theta}_B} = \text{mín}\{\text{DIC}_{\boldsymbol{\Theta}_1}, \ldots, \text{DIC}_{\boldsymbol{\Theta}_s}\}.$$

De manera análoga, el mejor modelo Bayesiano para el conjunto de datos $\mathbb{X}$ con vector de parámetros $\boldsymbol{\theta}$ y vector de hiperparámetros $\boldsymbol{\Theta}_M$ con respecto al criterio WAIC es aquel que satisface que:

$$\text{WAIC}_{\boldsymbol{\Theta}_M} = \text{mín}\{\text{WAIC}_{\boldsymbol{\Theta}_1}, \ldots, \text{WAIC}_{\boldsymbol{\Theta}_s}\},$$

donde $\text{DIC}_{\boldsymbol{\Theta}_i}$ y $\text{WAIC}_{\boldsymbol{\Theta}_i}$ representan el valor de los criterios Bayesianos DIC Y WAIC respectivamente para el modelo Bayesiano con vector de hiperparámetros $\boldsymbol{\Theta}_i$ para todo $i \in \{1, \ldots, s\}$.

2.1.5. Conceptos de Seguros

Prima pura de riesgo. La prima pura de riesgo en el proceso de tarificación de seguros depende de la tasa que se cobra por el monto del valor asegurado y el tiempo de exposición, de manera que permite cumplir con todas las obligaciones de los pagos de los siniestros. De esta manera, la prima pura de riesgo que se cobra a cada asegurado permite determinar el valor de $\mathbb{E}(S)$.

La prima pura de riesgo se define en seguros como el producto entre la frecuencia y la severidad. La frecuencia se entiende como el número esperado de siniestros en un nivel de exposición dado por el número de asegurados o tomadores, el tiempo de cobertura del contrato de seguros o la suma asegurada. La severidad es el monto esperado de la cuantía de los siniestros. De esta manera la tasa pura de riesgo P es:

$$P = \text{Frecuencia} \cdot \text{Severidad} = \frac{\mathbb{E}(N)}{\text{Exposición}} \cdot \mathbb{E}(\bar{X}), \tag{11}$$

donde N es el número de siniestros ocurridos, X es la cuantía de los siniestros. Además, se cuenta con una medida de exposición para el riesgo asegurado (p. ej. tiempo de cobertura del seguro, valor asegurado o número de asegurados).

Prima comercial. La tasa comercial C depende de la tasa pura de riesgo P de manera que

$$C = P(1 + \theta),$$

con θ tal que $\theta > 0$. El valor θ es un factor de recargo que tiene en cuenta los gastos y la rentabilidad esperada del negocio asegurador. Fasecolda (Federación de Aseguradores Colombianos) define el factor de recargo θ de la siguiente manera:

$$1 + \theta = \frac{1}{1 - GA - GC - RU},$$

donde GA corresponde al porcentaje de gastos de gestión administrativa, GC corresponde al porcentaje de gastos por gestión de comercial y RU es el porcentaje de utilidad. De esta manera, la tasa comercial C es:

$$C = \frac{P}{1 - GA - GC - RU}.$$

La prima comercial PC, es el valor que se cobra al tomador de la póliza por el riesgo asegurado. Fasecolda indica que el cálculo de la prima comercial está dado de la siguiente manera:

$$PC = \frac{SA \cdot C \cdot \text{tiempo de vigencia}}{365}, \tag{12}$$

donde SA es el valor asegurado, C es la tasa comercial expresada en porcentaje. Además, se cuenta con el tiempo de vigencia o de cobertura de los amparos de la póliza de seguros.

Reaseguro. Para eliminar un porcentaje de la pérdida causada por el pago del monto total de los siniestros, las compañías aseguradoras utilizan reaseguros. Estos reaseguros son conocidos en la práctica como Cuota Parte o de Exceso de Pérdida. A continuación, se precisan algunos detalles de estos conceptos.

Reaseguro Proporcional o Cuota Parte. Para Rincón (2012) en su obra [12], el reaseguro Cuota Parte es conocido como reaseguro proporcional. Según Rincón (2012), la función que define el reaseguro proporcional es h tal que:

$$h(x) = ax,$$

donde $h : [0, \infty) \to \mathbb{R}$ y $a \in (0, 1)$. El coeficiente a es conocido en la práctica como el porcentaje de retención para la compañía aseguradora y $1 - a$ como el porcentaje de cesión a la entidad reaseguradora. En consecuencia, si $S = \sum_{i=1}^{N} Y_i$ es el pago total de los siniestros, el monto SA retenido por la compañía aseguradora es:

$$\text{SA} = h(\textstyle\sum_{i=1}^{N} Y_i) = \sum_{i=1}^{N} aY_i.$$

El pago total de los siniestros cedido SR a la reaseguradora es:

$$\text{SR} = \textstyle\sum_{i=1}^{N}(1 - a)Y_i.$$

Reaseguro de Exceso de Pérdida o No proporcional. Este reaseguro entra en operación cuando el monto de un siniestro supera un valor $M > 0$ pactado en un contrato entre la aseguradora y el reasegurador. Según Rincón (2012) en [12], la función que define el reaseguro no proporcional es:

$$h(x) = \min\{x, M\}.$$

En consecuencia se define SA como el monto retenido por la aseguradora como

$$\text{SA} = \min\{S, M\}.$$

De esta manera el monto cedido a la reaseguradora es $\text{SR} = S - \text{SA}$ y presenta dos casos:

- Si $\min\{S, M\} = S$ entonces $\text{SR} = 0$
- Si $\min\{S, M\} = M$ entonces $\text{SR} = S - M$

2.2. Modelo de mezcla completamente Bayesiano para la Severidad de los Siniestros

La cuantificación del riesgo de suscripción en una compañía aseguradora de capital $u \geq 0$ depende del proceso estocástico de riesgo a tiempo continuo $\{C(t)\}_{t\geq 0}$ dado por:

$$C_t = \begin{cases} u & , \text{si } t = 0 \\ u + P(t) - \sum_{i=1}^{N(t)} X_i & , \text{si } t \geq 0, \end{cases}$$

donde $P(t)$ es una variable aleatoria que permite determinar el recaudo total de primas hasta el tiempo t, $N(t)$ es la variable aleatoria del proceso de conteo de siniestros ocurridos en el intervalo de tiempo $[0, t]$ con $t > 0$ y la muestra aleaoria $\{X_1, \ldots, X_{N(t)}\}$ es tal que X_i es la variable aleatoria de la cuantía del $i-$ésimo siniestro con $X_i \overset{\text{iid}}{=} p(X_i \mid \boldsymbol{\theta})$ para todo $i \in \{1, \ldots, N(t)\}$ con $\boldsymbol{\theta}$ como vector de parámetros.

De esta manera, interesa modelar el proceso de conteo $N(t)$ y la cuantía de los siniestros para cuantificar la medida de riesgo de la manera óptima. Con respecto al proceso de conteo $N(t)$, es importante definir unidades de tiempo donde se observa el total de siniestros ocurridos. La observación se suele hacer de manera mensual porque es así como se hace monitoreo y seguimiento al proceso estocástico $\{C_t\}$ para analizar los estados financieros en una compañía aseguradora. La variable aleatoria del número de siniestros ocurridos que se observa de manera mensual es $B_t = N_t - N_{t-1}$ para todo $t \in \{1, 2, \ldots\}$ donde

$N(0) = 0$. Si las variables aleatorias $B_1, B_2, \dots B_t$ son independientes e idénticamente distribuidas, el modelo Bayesiano se define de la siguiente manera:

$$\mathbb{B} \mid \lambda \overset{\text{iid}}{=} p(\mathbb{B} \mid \boldsymbol{\lambda}) \; ; \; \boldsymbol{\lambda} \mid \boldsymbol{\Lambda} \overset{\text{d}}{=} p(\boldsymbol{\lambda} \mid \boldsymbol{\Lambda}),$$

donde $\mathbb{B} = \{B_1, \dots, B_t\}$, $\boldsymbol{\lambda}$ es el vector de parámetros y $\boldsymbol{\Lambda}$ es el vector de hiperparámetros de $\boldsymbol{\lambda}$. Observe que $p(\mathbb{B} \mid \boldsymbol{\lambda})$ depende de la distribución discreta más apropiada para el número de siniestros ocurridos por mes. En el modelo de riesgo de Cramér-Lundberg, se asume que esta distribución es Poisson, sin embargo, esto depende de la naturaleza del conjunto de datos $\mathbb{B}$ porque no necesariamente en todos los casos se cuenta con la igualdad estadística entre la media y la varianza de los datos de conteo de los siniestros.

Por otra parte, la cuantía de los siniestros es otro aspecto importante para considerar en la cuantificación del riesgo de suscripción para una compañía aseguradora. Es por esto, que escoger de manera apresurada una variable aleatoria continua de soporte positivo para modelar la cuantía de los siniestros puede ser inapropiado para el estudio del proceso estocástico $\{C_t\}$ en una compañía aseguradora. El principal problema en el ajuste de un modelo de distribución es asumir que los riesgos asegurados son homogéneos y así subestimar las probabilidades de ocurrencia de siniestros con una severidad o cuantía muy alta. Esta clase de cuantías tan altas pueden generar la insolvencia de la compañía aseguradora en el largo plazo. Por lo anterior, hacer una clasificación apropiada de los riesgos asegurados y sus cuantías más probables es uno de los retos más grandes que tienen hoy en día las compañías aseguradoras al momento de realizar la suscripción de pólizas porque de ello depende el crecimiento de su capital.

Si en un portafolio de asegurados con coberturas para automóviles hay un grupo diverso de personas para conducir según su edad, género, región de residencia y marca de auto, entre otros; la cuantía de un siniestro para dos autos con un conductor del mismo género, edad y marca pero que residen en zonas de accidentalidad baja y alta, puede ser muy diferente. En ese sentido, se observa que la cuantía de los siniestros depende de un rasgo o variable latente que permite clasificar los riesgos según su perfil. En consecuencia, los riesgos asegurados no son homogéneos y existen grupos para clasificarlos de manera que: asegurados de un mismo grupo son homogéneos y asegurados de diferentes grupos son heterogéneos. Estos grupos o clústeres pueden ser interpretados como tipos de riesgo que para un ejemplo pueden ser: bajo, medio y alto. Sin embargo, asumir que siempre se tendrán tres grupos de clasificación de riesgo puede llevar a una cuantificación equívoca del riesgo de suscripción.

El modelo de distribución de probabilidad que mejor se ajusta para la cuantía de los siniestros según las condiciones mencionadas anteriormente, es un modelo de mezcla que tiene un número de componentes sobre el cual existe incertidumbre porque no se sabe a priori cuantos grupos de clasificación de riesgos existen para un portafolio de asegurados. La distribución muestral para la cuantía x_i del $i-$ésimo siniestro con $i \in \{1, \dots, N(t)\}$ está dada por un modelo de mezcla de H componentes con $H \in \{2, 3, \dots\}$, definido de la siguiente manera:

$$f(x_i \mid \theta) = \textstyle\sum_{h=1}^{H} W_h p_h(x_i \mid \boldsymbol{\theta}_h),$$

donde las $p_h(x_i \mid \boldsymbol{\theta}_h)$ son las componentes de la mezcla, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H)$ con $\boldsymbol{\theta}_h$ como el vector de parámetros para la componente h y los W_h se conocen como pesos de manera que $W_h > 0$ y $\sum_{h=1}^{H} W_h = 1$ para todo $h \in \{1, \dots H\}$.

Observe que $p_h(x_i \mid \boldsymbol{\theta}_h)$ es la función de densidad de una variable aleatoria con soporte continuo para la cuantía de los siniestros de riesgos asegurados que se clasifican en el h-ésimo grupo con $h \in \{1, \dots, H\}$. De esta manera, con el modelo de mezcla se garantiza que las cuantías de los riesgos asegurados dependen de los grupos en los cuales se encuentran clasificados los riesgos, de manera que cuantías en un mismo grupo son homogéneas y cuantías de diferentes grupos resultan ser heterogéneas.

La interpretación de los pesos de la mezcla sugiere que: no hay grupos de riesgo sin un número de riesgos asegurados y la proporción de asegurados en el $h-$ésimo grupo se representa con el valor W_h para todo

$h \in \{1, \ldots, H\}$. En la práctica cuando se presentan tres tipos de riesgo como bajo, medio y alto, se suele observar que la proporción de riesgo alto es la más baja pero merece especial atención porque las cuantías de sus riesgos asegurados puede generar la insolvencia de la compañía aseguradora.

El modelo de mezcla es consecuencia de la aplicación del teorema de probabilidad total tal y como se presenta a continuación: Sean H grupos de riesgos en un portafolio de asegurados de un ramo de autos para una cobertura de sustracción con $H \in \mathbb{Z}^+$ y $H \geq 2$. Si la cuantía de los siniestros del $h-$ésimo grupo es una variable aleatoria continua X_h de soporte positivo tal que $X_h \mid \boldsymbol{\theta}_h \overset{\text{iid}}{=} p(X_h \mid \boldsymbol{\theta}_h)$ y las proporciones del los grupos son $W_1, \ldots, W_h$ con $W_h > 0$ para todo $h \in \{1, \ldots, H\}$, la probabilidad de ocurrencia del evento A donde la cuantía del siniestro es al menos el capital de la aseguradora u con $u \geq 0$, está dada por:

$$\mathbb{P}\text{r}(A) = \mathbb{P}\text{r}(X > u) = \textstyle\sum_{h=1}^{H} W_h \mathbb{S}_h(X_h = u \mid \boldsymbol{\theta}_h),$$

donde $\mathbb{S}_h$ es la función de sobrevivencia de la variable aleatoria X_h. Al observar $N(t)$ siniestros en el intervalo $[0, t]$ con $N(t) = n$ y $t > 0$, el modelo de mezcla completamente Bayesiano para la severidad de los siniestros se puede expresar jerárquicamente de la siguiente manera:

$$\begin{aligned}
x_i \mid z_i &\overset{\text{iid}}{=} p_h(x_i \mid \boldsymbol{\theta}_{z_i}) \\
z_i \mid \omega &\overset{\text{d}}{=} \text{Multinomial}(\omega) \\
\omega \mid \boldsymbol{\alpha} &\overset{\text{d}}{=} \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\theta}_h \mid \boldsymbol{\phi} &\overset{\text{d}}{=} p(\boldsymbol{\theta}_h \mid \boldsymbol{\phi}) \\
\boldsymbol{\phi} &\overset{\text{d}}{=} p(\boldsymbol{\phi}),
\end{aligned}$$

para todo $i \in \{1, \ldots, n\}$, donde los $z_i \in \{1, 2, ..., H\}$ son variables latentes o grupos de riesgos asegurados que indican con cual componente de la mezcla está asociado el siniestro de cuantía x_i, el vector $\omega = \{w_1, ..., w_H\}$ contiene los pesos de la mezcla y ϕ denota otros posibles parámetros de los cuales depende $\boldsymbol{\theta}$.

La distribución muestral y la distribución previa del modelo presentado anteriormente se describen a continuación:

1. Distribución muestral: La distribución de la cuantía de los siniestros corresponde a un modelo de mezcla de H componentes con distribuciones de variables aleatorias continuas de soporte positivo, para $H \in \mathbb{Z}^+$ y $H \geq 2$. De esta manera si se observan n siniestros de variables aleatorias $X_1, \ldots, X_n$ idénticamente distribuidas en el intervalo $[0, t]$ con $t > 0$, la función de verosimilitud está dada por:

   $$\mathbb{P}\text{r}(X_1 = x_1, \ldots, X_n = x_n) = \textstyle\prod_{i=1}^{n} \left[\sum_{h=1}^{H} \omega_h p_h(x_i \mid \boldsymbol{\theta}_h) \right].$$

   Sin embargo, el resultado anterior es un reto analítico complejo con un elevado costo computacional al momento de calcular las distribuciones condicionales completas del modelo Bayesiano. Por tanto, se propone una función de verosimilitud de manera que al dejar fija una clase latente h se eliminen los términos del modelo de mezcla sobre la sumatoria cuando los siniestros asociados no pertenecen a la clase h. La verosimilitud propuesta es:

   $$\mathbb{P}\text{r}(X_1 = x_1, \ldots, X_n = x_n) = \textstyle\prod_{h=1}^{H} \prod_{i=1}^{n} \left[\omega_h p_h(x_i \mid \boldsymbol{\theta}_h) \right]^{z_{ih}},$$

   donde z_{ih} es una variable aleatoria indicadora para la observación i. Esta función indicadora toma el valor de uno si se asocia con la clase h o cero en otro caso:

$$z_{ih} = \begin{cases} 1 & \text{, si } i \in h \\ 0 & \text{, si } i \notin h. \end{cases}$$

Si existen H componentes, entonces cada observación x_i tiene una probabilidad de pertenecer a cada una de las clases latentes o grupos de clasificación de riesgos asegurados. De esta manera, la distribución más adecuada para encontrar la clase latente de cada observación i con $i \in \{1, \ldots, n\}$ es multinomial. Note que el total de riesgos n_h de la clase h está dado por:

$$n_h = \textstyle\sum_{i=1}^{n} z_{ih},$$

para todo $h \in \{1, \ldots, H\}$. De esta manera, la función de densidad conjunta es:

$$\Pr(N_1 = n_1, \ldots, N_H = n_H) = \frac{n!}{\prod_{h=1}^{H} n_h!} \prod_{i=1}^{n} \prod_{h=1}^{H} \omega_h^{z_i h} = \frac{n!}{\prod_{h=1}^{H} n_h!} \prod_{h=1}^{H} \omega_h^{n_h},$$

donde N_h es la variable aleatoria que indica el número de riesgos asociados a la clase h con $h \in \{1, \ldots .H\}$ y $N_h \overset{\text{iid}}{=} \text{Multinomial}(n, \omega_1, \ldots, \omega_h)$. Observe que cuando existen únicamente dos clases de riesgos asegurados, la distribución anterior es binomial.

2. Distribución previa: Como se definen dos distribuciones muestrales, una para la cuantía de los siniestros y otra para las clases latentes de cada observación, entonces es necesario proponer dos distribuciones previas de manera que el vector de parámetros $\boldsymbol{\theta}_h$ de la cuantía sea una realización del vector aleatorio $\boldsymbol{\phi}$ y el vector ω de las clases latentes sea una realización del vector aleatorio $\boldsymbol{\alpha}$. La mejor propuesta para cada distribución previa en este caso es una conjugada porque facilita el cómputo y el cálculo analítico de la distribución condicional completa de cada parámetro. Sin embargo como la distribución muestral de la cuantía de los siniestros aún no se encuentra definida, se dice que su previa es $p(\boldsymbol{\theta}_h \mid \boldsymbol{\phi})$, mientras que para distribución muestral $z_i \mid \text{Multinomial}(1, \omega_1, \ldots, \omega_H)$ su previa conjugada $p(\boldsymbol{\omega})$ es Dirichlet (generalización multivariante de una distribución Beta):

$$p(\boldsymbol{\omega} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{h=1}^{H} \alpha_h)}{\prod_{h=1}^{H} \Gamma(\alpha_h)} \prod_{h=1}^{H} \omega_h^{\alpha_h - 1},$$

donde $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_H)$. Observe que si solo existen dos grupos de riesgos asegurados, $\boldsymbol{\omega} \mid \boldsymbol{\alpha}$ se distribuye Beta con vector de parámetros $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$.

2.2.1. Modelo de mezcla completamente Bayesiano con distribución lognormal y número de componentes H fijo.

Según la Normativa Europea de Solvencia II (EIOPA, 2014), una posible propuesta para la cuantía de los siniestros es la distribución lognormal para cada una de las h componentes del modelo de mezcla con $h \in \{1, 2, 3, \ldots, H\}$. En este primer estudio de caso, el número de componentes H que permiten clasificar los riesgos en un portafolio de asegurados, es un número entero fijo tal que $H \geq 2$. Sin embargo, es importante considerar la incertidumbre de H en un escenario aleatorio porque siempre existirán dudas frente al número de grupos que permiten clasificar los riesgos. El modelamiento jerárquico para la variable aleatoria de la cuantía de los siniestros X se plantea de la siguiente manera con un modelo de mezcla de H componentes:

$$x_i \mid z_i \overset{\text{iid}}{=} \text{lognormal}(x_i \mid \theta_h, \sigma_h^2)$$
$$z_i \mid w \overset{\text{iid}}{=} \text{Multinomial}(1; \omega_1, \omega_2, \ldots, \omega_H)$$

$$\begin{aligned}\boldsymbol{\omega} &\overset{\mathrm{d}}{=} \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_H)\\ \theta_h &\overset{\mathrm{d}}{=} \text{Normal}(\mu_h, \tau_h^2)\\ \sigma_h^2 &\overset{\mathrm{d}}{=} \text{Inv-Gamma}(a_h, b_h),\end{aligned} \tag{13}$$

donde $z_i, h \in \{1, \ldots, H\}$ y $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_H)$. Para el modelo presentado anteriormente, solo es necesario describir la previa conjugada para θ_h y σ_h^2 en virtud de que en la sección anterior se describieron las distribuciones muestrales y la distribución previa conjugada para una distribución multinomial. En primera instancia, es razonable proponer que θ_h es una realización de una variable aleatoria con distribución normal porque $-\infty < \theta_h < \infty$. Así mismo resulta adecuado proponer que σ_h^2 sea una realización de una variable aleatoria con distribución inversa gamma porque $\sigma_h^2 > 0$. Por otra parte, es trivial comprobar que una normal y una inversa gamma como previas respectivas para θ_h y σ_h^2 son conjugadas.

Distribuciones condicionales completas. Sea $X_1, \ldots, X_n$ una muestra aleatoria de la cuantía de siniestros observados en un intervalo de tiempo $[0, t]$ con $t > 0$ y $X_i \overset{\text{iid}}{=} p(X_i \mid \theta_h, \sigma_h^2)$. La distribución muestral $p(X_i \mid \theta_h, \sigma_h^2)$ se encuentra condicionada a un modelo de mezcla completamente Bayesiano como el presentado en la ecuación (13) con número de componentes H fijo y $H \geq 2$. De esta manera, si $\boldsymbol{\theta}$ es un vector de parámetros tal que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H, \sigma_1^2, \ldots, \sigma_H^2, \omega, Z_1, \ldots, Z_n)$, entonces el vector de hiperparámetros $\boldsymbol{\Theta}$ es:

$$\boldsymbol{\Theta} = (\alpha, \mu_1, \ldots, \mu_H, \tau_1^2, \ldots, \tau_H^2, a_1, \ldots, a_h, b_1, \ldots, b_h).$$

En consecuencia, si $\mathbb{X} = (X_1, \ldots, X_n)$ la distribución posterior $p(\boldsymbol{\theta} \mid \mathbb{X}, \boldsymbol{\Theta})$ es:

$$p(\boldsymbol{\theta} \mid \mathbb{X}, \boldsymbol{\Theta}) \propto p(x_i \mid z_i) \cdot p(z_i \mid \boldsymbol{\omega}) \cdot \prod_{h=1}^{H} \omega_h^{\alpha-1} \cdot \prod_{h=1}^{H} p(\theta_h \mid \mu_h, \tau_h^2) \cdot \prod_{h=1}^{H} p(\sigma_h^2 \mid a_h, b_h)$$

donde $p(x_i \mid z_i)$ y $p(z_i \mid \boldsymbol{\omega})$ son distribuciones muestrales tales que:

$$\begin{aligned}p(x_i \mid z_i) &= \prod_{i=1}^{n} \prod_{h=1}^{H} \left[\text{lognormal}(X_i \mid \theta_h, \sigma_h^2)\right]^{Z_{ih}}\\ p(z_i \mid \boldsymbol{\omega}) &\propto \prod_{i=1}^{n} \prod_{h=1}^{H} \omega_h^{z_{ih}}.\end{aligned}$$

Distribución condicional completa para z_i. La distribución condicional completa $p(Z_i = h \mid \text{resto})$, con i fijo y $h \in \{1, \ldots, H\}$, es:

$$p(Z_i = h \mid \text{resto}) = \frac{\omega_h \text{lognormal}(X_i \mid \theta_h, \sigma_h^2)}{\sum_{l=1}^{H} \omega_l \text{lognormal}(X_i \mid \theta_l, \sigma_l^2)}, \tag{14}$$

donde resto es tal que:

$$\text{resto} = (\alpha, \mu_1, \ldots, \mu_H, \tau_1^2, \ldots, \tau_H^2, a_1, \ldots, a_H, b_1, \ldots, b_H, \boldsymbol{\omega}, \mathbb{X}).$$

Observe que el cómputo de la distribución condicional completa de la ecuación (14) consiste en realizar el cociente entre la densidad de la componente h evaluada en la cuantía del siniestro de la $i-$ésima observación y la función de densidad del modelo de mezcla evaluada en el mismo punto.

Distribución condicional completa para $\boldsymbol{\omega}$. La distribución condicional completa $p(\boldsymbol{\omega} \mid \text{resto})$ es:

$$p(\boldsymbol{\omega} \mid \text{resto}) \propto \prod_{h=1}^{H} \omega^{n_h + \alpha_h - 1},$$

donde resto se define como:

$$\text{resto} = (\alpha, \mu_1, \ldots, \mu_H, \tau_1^2, \ldots, \tau_H^2, a_1, \ldots, a_H, b_1, \ldots, b_H, \mathbb{X}, \mathbb{Z}).$$

En consecuencia, se deduce que $\boldsymbol{\omega} \mid \text{resto} \overset{\text{d}}{=} \text{Dirichlet}(n_1 + \alpha_1, \ldots, n_H + \alpha_H)$.

Distribución condicional completa para $\boldsymbol{\theta_h}$. Si h es fijo con $h \in \{1, \ldots, H\}$, entonces la distribución condicional completa $p(\theta_h \mid \text{resto})$ para θ_h es:

$$p(\theta_h \mid \text{resto}) \propto \exp\left[-\frac{1}{2}\left\{\left(\frac{n_h}{\sigma_h^2} + \frac{1}{\tau_h^2}\right)\theta_h^2 - 2\left(\frac{\sum_{i:Z_i \in h}\log(X_i)}{\sigma_h^2} + \frac{\mu_h}{\tau_h^2}\right)\right\}\right], \tag{15}$$

donde $\text{resto} = (\alpha, \tau_1^2, \ldots, \tau_H^2, a_1, \ldots, a_H, b_1, \ldots, b_H, \boldsymbol{\omega}, \mathbb{X}, \mathbb{Z})$.

Por tanto, se deduce que $\theta_h \mid \text{resto} \overset{\text{d}}{=} \text{Normal}(\widehat{m}\widehat{V}^{-1}, \widehat{V}^{-1})$, con n_h como el total de riesgos asegurados en la clase latente h, y $\widehat{m}, \widehat{V}$ tales que:

$$\widehat{m} = \frac{\sum_{i:Z_i \in h}\log(X_i)}{\sigma_h^2} + \frac{\mu_h}{\tau_h^2} \; ; \; \widehat{V} = \frac{n_h}{\sigma_h^2} + \frac{1}{\tau_h^2}.$$

Distribución condicional completa para $\boldsymbol{\sigma_h^2}$. Si h es fijo con $h \in \{1, \ldots, H\}$, entonces la distribución condicional completa $p(\sigma_h^2 \mid \text{resto})$ de σ_h^2 es:

$$p(\sigma_h^2 \mid \text{resto}) \propto (\sigma_h^2)^{\frac{n_h}{2}+a_h+1} \cdot \exp\left[-\left(\frac{1}{2}\sum_{i:Z_i \in h}(\log(X_i) - \theta_h)^2 + b_h\right)/\sigma_h^2\right], \tag{16}$$

donde $\text{resto} = (\alpha, \theta_1, \ldots, \theta_h, a_1, \ldots, a_H, b_1, \ldots, b_H, \boldsymbol{\omega}, \mathbb{X}, \mathbb{Z})$.

En consecuencia de lo anterior, $\sigma_h^2 \mid \text{resto} \overset{\text{d}}{=} \text{Inv-Gamma}\left(\frac{n_h}{2} + a_h, \frac{1}{2}\sum_{i:Z_i \in h}(\log(X_i) - \theta_h)^2 + b_h\right)$ con n_h como el total de riesgos asegurados en la clase latente h.

2.2.2. Modelo de mezcla completamente Bayesiano con distribución lognormal y número de componentes aleatorio entre 2 y H.

Uno de los problemas que presenta el modelo propuesto en la sección anterior es que el número H de componentes de la mezcla es fijo. En la práctica el número de componentes depende de la incertidumbre y la aleatoriedad de los riesgos asegurados. La naturaleza de estos riesgos, así como las condiciones del ramo de seguros no debe asociarse con un número fijo de componentes porque esto puede derivar en la subestimación de las cuantías más severas de siniestros que se encuentran concentradas en los grupos de riesgo más alto y pueden generar insolvencia en la compañía aseguradora. En la presente sección se propone un modelo de mezcla completamente Bayesiano para la severidad de los siniestros con un número aleatorio de componentes entre 2 y H. Este último supuesto resulta de vital importancia porque se requieren como mínimo dos componentes para la aplicación de un modelo de mezcla. En caso de que todos los grupos se fusionen en un mismo componente, se puede ajustar una distribución lognormal para modelar la cuantía de los siniestros.

Basados en el modelo de mezcla completamente Bayesiano con número de componentes H fijo presentado en (13), según Gelman, A. et al (2014) se realiza el siguiente ajuste para tener un número de componentes aleatorio entre 2 y H con $H \in \mathbb{Z}$ y $H \geq 2$:

$$x_i \mid z_i \overset{\text{iid}}{=} \text{lognormal}(x_i \mid \theta_h, \sigma_h^2)$$
$$z_i \mid w \overset{\text{iid}}{=} \text{Multinomial}(1; \omega_1, \omega_2, \ldots, \omega_H)$$

$$\begin{aligned}
\omega_h &= \frac{\lambda_h}{\sum_{l=1}^{H} \lambda_l} \\
\theta_h &\overset{\mathrm{d}}{=} \text{Normal}(\mu_h, \tau_h^2) \\
\sigma_h^2 &\overset{\mathrm{d}}{=} \text{Inv-Gamma}(a_h, b_h) \\
\lambda_h &\overset{\mathrm{d}}{=} \text{Gamma}\left(\frac{n_0}{H}, 1\right),
\end{aligned} \tag{17}$$

para todo $i \in \{1, \ldots, n\}$ y $h \in \{1, \ldots, H\}$ con $\lambda_h > 0$.

En la propuesta del modelo con número de componentes H fijo, el vector del probabilidades $\boldsymbol{\omega}$ tiene una distribución previa Dirichlet tal que $\boldsymbol{\omega} \mid \boldsymbol{\alpha} \overset{\mathrm{d}}{=} \text{Dirichlet}(\alpha_1, \ldots, \alpha_H)$, con $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_H)$ y $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_H)$. De esta manera:

$$\mathbb{E}(\boldsymbol{\omega} \mid \boldsymbol{\alpha}) = \left(\frac{\alpha_1}{\sum_{h=1}^{H} \alpha_h}, \ldots, \frac{\alpha_H}{\sum_{h=1}^{H} \alpha_h} \right).$$

En consecuencia, resulta natural definir ω_h tal que $\omega_h = \frac{\lambda_h}{\sum_{l=1}^{H} \lambda_l}$ para todo $h \in \{1, \ldots, H\}$. Observe que la propuesta de este modelo consiste en definir la variable aleatoria ω_h como su valor esperado cuando su distribución previa es Dirichlet, haciendo $\lambda_h = \alpha_h$. Por otra parte, si $\alpha = \alpha_1 = \cdots = \alpha_H$, se puede definir α como un valor inversamente proporcional al número de grupos H:

$$\alpha = \frac{n_0}{H},$$

donde n_0 es una constante. Observe que si $n_0 = 1$, entonces $\alpha = \frac{1}{H}$. En estos términos, bajo la hipótesis donde $\lambda_h = \alpha_h = \alpha$, se tiene que λ_h es realización de una variable aleatoria que depende de $\frac{n_0}{H}$ de modo que:

$$\mathbb{E}(\lambda_h) = \frac{n_0}{H},$$

para todo $h \in \{1, \ldots, H\}$. De esta manera, una propuesta natural al hecho presentado anteriormente es:

$$\lambda_h \overset{\mathrm{d}}{=} \text{Gamma}\left(\frac{n_0}{H}, 1\right).$$

2.3. Aplicación de la metodología propuesta para modelo de mezcla completamente Bayesiano con número de componentes aleatorio.

Una vez propuestos m modelos $\mathscr{M}_1, \ldots, \mathscr{M}_m$ para la frecuencia $\mathbb{N}(t)$ y modelos de mezcla con número de componentes aleatorio completamente Bayesianos para la severidad $\mathbb{X}$ de los siniestros ocurridos en un intervalo de tiempo $[0, t]$ con las técnicas mencionadas en la sección anterior, se verifican: la bondad de ajuste por medio de los valores p predictivos, los criterios Bayesianos WAIC y DIC, y la calidad del pronóstico por medio de la técnica k-fold, CV (*k - fold cross validation*) utilizada por Gelman, A. et al. (2014). A continuación, se describe el procedimiento para realizar la verificación de los ítems mencionados anteriormente:

1. Calcular los valores p predictivos $p_{\mathbb{N},(\boldsymbol{\lambda}),T}$ y $p_{\mathbb{X},(\boldsymbol{\theta}),T}$ para cada modelo Bayesiano $\mathscr{M}_i$ con $i \in \{1, \ldots, m\}$, escogiendo como estadística T: la media, la mediana, la varianza y el coeficiente de variación.

2. Calcular los criterios Bayesianos: WAIC($\mathscr{M}_i$) y DIC($\mathscr{M}_i$) para todo modelo candidato $\mathscr{M}_i$ con $i \in \{1, \ldots, m\}$.

3. Implementar la técnica de validación cruzada de $k-$fold para cada modelo $\mathbb{M}_i$ con $i \in \{1, \ldots, m\}$ por medio de los siguientes pasos:

   - Generar k particiones de igual tamaño mediante un mecanismo de muestreo para: el número de siniestros ocurridos con observaciones $j \in \{1, \ldots, p\}$ y de la cuantía de los siniestros con observaciones $i \in \{1, \ldots, n\}$. Seguido de lo anterior, se evalúa la función de verosimilitud en las observaciones de la $v-$ésima partición con $v \in \{1, \ldots, k\}$, muestreando S valores de la distribución posterior de los vectores de parámetros $\boldsymbol{\lambda}$ y $\boldsymbol{\theta}$ utilizando las observaciones de las restantes $v-1$ particiones para calcular el indicador $\text{lppd}_{\text{k-CV}}$ expuesto en el siguiente paso, tanto para frecuencia como severidad, respectivamente. Los vectores $\boldsymbol{\theta}^{is}$ y $\boldsymbol{\lambda}^{js}$ denotan los valores muestreados de la distribución posterior cuando se excluyen respectivamente las observaciones de las particiones i, j en la $s-$ésima iteración con $s \in \{1, \ldots, S\}$ e $i, j \in \{1, \ldots, k\}$.
   - Calcular el indicador predictivo de validación cruzada $\text{lppd}_{\text{k-fold,CV}}$ definido de la siguiente manera para la frecuencia $\mathbb{N}$ y la severidad $\mathbb{X}$:

$$\text{lppd}_{\text{k-fold,CV},\mathbb{N}} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{p_i}\log\left(\frac{1}{S}\sum_{s=1}^{S} p(n_j \mid \boldsymbol{\lambda}^{is})\right)$$
$$\text{lppd}_{\text{k-fold,CV},\mathbb{X}} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log\left(\frac{1}{S}\sum_{s=1}^{S} p(x_i \mid \boldsymbol{\theta}^{is})\right),$$

     donde p_i, n_i son los tamaños de muestra en la $i-$ésima partición para las observaciones de frecuencia y severidad respectivamente con $i \in \{1, \ldots, k\}$. Note que $\sum_{i=1}^{k} p_i = p$ y $\sum_{i=1}^{k} n_i = n$.
   - Calcular el error cuadrático medio CME y el error porcentual absoluto medio MAPE en las observaciones de la $v-$ésima partición con $v \in \{1, \ldots, k\}$ generando p_v observaciones $\widehat{n}_j$ y n_v observaciones $\widehat{x}_j$ por medio de sus respectivas distribuciones predictivas:

$$\text{CME}_{v,\mathbb{N}} = \frac{1}{p_v}\sum_{j=1}^{p_v}(n_j - \widehat{n}_j)^2 \; ; \; \text{CME}_{v,\mathbb{X}} = \frac{1}{n_v}\sum_{i=1}^{n_v}(x_i - \widehat{x}_j)^2$$
$$\text{MAPE}_{v,\mathbb{N}} = \frac{1}{p_v}\sum_{j=1}^{p_v}\left|\frac{n_j - \widehat{n}_j}{n_j}\right| \; ; \; \text{MAPE}_{v,\mathbb{X}} = \frac{1}{n_v}\sum_{i=1}^{n_v}\left|\frac{x_i - \widehat{x}_i}{x_i}\right|.$$

   - Calcular los valores medios de CME y MAPE en la muestra de test evaluando los siguientes indicadores:

$$\text{CME}_{\text{test},\mathbb{N}} = \frac{1}{k}\sum_{v=1}^{k}\text{CME}_{v,\mathbb{N}} \; ; \; \text{MAPE}_{\text{test},\mathbb{N}} = \frac{1}{k}\sum_{v=1}^{k}\text{MAPE}_{v,\mathbb{N}}$$
$$\text{CME}_{\text{test},\mathbb{X}} = \frac{1}{k}\sum_{v=1}^{k}\text{CME}_{v,\mathbb{X}} \; ; \; \text{MAPE}_{\text{test},\mathbb{X}} = \frac{1}{k}\sum_{v=1}^{k}\text{MAPE}_{v,\mathbb{X}}.$$

Los tres pasos mencionados anteriormente permiten identificar el mejor modelo Bayesiano para predecir en el intervalo de tiempo $[0, t]$, el monto de pérdida agregada de los siniestros denotado por $S(t)$. Para la metodología implementada se utilizará un número de $k = 5$-fold.

Función de recaudo de primas. El mejor modelo Bayesiano $\mathscr{M}_k$ escogido para predecir el monto total de los siniestros hasta el tiempo t permite calcular la función de recaudo de primas $P(t)$ porque esta depende del monto que cobra a los tomadores por asegurar sus riesgos. Es decir, que la función $P(t)$ depende del proceso de tarifación en el cual se realiza el cálculo de la tasa pura de riesgo y la tasa comercial. La tasación de seguros de no vida se realiza utilizando la técnica de credibilidad Bayesiana:

$$\widehat{\mathbb{E}}(S) = Z \cdot \bar{S} + (1 - Z) \cdot \mathbb{E}(S \mid \boldsymbol{\theta}_{\text{post}}, \boldsymbol{\lambda}_{\text{post}}),$$

donde Z es el factor de credibilidad, S es tal que $S = \sum_{i=1}^{N(t)} X_i$, $N(t)$ es el proceso de conteo de siniestros, X_i es la cuantía del $i-$ésimo siniestro y $\boldsymbol{\theta}_{\text{post}}, \boldsymbol{\lambda}_{\text{post}}$ son los vectores de parámetros muestreados de las distribuciones posteriores $p(\lambda \mid \mathbb{N})$ y $p(\theta \mid \mathbb{X})$. Suponiendo que se tienen S_{MCMC} muestras $\boldsymbol{\lambda}^{(s)}$ y $\boldsymbol{\theta}^{(s)}$ de $p(\lambda \mid \mathbb{N})$ y $p(\theta \mid \mathbb{X})$ con $s \in \{1, \ldots, S_{\text{MCMC}}\}$, el proceso para la obtención de $\widehat{\mathbb{E}}(S)$ se describe a continuación:

1. Para cada $s \in \{1, \ldots, S_{\text{MCMC}}\}$, guardar $\mathbb{V}\text{ar}\left(S^{(s)}\right)$ y $\mathbb{E}\left(S^{(s)}\right)$ de la siguiente manera:

$$\begin{aligned} \mathbb{E}\left(S^{(s)}\right) &\leftarrow \mu_N\left(\boldsymbol{\lambda}^{(s)}\right) \cdot \mu_x\left(\boldsymbol{\theta}^{(s)}\right) \\ \mathbb{V}\text{ar}\left(S^{(s)}\right) &\leftarrow \sigma_x^2\left(\boldsymbol{\theta}^{(s)}\right) \cdot \mu_N\left(\boldsymbol{\lambda}^{(s)}\right) + \sigma_N^2\left(\boldsymbol{\lambda}^{(s)}\right) \cdot \mu_x^2\left(\boldsymbol{\theta}^{(s)}\right). \end{aligned}$$

2. Obtener el valor k definido de la siguiente manera:

$$k = \frac{\frac{1}{S_{\text{MCMC}}} \sum_{s=1}^{S_{\text{MCMC}}} \mathbb{V}\text{ar}\left(S^{(s)}\right)}{\frac{1}{S_{\text{MCMC}}-1} \sum_{s=1}^{S_{\text{MCMC}}} \left[\mathbb{E}\left(S^{(s)}\right) - \frac{1}{S_{\text{MCMC}}} \sum_{s=1}^{S_{\text{MCMC}}} \mathbb{E}\left(S^{(s)}\right)\right]^2} \tag{18}$$

3. Calcular el factor de credibilidad Z:

$$Z = \frac{m_N}{m_N + k},$$

donde m_N es el total de tomadores de póliza en un portafolio de asegurados.

4. Para cada $s \in \{1, \ldots, S_{\text{MCMC}}\}$ generar una muestra para el monto total de siniestros $S^{(s)}$ de la siguiente manera:

   - Obtener una muestra para $N(t)^{(s)}$ de manera que:

$$N(t)^{(s)} \overset{\text{d}}{=} p\left(N(t) \mid \boldsymbol{\lambda}^{(s)}\right).$$

   - Obtener $N(t)^{(s)}$ muestras de la cuantía de los siniestros $X_i^{(s)}$ con $i \in \{1, \ldots, N(t)^{(s)}\}$ de manera que:

$$X_i^{(s)} \overset{\text{iid}}{=} p\left(X_i \mid \boldsymbol{\theta}^{(s)}\right).$$

   - Guardar $S^{(s)}$ como:

$$S^{(s)} \leftarrow \sum_{i=1}^{N(t)^{(s)}} X_i^{(s)}.$$

5. Calcular $\widehat{\mathbb{E}}(S)$:

$$\widehat{\mathbb{E}}(S) = Z \cdot \frac{\sum_{i=1}^{N(t)} X_i}{m_N} + (1 - Z) \cdot \frac{1}{S_{\text{MCMC}}} \sum_{s=1}^{S_{\text{MCMC}}} S^{(s)}.$$

6. Cuantificar una medida de riesgo M para el portafolio de asegurados con m_N tomadores de pólizas, teniendo en cuenta el tiempo de vigencia o exposición del riesgo así como la suma asegurada:

$$M = \frac{1}{365} \sum_{i=1}^{m_N} \text{SA}_i \cdot t_i,$$

donde SA_i es la suma asegurada en una unidad monetaria para el riesgo i y t_i es el tiempo de exposición del riesgo i o tiempo de vigencia de su póliza en días.

7. La tasa pura de riesgo p para el portafolio de asegurados y expresada porcentualmente es:

$$p = 100 \times \frac{\widehat{\mathbb{E}}(S)}{M},$$

donde M es la medida de riesgo de exposición para el portafolio de asegurados.

8. La tasa comercial c para el portafolio de asegurados y expresada porcentualmente es:

$$c = \frac{p}{1 - \text{GA} - \text{GC} - \text{RU}},$$

donde p es la tasa pura de riesgo, GA es el porcentaje de gastos administrativos, GC es el porcentaje de gastos comerciales y RU esl porcentaje de utilidad.

9. Calcular la prima comercial PC_i que se cobra a cada tomador i por asegurar su riesgo. El cálculo de PC_i se define de la siguiente manera:

$$\text{PC}_i = \frac{\text{SA}_i \times \frac{c}{100} \times t_i}{365},$$

donde c es la tasa comercial, SA_i es la suma asegurada del tomador i y t_i el tiempo de vigencia de su póliza.

La función $P(t)$ de recaudo de primas que permite calcular el total de primas recibidas hasta en el intervalo de tiempo $[0, t]$ para $N_{m,t}$ tomadores se define de la siguiente manera si el pago se realiza de manera anticipada:

$$P(t) = \sum_{i=1}^{N_{m,t}} \text{PC}_i.$$

Si las primas se reciben en cuotas y de forma fraccionada, se proyectan los flujos de las cuotas recibidos hasta el tiempo t.

3. Análisis descriptivo.

EL conjunto de datos implementado para ilustrar la metodología de desarrollo del modelo Bayesiano jerárquico de mezcla al momento de cuantificar el riesgo de suscripción en seguros de no vida corresponde a la base dataCar ubicada en la librería insuranceData del software estadístico R. Esta base contiene 67803 registros de pólizas de seguros de autos a un año entre 2004 y 2005 reportadas por algunas de las compañías aseguradoras en Estados Unidos. La descripción de las variables se presenta por medio del siguiente diccionario de datos:

Tabla 1: *Tabla de diccionario para las variables del conjunto de datos implementado en la presente metodología. La primera columna indica el nombre de la variable, la segunda presenta su descripción, la tercera el tipo de variable, la cuarta su rango de valores y la quinta su uso para la metodología.*

Variable	**Descripción**	**Tipo**	**Rango**	**Uso**
veh_value	Indica el valor asegurado del vehículo en dólares	Cuantitativa continua	[1800,345600]	Análisis
exposure	Indica el tiempo de cobertura del en días	Cuantitativa continua	[0.99,364.75]	Análisis
clm	Indica si la póliza tuvo al menos un siniestro	Categórica nominal	1, si hubo siniestro 0, en otro caso	Respuesta
numclaims	Indica el número de siniestros que tuvo la póliza	Cuantitativa discreta	$\{0, 1, 2, 3, 4\}$	Análisis
claimcst0	Indica el monto total agregado de la cuantía de los siniestros	Cuantitativa continua	[0,55922.12]	Respuesta
veh_body	Indica el tipo de vehículo asegurado	Categórica nominal	{RDSTR,BUS,CONVT MCARA,MIBUS,RANVN COUPE,HDTOP,TRUCK UTE,STNWG,HBACK SEDAN}	Análisis
veh_age	Indica la edad en años del vehículo asegurado	Cuantitativa discreta	$\{1, 2, 3, 4\}$	Análisis
gender	Indica el género del conductor	Categórica nominal	Three Drie	Análisis
area	Indica el área donde se moviliza el vehículo asegurado	Categórica nominal	$\{A, B, C, D, E, F\}$	Análisis
agecat	Indica el tiempo en años de uso del vehículo	Cuantitativa discreta	$\{1, 2, 3, 4, 5, 6\}$	Análisis

En el estudio descriptivo y exploratorio de la base de datos se encontró que la variable del monto del valor asegurado del vehículo presenta 53 registros faltantes, los cuales corresponden a un 0.078107 %. La imputación de estos registros resulta vital para realizar la tarifación del seguro al momento de construir una función de recaudo de primas que permita estimar las trayectorias del proceso estocástico de riesgo a tiempo continuo para el crecimiento del capital de una compañía aseguradora durante un año. Estos datos no se pueden muestrear directamente de la función de verosimilitud del modelo Bayesiano propuesto en la metodología dado que este modelo estima la cuantía de los siniestros y no valores asegurados. En virtud de lo anterior, se utilizan modelos de Machine Learning como el vecino más cercano para estimar los valores perdidos del monto asegurado implementando el mejor modelo predictivo utilizando como métrica el error cuadrático medio.

A continuación, se presentan los resultados gráficos de un análisis descriptivo y exploratorio del conjunto de datos categóricos con su respectiva interpretación desde el punto de vista del negocio asegurador:

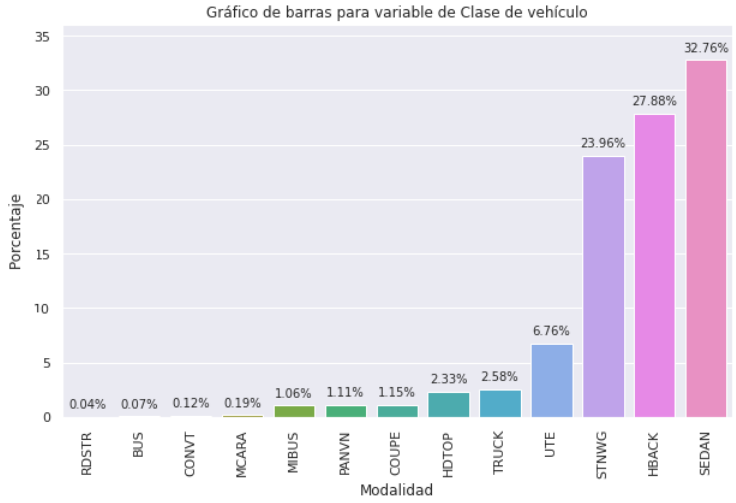

Figura 1: Gráfico de barras de riesgos asegurados por tipo de vehículo.

Por medio del gráfico anterior se evidencia que en términos de los tipos de vehículos que se encuentran dentro del portafolio de 67803 asegurados, existe una concentración en la producción de pólizas de riesgos asociados a vehículos tales como: SEDAN, HBACK y STNWG con porcentajes del 32.76 %, 27.88 % y 23.96 %, respectivamente. Aunque en mayor medida se destaca la producción de pólizas de automóviles como SEDAN, HBACK, SWITNG, desde el punto de vista del negocio asegurador, la concentración de los siniestros ocurridos o de las cuantías más altas no necesariamente se encuentra en la moda de la producción.

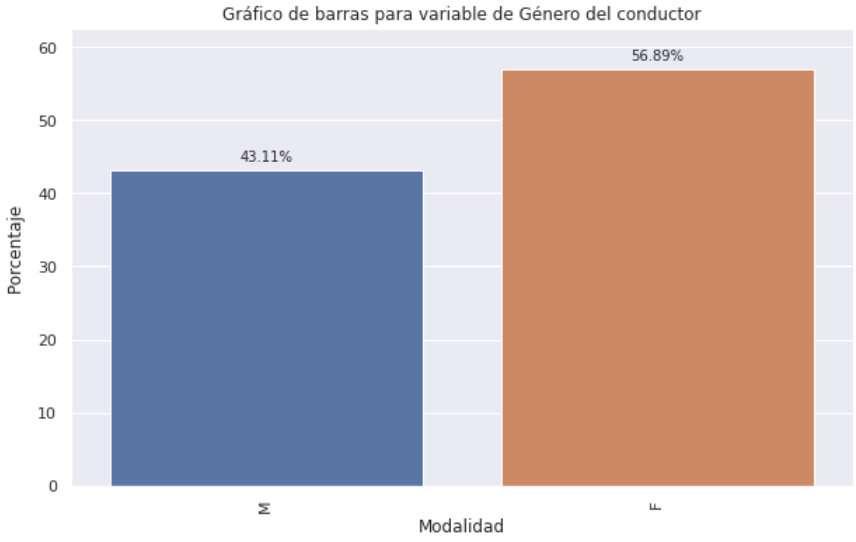

Figura 2: Gráfico de barras del género del conductor de los vehículos asegurados.

Bajo la hipótesis de que los tomadores de las pólizas de automóviles corresponden a sus conductores, se aprecia que la mayoría de los tomadores de póliza del portafolio de asegurados son mujeres y su porcentaje de participación es del 56.89 %. De esta manera, la composición del portafolio se tiene en mayor medida por conductores de género femenino.

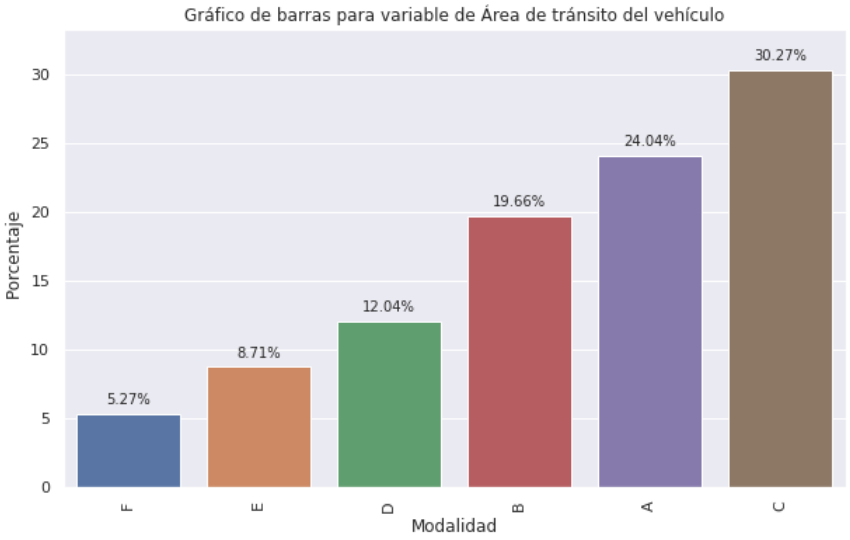

Figura 3: Gráfico de barras del área de movilización de los vehículos asegurados.

Con respecto al área donde se movilizan los vehículos del portafolio de asegurados, el gráfico anterior permite evidenciar que los automóviles transitan en mayor medida zonas como C, B y A con porcentajes de participación del 30.27 %, 24.04 % y 19.66 %, respectivamente. De esta manera, la concentración de los siniestros o sus cuantías puede estar asociada o no a la moda de las zonas donde transitan los vehículos asegurados.

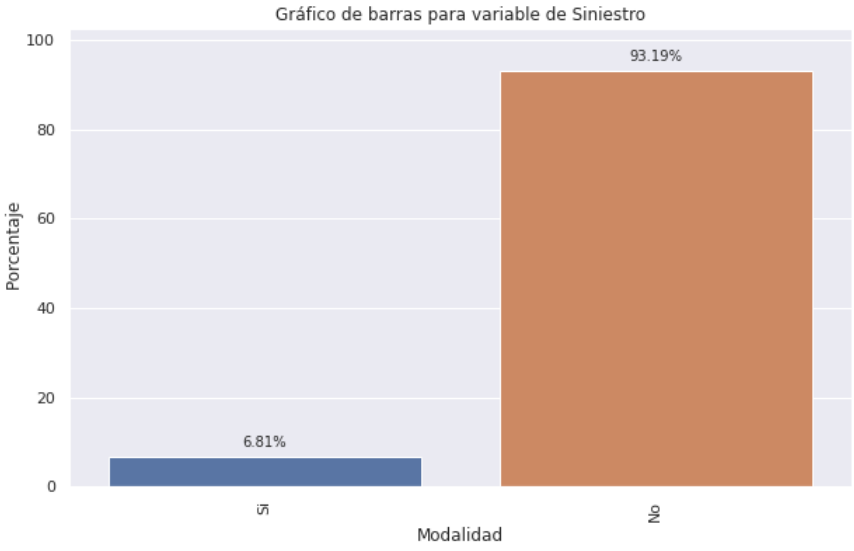

Figura 4: Gráfico de barras para la ocurrencia del siniestro.

Finalmente, en este último gráfico descriptivo para la información categórica se aprecia que el 93.19 % de las pólizas de automóviles asegurados no tienen siniestros, mientras que el restante 6.81 % de las pólizas tuvo al menos un siniestro.

Para verificar la dependencia o asociación estadística que existe entre las variables categóricas descritas anteriormente, y en particular, la dependencia que existe entre las variables asociadas a la ocurrencia

de los siniestros en pólizas de automóviles, se implementa la prueba chi cuadrado utilizando el siguiente sistema de hipótesis:

$$H_0 : \text{El par de variables categóricas son independientes.}$$
$$H_1 : \text{El par de variables categóricas son dependientes.}$$

Si interesan aquellas asociaciones estadísticamente significativas entre pares de variables categóricas, es necesario buscar aquellas en donde se rechaza la hipótesis nula. Seleccionando un nivel de significancia del 10 %, el criterio de rechazo de la hipótesis nula indica que si el valor p de la prueba chi cuadrado es tal que $p < \alpha = 0.10$, existen evidencias estadísticas suficientes para afirmar que existe asociación o dependencia entre el par de variables categóricas estudiadas. A continuación, se presenta un mapa de calor que contiene los p valores asociados a la prueba chi cuadrado por cada para de variables categóricas para verificar si bajo el criterio de rechazo de la hipótesis nula, existe o no asociación entre ellas. En el gráfico, se agregan nuevas variables cualitativas (catagecat,catvehage) para categorizar la edad y el tiempo de uso del vehículo creando dos modalidades haciendo una partición con la mediana.

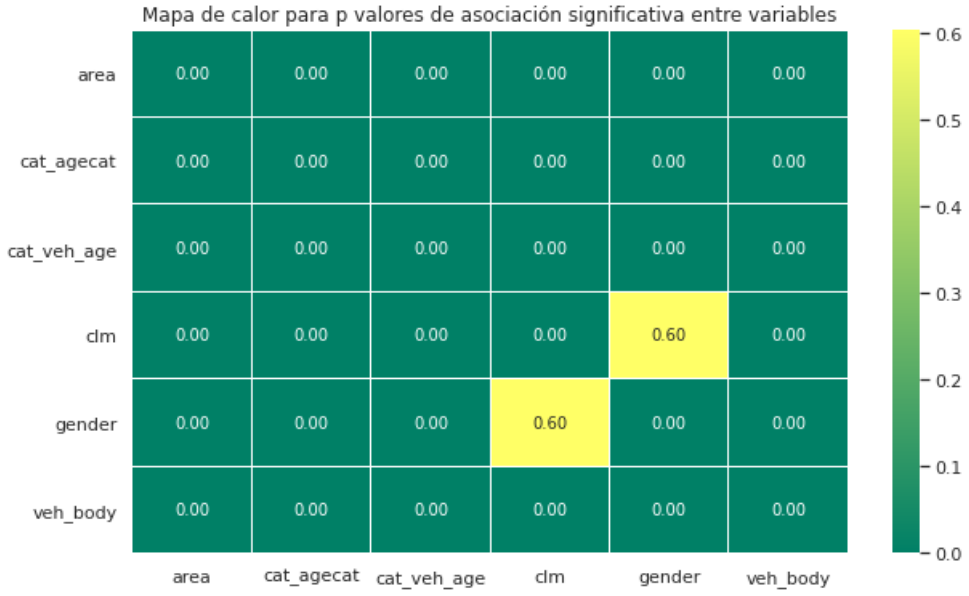

Figura 5: Mapa de calor de valores p de la prueba chi cuadrado para verificar asociación o dependencia estadística entre variables categóricas.

Observando el gráfico anterior, se aprecia que variables cualitativas como área de tránsito, tipo de vehículo, edad y tiempo de uso del vehículo categorizadas presentan una asociación estadísticamente significativa con la ocurrencia del siniestro. Por otra parte, se aprecia que el género no se asocia estadísticamente con la ocurrencia del siniestro. Este resultado es de vital importancia al momento de implementar la tarifación de seguros y de explicar los atributos que definen las clases latentes asociadas a las cuantías de los siniestros dado que el precio por la cobertura de una póliza de automóviles puede tener variaciones según el tipo del vehículo, su tiempo de uso, su edad o el área donde transita.

La prueba chi cuadrado, aunque indica la asociación o dependencia de las variables categóricas, no describe como son las relaciones entre las variables. Para describir estas relaciones, se presentan gráficos de perfil para observar la siniestralidad con respecto a cada variable categórica asociada:

En el análisis descriptivo de la producción de pólizas de automóviles se verificó que la mayor concentración se presenta en tipos de vehículo como SEDAN, HBACK y STNWG. Sin embargo, en términos de

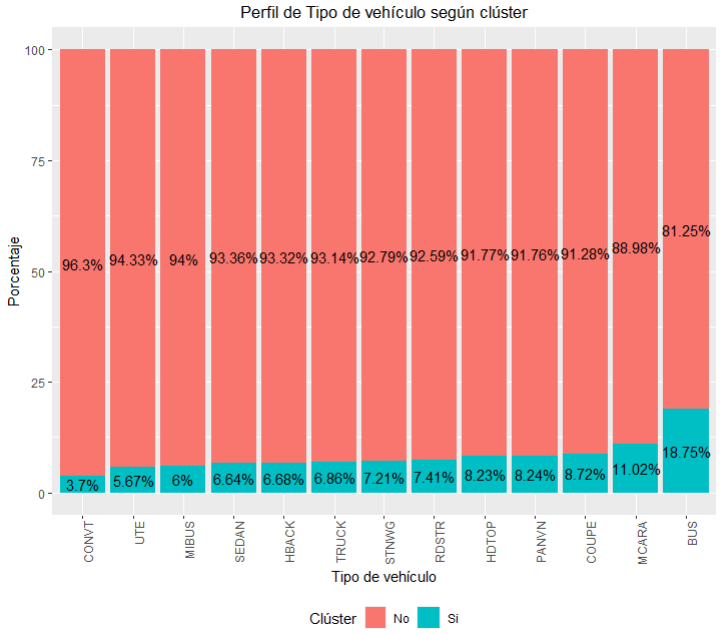

Figura 6: Gráfico de perfil de tipo de vehículo asegurado según siniestralidad.

proporciones, la mayor siniestralidad se presenta para automóviles con menor producción de pólizas y corresponden a MCARA y BUS con porcentajes correspondientes a 11.02 % y 18.75 %. De esta manera, se observa que, dependiendo del tipo de vehículo, existe una variación de la siniestralidad dado que no tiende a ser homogénea según el tipo de automóvil asegurado.

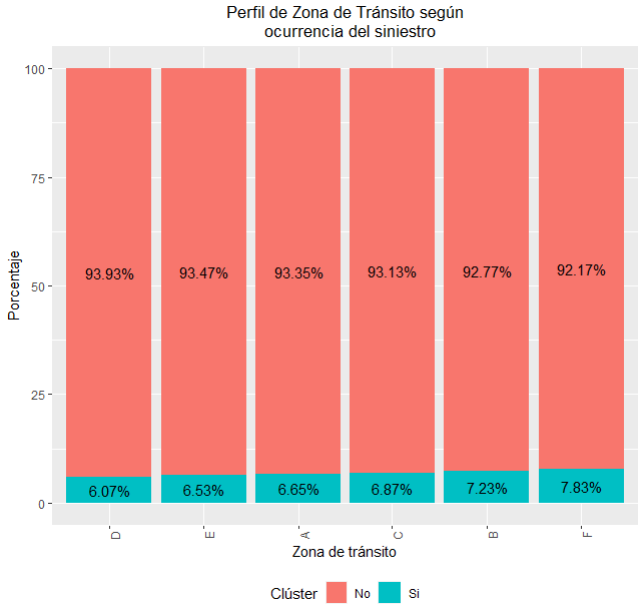

Figura 7: Gráfico de perfil de zonas de tránsito de vehículos asegurados según siniestralidad.

En este gráfico se puede apreciar que los vehículos asegurados en la zona D presentan la menor siniestralidad del portafolio, mientras que los vehículos asegurados para transitar en la zona F, presenta la

mayor siniestralidad. Es importante destacar el hecho de que la siniestralidad se concentra en la zona F, y esta corresponde a la región geográfica donde menos se expiden pólizas de automóviles.

A continuación, se presentan los perfiles de edad del vehículo asegurado y su tiempo de uso según la siniestralidad observada:

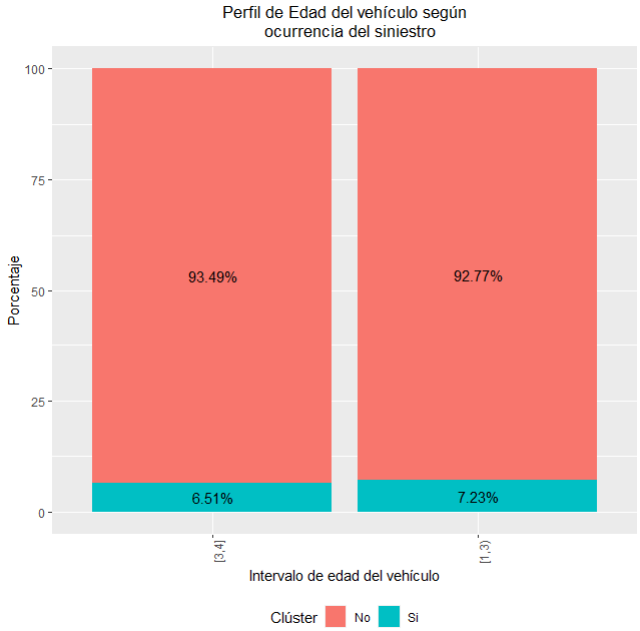

Figura 8: Gráfico de perfil de edad de vehículos asegurados asegurado según siniestralidad.

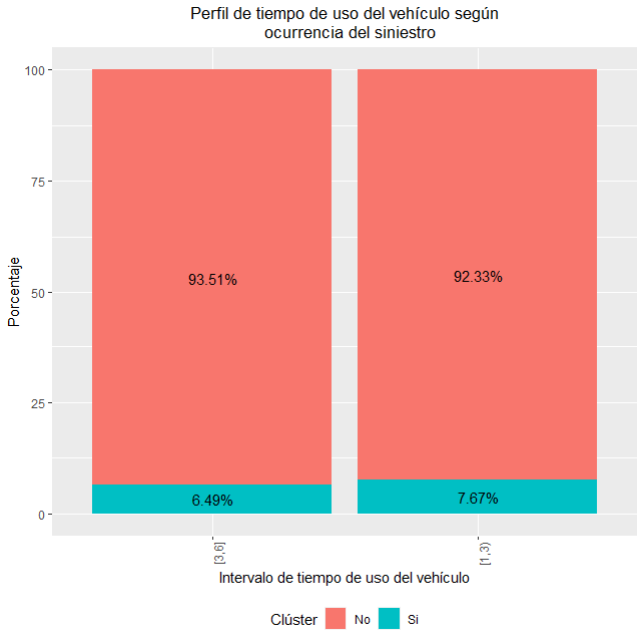

Figura 9: Gráfico de perfil de tiempo de uso de vehículos asegurados según siniestralidad.

En términos de la edad del vehículo, se aprecia que los autos con edad entre 1 y 3 años presentan una

menor siniestralidad que los autos que tienen una edad entre 3 y 4 años. Finalmente, el mismo patrón de siniestralidad se observa cuando se estudia el tiempo de uso del vehículo en virtud de que hay una mayor tendencia a la ocurrencia de siniestros en pólizas de automóviles que tienen un menor tiempo de uso de tres años.

Con respecto a la información cuantitativa encontrada en la base de datos, se presenta a continuación el gráfico de correlaciones de Pearson para verificar posibles relaciones directas o inversas entre las variables, teniendo un mayor interés por las relaciones asociadas a la cuantía de los siniestros:

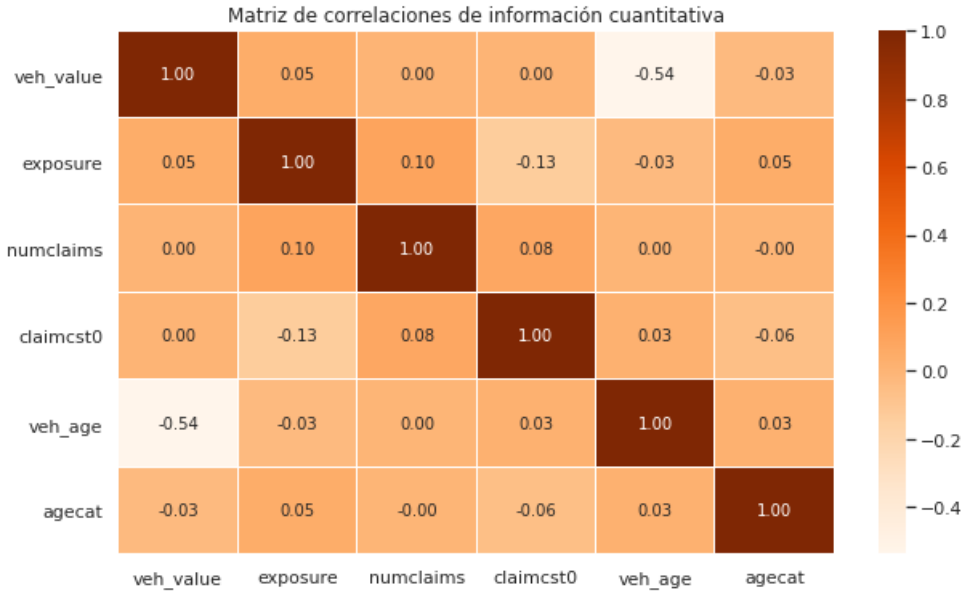

Figura 10: Matriz de correlaciones de Pearson en información cuantitativa.

La matriz de correlaciones presentada anteriormente permite afirmar que, al parecer, las correlaciones entre el monto de los siniestros y variables como el valor asegurado, tiempo de cobertura o exposición del riesgo, número de siniestros de la póliza, edad del vehículo y tiempo de uso, tienden a ser nulas. Sin embargo, en términos de las correlaciones observadas más 'fuertes' se evidencia que entre mayor es el tiempo en días de la cobertura de la póliza, menor tiende a ser la cuantía de los siniestros, mientras que a mayor número de siniestros se presente en la póliza, mayor es la cuantía final asociada al monto total causado por el siniestro. Por otra parte, es importante destacar la relación inversa entre el valor asegurado y la edad del vehículo, de manera que, a mayor edad, menor es el valor comercial de la suma con la cual se aseguraba un automóvil en Estados Unidos entre 2004 y 2005.

Una vez analizadas las variables cuantitativas y su correlación con el monto de los siniestros, es necesario extender el estudio descriptivo y exploratorio de esta variable en mención con respecto a variables categóricas como el tipo de vehículo asegurado, género del conductor y área de tránsito del vehículo. Para cumplir con este propósito se proponen a continuación diferentes visualizaciones de gráficos de boxplot para conjeturar posibles diferencias que existen entre el monto de los siniestros según las modalidades de cada variable categórica.

El siguiente gráfico permite observar que al parecer existen diferencias entre el monto de los siniestros según el tipo de vehículo asegurado al establecer una comparación entre las medianas del logaritmo de los montos. Este posible resultado es a penas natural porque en el contexto del sector asegurador, en el ramo de automóviles las cuantías pueden variar dependiendo de si el vehículo es liviano o pesado. En este sentido, para respaldar la afirmación anterior, se observa que el microbús presenta la mayor cuantía en términos de la comparación de medianas. De esta manera, la tarifación de seguros en función de las

tasas puras de riesgo y comercial puede depender en gran medida del tipo de vehículo asegurado.

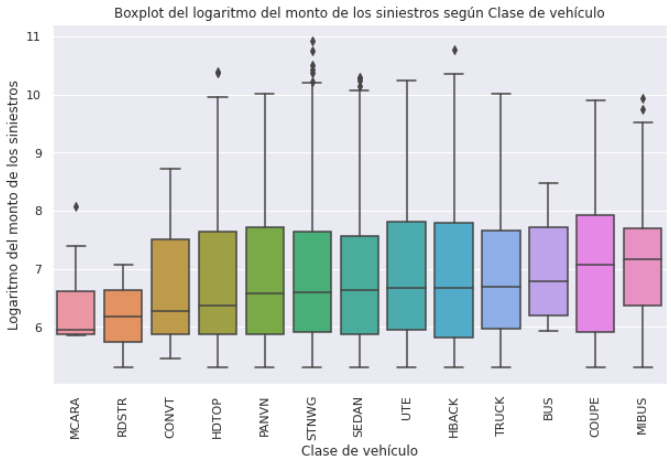

Figura 11: Boxplot del logaritmo del monto de los siniestros según el tipo de vehículo asegurado.

El siguiente boxplot permite observar las diferencias entre el logaritmo del monto de las cuantías de los siniestros según el género del conductor del vehículo asegurado:

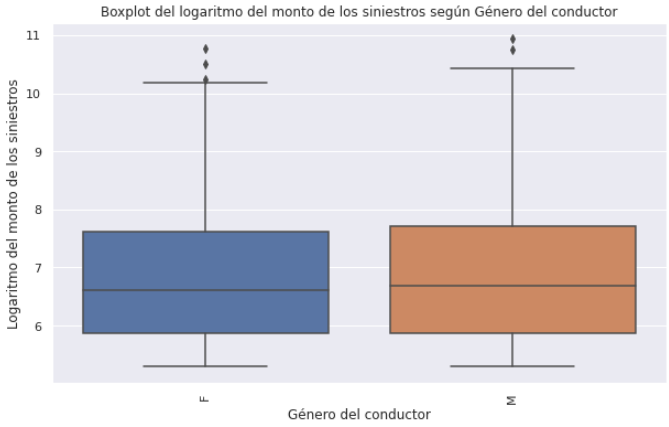

Figura 12: Boxplot del logaritmo del monto de los siniestros según el género del conductor del vehículo asegurado.

Comparando el monto de los siniestros de los automóviles asegurados en 2004 y 2005 en Estados Unidos por género del conductor, se puede evidenciar por medio del gráfico anterior que al parecer no existen diferencias entre las medianas de las cuantías. Este resultado puede ser un indicador de que el género del conductor del vehículo asegurado no necesariamente es un factor significativo al momento de hacer la tarifación de seguros de automóvil según el comportamiento de la siniestralidad observado en la industria del sector asegurador del ramo de automóviles que incluye tanto vehículos livianos como pesados.

Finalmente, los montos de la siniestralidad según la zona donde transitan regularmente los vehículos asegurados en Estados Unidos durante 2004 y 2005 parecen no presentar grandes diferencias al comparar sus medianas, pero un hecho apreciable por medio del siguiente gráfico es que la severidad de las cuantías que se presentan en la zona F es la mayor entre las zonas geográficas observadas.

De este modo, parece que la zona geográfica puede no ser un factor muy influyente para tener en cuenta dentro del proceso de tarifación de seguros de automóvil. Sin embargo, es importante apreciar si se

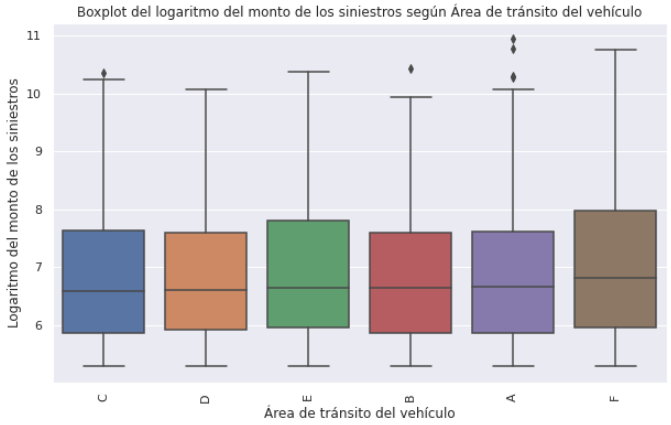

Figura 13: Boxplot del logaritmo del monto de los siniestros según el área de tránsito del vehículo asegurado.

presentan estas mismas diferencias entre la siniestralidad de las diferentes zonas geográficas al momento de establecer comparación en los tipos de vehículo asegurados. En virtud de lo anterior se aprecia la siguiente visualización:

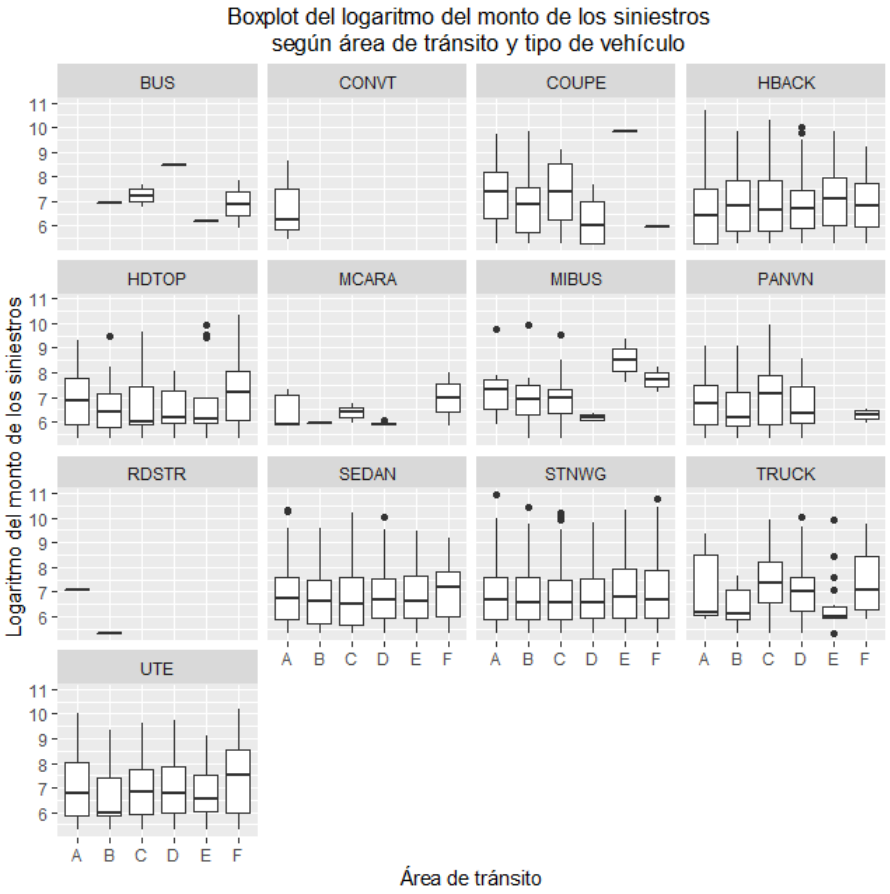

Figura 14: Boxplot del logaritmo del monto de los siniestros según el tipo de vehículo asegurado y su área de tránsito.

Por medio del gráfico de boxplot presentado anteriormente se comparan los montos de la siniestralidad por zona geográfica según el tipo de vehículo asegurado. La visualización permite apreciar que la siniestralidad del auto convertible solo se presenta en la zona A y esto atribuye a que solo hay producción de pólizas para este auto en esta zona o porque en las demás zonas la siniestralidad es nula. Por otra parte, para tipos de vehículo como SEDAN y STWNG parecen no existir grandes diferencias entre las cuantías de los siniestros por zona geográfica mientras que para los demás tipos de vehículo parecen existir diferencias entre las cuantías tal y como se observa en el caso de microbús, siendo las zonas F y E, respectivamente aquellas que presentan mayor severidad en los valores de los siniestros. De esta manera, aunque la comparación de los montos de la siniestralidad entre zonas geográficas ignorando el tipo de vehículo parece no mostrar diferencias, la interacción entre la zona geográfica y el tipo de vehículo asegurado pueden ser una combinación que incida significativamente en la tarifación de seguros de automóvil.

Hecha la comparación de los montos de la siniestralidad por tipo de vehículo asegurado, tiene sentido realizar la comparación entre el género del conductor por tipo de vehículo.

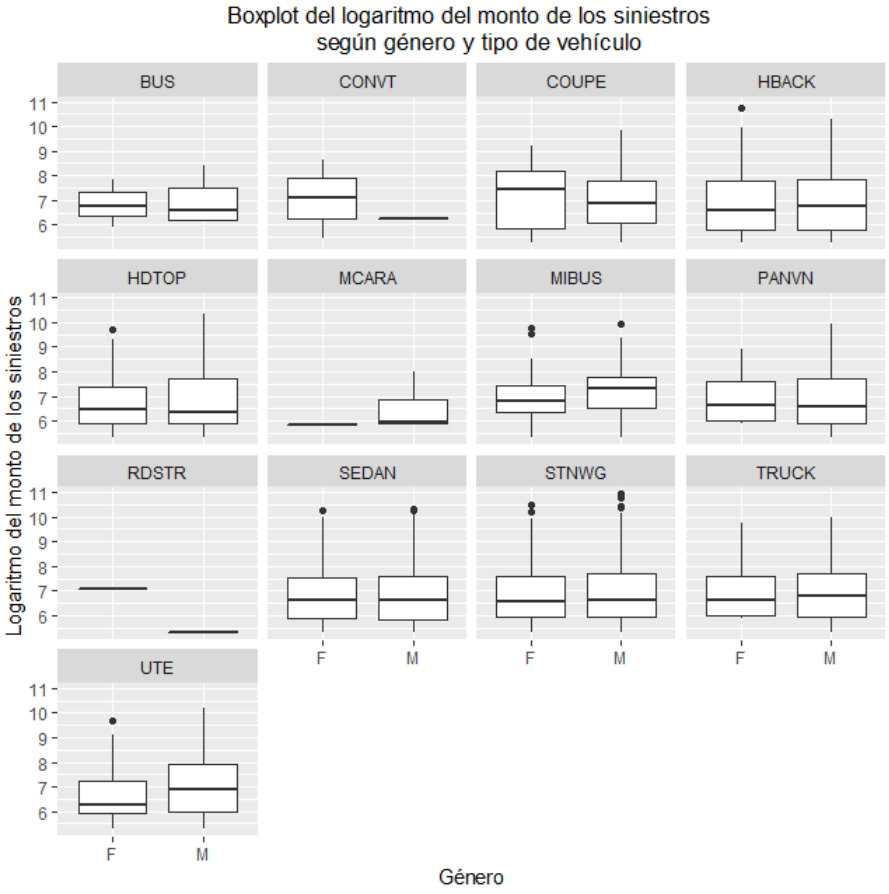

Figura 15: Boxplot del logaritmo del monto de los siniestros según el tipo de vehículo asegurado y género de su conductor.

La visualización presentada anteriormente permite apreciar que, aunque al parecer, en la mayoría de los tipos de vehículos asegurados no se observan diferencias entre las medianas de las cuantías de los siniestros por género del conductor, se tienen algunos tipo de vehículo como convertible, COUPE, MCARA, microbús, RDSTR y UTE donde la diferencia entre las cuantías parece no ser despreciable. En estos casos, cuando el tipo de vehículo asegurado es: convertible, COUPE y RDSTR parece existir una incidencia de las mujeres en cuantías de siniestros más severas que las de los hombres; mientras que en vehículos como microbús o UTE el efecto en función de cuantías más severas resulta contrario.

Una vez hecho el análisis descriptivo y exploratorio tanto de información cuantitativa como categórica que puede estar asociada a la ocurrencia de siniestros de automóviles o bien a sus cuantías, se estudian de forma independiente la frecuencia y la severidad de los siniestros. Con respecto a la frecuencia de los siniestros, indicadores descriptivos tales como media y varianza del número de siniestros ocurridos por semana proporcionan los siguientes resultados:

Tabla 2: *Tabla de estadísticos de media y varianza para el número de siniestros ocurridos por semana.*

Estadístico	**Valor**
Media	88, 92308
Varianza	1730, 661

De forma exploratoria se evidencia que la varianza es mayor que la media, de manera que resulta razonable proponer el ajuste de una distribución binomial negativa para la frecuencia de los siniestros. Sin embargo, al observar el número de siniestros ocurridos y agregados por póliza en cada semana se aprecia una posible similaridad entre los estadísticos de media y varianza, aunque se encuentran en un comportamiento creciente con tendencia lineal con respecto al horizonte temporal cuando se define un mes como una unidad de tiempo de cuatro semanas:

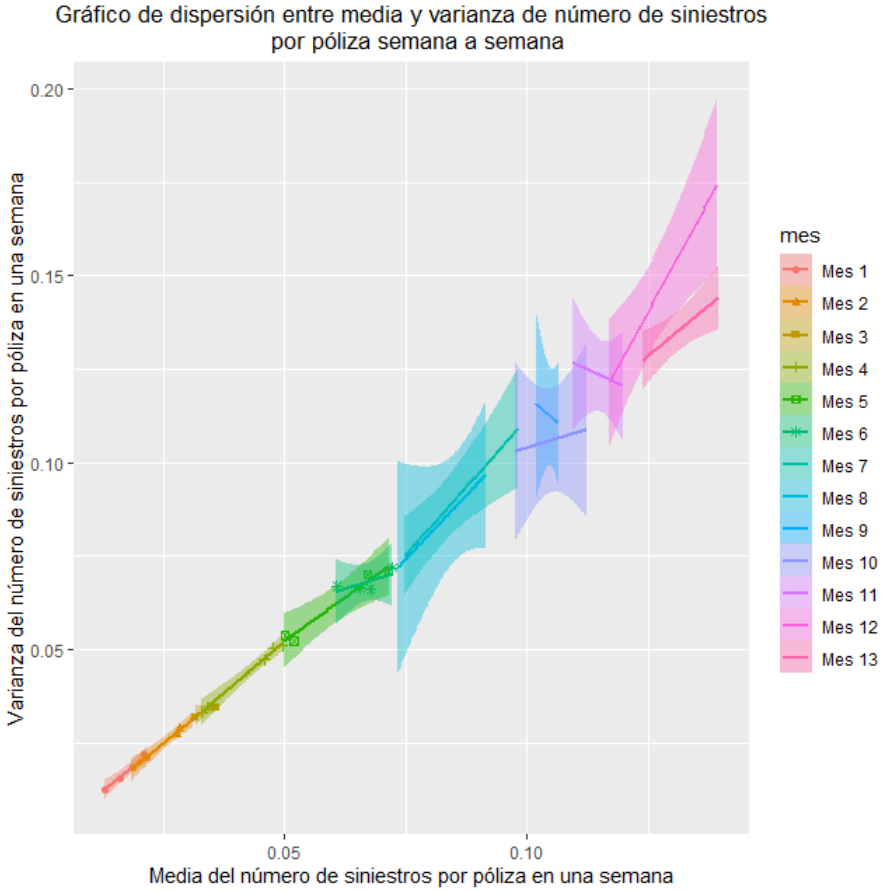

Figura 16: Gráfico de dispersión de media y varianza del número de siniestros ocurridos agregados por póliza en unidades de tiempo semanales.

De esta manera, la propuesta de un modelo Poisson también resulta razonable si se explica correctamente su parámetro como una realización de una variable aleatoria que depende del tiempo.

Por otra parte, las funciones de densidad de la severidad y el logaritmo de la severidad de los siniestros permiten apreciar que:

- La distribución del monto de los siniestros ocurridos es asimétrica a la derecha dado que la mayoría de los siniestros presentan cuantías que no exceden el monto de los 10.000 dólares, mientras que la minoría de los siniestros excede este valor.
- La distribución del logaritmo del monto de los siniestros ocurridos parece tener al menos tres modas, de manera que es posible construir clústeres o grupos de riesgos asociados asociados a densidades de funciones de mixtura.

Los resultados mencionados anteriormente se visualizan por medio del siguiente gráfico exploratorio:

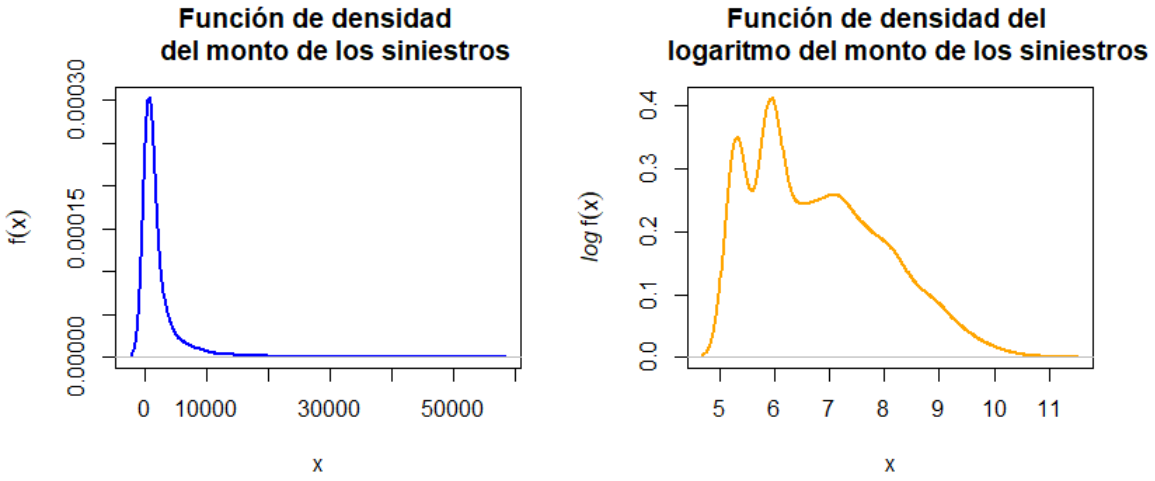

Figura 17: Funciones de densidad del monto de los siniestros y el logaritmo del monto de los siniestros.

4. Modelo Bayesiano para la frecuencia de los siniestros ocurridos.

Dado que la varianza de los siniestros ocurridos por semana es mayor que su media tal y como se evidenció en el análisis descriptivo presentado en la sección anterior, la propuesta de un modelo binomial negativo resulta razonable al momento de modelar la frecuencia de los siniestros.

4.1. Propuesta de modelos Bayesianos para la frecuencia de siniestros.

4.1.1. Modelo Bayesiano Binomial Negativo.

Sea N una variable aleatoria que representa la frecuencia del número de siniestros ocurridos por semana de manera que:

$$N \overset{\text{iid}}{=} \text{Binomial Negativa}(\theta, k),$$

donde N es el número de ensayos que debe realizarse un experimento aleatorio hasta obtener k éxitos, de manera que cada ensayo tiene una probabilidad de éxito θ. En estos términos, el soporte para θ se encuentra entre 0 y 1, razón por la cual θ se propone como una realización de una distribución Beta con parámetros α y β, y k como una realización de una distribución exponencial de parámetro α_0:

$$\begin{aligned} \theta &\overset{\text{d}}{=} \text{Beta}(\alpha, \beta) \\ k &\overset{\text{d}}{=} \text{Exponencial}(\alpha_0). \end{aligned}$$

De acuerdo con lo anterior, los parámetros del modelo bayesiano son θ y k, mientras que los hiperparámetros son α, β y α_0. Como el número de siniestros ocurridos no presenta información previa, los valores propuestos para los hiperparámetros corresponden a distribuciones Beta y exponencial no informativas. Una propuesta de α y β para una distribución Beta no informativa es tal que $\alpha = 1$ y $\beta = 1$ porque en este caso, la distribución resultante es uniforme continua con soporte entre 0 y 1, de manera que la probabilidad θ es una realización aleatoria entre 0 y 1 de la cual no se tiene un conocimiento previo. Por otra parte, una distribución exponencial es no informativa si $\alpha_0 = 1$, porque de esta manera su coeficiente de variación es del 100 %.

Si se tienen m observaciones de $n_1, \ldots, n_m$ de la variable aleatoria N, la distribución condicional completa para $\theta \mid$ resto es tal que:

$$\theta \mid \text{resto} \overset{\text{d}}{=} \text{Beta}\left(\alpha + mk, \beta + \textstyle\sum_{i=1}^{m} n_i\right), \tag{19}$$

donde resto $= (N, \alpha, \beta, k, \alpha_0)$.

La distribución condicional completa para $k \mid$ resto es tal que:

$$p_k(k \mid \text{resto}) \propto \frac{\prod_{i=1}^{m} \Gamma(k + n_i)}{[\Gamma(k)]^m} \cdot \theta^{mk} \cdot e^{-\alpha_0 \cdot k}, \tag{20}$$

donde resto $= (N, \alpha, \beta, \theta, \alpha_0)$. Como no existe una distribución conocida para $k \mid$ resto, para muestrear sus valores se implementa un algoritmo de Metrópolis adaptativo utilizando la distribución normal con media k y varianza δ^2 bajo la restricción de que $k > 0$. En este sentido, conocer el logaritmo de $p(k \mid$ resto) permite generar posteriormente un condicional para la aceptación de las muestras de valores de la

distribución bajo el algoritmo adaptativo de Metrópolis tal y como lo sugiere Gelman, A. et al. (2014) mediante algunas extensiones para MCMC. De esta manera se tiene que:

$$\log p_k(k \mid \text{resto}) = \sum_{i=1}^{m} \log\Gamma(k+n_i) - m\log\,\Gamma(k) + mk\log\,\theta - \alpha_0 k.$$

Al momento de muestrear los valores candidatos k para la distribución posterior condicional completa $p(k \mid \text{resto})$ de una normal, es necesario estar sujeto a la restricción $k > 0$. En este sentido, la aplicación de una función g como enlace permite obtener siempre muestras de k sujetas a la restricción en mención de manera que $g : (0, \infty) \to \mathbb{R}$. Por tanto, definiendo $g(k) = \Phi$, se muestrea un valor Φ^* de manera que:

$$\Phi^* \overset{\text{d}}{=} \text{Normal}(\Phi, \delta^2).$$

En caso de aceptación del valor muestreado de la distribución condicional completa $p(k \mid \text{resto})$, se considera como muestra a e^{Φ^*}. De manera natural, una función adecuada para g es el logaritmo natural y por tanto $\Phi = \log\,(k)$. En vista de la aplicación de la función g, se requiere considerar la función de densidad de Φ, la cual es obtenida mediante el teorema de la transformación expuesto en [11]:

$$\begin{aligned} p_\Phi(\phi) &= \mid J(\phi) \mid \cdot p_k(k = h(\phi) \mid \text{resto}) \\ &= e^{\phi} \cdot p_k(e^{\phi} \mid \text{resto}), \end{aligned}$$

donde h es la función inversa de g y es tal que $h(\phi) = e^{\phi}$. En virtud de lo anterior el logaritmo de la función de densidad es:

$$\log p_\Phi(\Phi = \phi) = \phi + \sum_{i=1}^{m} \log\Gamma(e^{\phi}+n_i) - m\log\,\Gamma(e^{\phi}) + m \cdot e^{\phi} \cdot \log\,\theta - \alpha_0 \cdot e^{\phi}.$$

Como consecuencia, el logaritmo de la razón de aceptación en el algoritmo adaptativo de Metrópolis para el valor candidato e^{Φ^*} muestreado de la distribución condicional completa $p_k(k \mid \text{resto})$ en la $j-$ésima iteración para todo $j \in \mathbb{Z}^+$, corresponde a:

$$\begin{aligned} \log r = &\left[\sum_{i=1}^{m} \log\Gamma(e^{\Phi^*}+n_i) - m\log\,\Gamma(e^{\Phi^*}) + m \cdot e^{\Phi^*} \cdot \log\,\theta - \alpha_0 \cdot e^{\Phi^*}\right] \\ &- \left[\sum_{i=1}^{m} \log\Gamma(e^{\Phi}+n_i) - m\log\,\Gamma(e^{\Phi}) + m \cdot e^{\Phi} \cdot \log\,\theta - \alpha_0 \cdot e^{\Phi}\right] + (\Phi^* - \Phi). \end{aligned}$$

El algoritmo de Metrópolis adaptativo bajo iteraciones de MCMC de la propuesta de modelo Bayesiano Binomial Negativo presentado anteriormente bajo el enfoque de Gelman, A. et al. (2014) presenta los siguientes pasos:

1. En la iteración inicial, generar una realización para θ, k respectivamente de la siguiente manera:

$$\begin{aligned} \theta^{(0)} &\overset{\text{d}}{=} \text{Beta}(\alpha = 1, \beta = 1) \\ k^{(0)} &\overset{\text{d}}{=} \text{Exponencial}(\alpha_0 = 1). \end{aligned}$$

2. Optimizar el valor de δ^2 de manera que la razón de aceptación se encuentre entre el 20 % y el 50 %.
3. Para la iteración del paso s con $s \geq 1$, realizar el siguiente procedimiento:
   - Muestrear un valor para $\theta^{(s)}$ tal que:

$$\theta^{(s)} \overset{\text{d}}{=} \text{Beta}\left(\alpha + m \cdot k^{(s-1)}, \beta + \sum_{i=1}^{m} n_i\right)$$

- Obtener una muestra para $k^{(s)}$ | resto utilizando los pasos del algoritmo del Metrópolis adaptativo de la siguiente manera:

  a) Guardar $\Phi^{(s-1)}$ como:

  $$\Phi^{(s-1)} \leftarrow \log\left(k^{(s-1)}\right).$$

  b) Muestrear Φ^* tal que $\Phi^* \overset{\text{d}}{=} \text{Normal}\left(\Phi^{(s-1)}, \delta^2\right)$.

  c) Computar el logaritmo de la razón de aceptación r_s mediante la aplicación:

  $$\begin{aligned}\log r_s = &\left[\sum_{i=1}^{m} \log\Gamma(e^{\Phi^*} + n_i) - m\log\,\Gamma(e^{\Phi^*}) + m \cdot e^{\Phi^*} \cdot \log\,\theta - \alpha_0 \cdot e^{\Phi^*}\right] \\ &- \left[\sum_{i=1}^{m} \log\Gamma(e^{\Phi^{(s-1)}} + n_i) - m\log\,\Gamma(e^{\Phi^{(s-1)}}) + m \cdot e^{\Phi^{(s-1)}} \cdot \log\,\theta - \alpha_0 \cdot e^{\Phi^{(s-1)}}\right] \\ &+ (e^{\Phi^*} - e^{\Phi^{(s-1)}}).\end{aligned}$$

  d) Muestrear un número aleatorio u_s tal que $u_s \overset{\text{d}}{=} \text{Uniforme}(0,1)$. Calcular el estado actual para $k^{(s)}$ de la siguiente manera:

  $$k^{(s)} = \begin{cases} e^{\Phi^*} & \text{, si } \log(u_s) \leq \log(r_s) \\ k^{(s-1)} & \text{, en otro caso.} \end{cases}$$

4. Guardar el vector de parámetros $\boldsymbol{\Theta}^{(k)}$ tal que:

$$\boldsymbol{\Theta}^{(s)} \leftarrow \left(\theta^{(s)}, k^{(s)}\right).$$

5. Generar los pasos 3 y 4 en S iteraciones, de manera que $S = 210.000$

6. Quemar las primeras 10.000 iteraciones y muestrear las observaciones de manera sistemática con salto $a = 3$.

Una vez implementado el algoritmo MCMC para la propuesta de modelo Binomial Negativo para la frecuencia de los siniestros ocurridos por semana, se presentan a continuación los tamaños efectivos de muestra, media posterior e intervalos de credibilidad por parámetro:

Tabla 3: *La tabla reporta en su primera columna el parámetro de interés del modelo Bayesiano Binomial Negativo, la segunda columna reporta el tamaño efectivo de muestra en el algoritmo MCMC, la tercera y quinta columna presentan los límites inferior y superior de los intervalos de credibilidad del* 95 %*, mientras que la cuarta columna proporciona la media posterior.*

Parámetro	Tamaño Efectivo	Límite inferior Intervalo de Credibilidad 95 %	Media Posterior	Límite superior Intervalo de Credibilidad 95 %
θ	3652,772000	0,034471	0,052914	3,246103
k	3420,391000	0,074967	4,959306	7,080416

Con respecto a la bondad del ajuste del modelo, al explorar estadísticas como media, varianza, coeficiente de variación y desviación estándar se observaron los siguientes valores p predictivos asociados a estas estadísticas:

Tabla 4: *La primera columna muestra el estadístico de bondad de ajuste para el modelo Bayesiano Binomial Negativo para la frecuencia de los siniestros y la segunda columna su respectivo valor p predictivo.*

Estadístico	PPI
Media	0,4810176
Varianza	0,4370378
Coeficiente de variación	0,4481228
Desviación estándar	0,4370378

Los valores p predictivos de cada uno de los estadísticos se encuentran alrededor de $0,5$, razón por la cual se considera razonable una bondad del ajuste adecuada de este modelo para la frecuencia de los siniestros ocurridos por semana. Por otra parte, la tasa de aceptación de valores muestreados en el algoritmo de Metrópolis adaptativo para el parámetro k corresponde a un valor del $39,50905\,\%$. Finalmente, las cadenas de Markov de cada uno de los parámetros luego de la etapa de quemado de observaciones y de muestreo sistemático son estacionarias en media tal y como se aprecia a continuación:

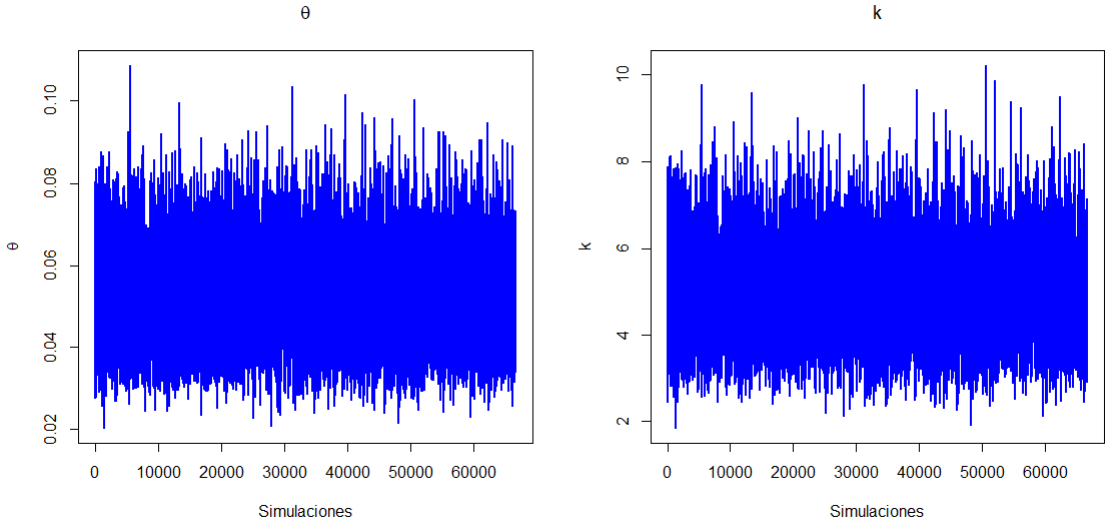

Figura 18: Cadenas de Markov para θ y k en algoritmo MCMC luego de la etapa de quemado y muestreo sistemático de las observaciones.

4.1.2. Modelo lineal generalizado Poisson.

En el análisis descriptivo de la frecuencia de los siniestros ocurridos observado mediante la ilustración 16 es posible apreciar que, al parecer existe una 'similaridad' entre la media y la varianza del número agregado de siniestros por pólizas o riesgos asegurados por semana, de manera que la media o la varianza presentan un crecimiento lineal con respecto al tiempo. De esta manera, la propuesta de un modelo lineal generalizado de Poisson Bayesiano puede resultar razonable. La propuesta considera predictores lineales tales como: el intercepto constante e igual a uno y el tiempo de ocurrencia de los eventos. Estos predictores se encuentran en una matriz de diseño X de tamaño $m \times 2$ dado que se tienen m observaciones $n_1, \ldots . n_m$ de la variable aleatoria N que representa la frecuencia del número de siniestros ocurridos por semana. De esta manera, la función de verosimilitud según Gelman, A. et al. (2014) está dada por:

$$n_i \mid \theta_i \overset{\text{d}}{=} \text{Poisson}(\theta_i \cdot N_i),$$

donde N_i representa el número total de pólizas o riesgos de automóviles asegurados en la semana i, y θ_i es la tasa de siniestralidad de eventos ocurridos durante la semana i. Como la tasa de siniestralidad θ_i está sujeta a la restricción $\theta_i > 0$, entonces se requiere de una función de enlace para θ_i de manera que una vez hecha la aplicación de está función se encuentre explicado por los predictores lineales y un conjunto de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1)$:

$$\log\,(\theta_i) = \boldsymbol{\beta}^T \cdot X_i.$$

Finalmente, el vector de parámetros $\boldsymbol{\beta}$ bajo el enfoque de Gelman, A. et al. (2014) es una realización de una distribución normal p variante con $p = 2$ tal que:

$$\boldsymbol{\beta} \overset{\text{d}}{=} \text{Normal}_p(\boldsymbol{\mu_0}, \Sigma),$$

donde $\boldsymbol{\mu_0}$ es un vector columna de tamaño $p \times 1$ con cada uno de sus componentes iguales a cero, y Σ es una matriz de varianzas y covarianzas definida como $\Sigma = S \cdot I$, donde $S \in \mathbb{R}^+$ e I es una matriz identidad de tamaño $p \times p$. Dado que no se conoce información previa sobre $\boldsymbol{\beta}$, se propone una distribución previa no informativa de manera que S toma un valor "grande", por ejemplo $S = 1.000$.

La distribución condicional completa para $\boldsymbol{\beta}$ se obtiene de la siguiente manera:

$$p(\boldsymbol{\beta} \mid N, X) \propto \prod_{i=1}^{m} \frac{\exp\left\{\left(-e^{\boldsymbol{\beta}^T \cdot X_i}\right) \cdot N_i\right\}\left(-e^{\boldsymbol{\beta}^T \cdot X_i \cdot N_i}\right)^{n_i}}{n_i!} \cdot \prod_{k=1}^{p} \frac{1}{S \cdot \sqrt{2\pi}} \exp\,\{-\frac{1}{2S^2}\beta_k^2\}.$$

Como la distribución condicional completa no presenta un kernel conocido, se utiliza un algoritmo de Metrópolis adapatativo para muestrear valores de esta distribución. En estos términos, se requiere el logaritmo de la expresión anterior:

$$\begin{aligned}\log\, p(\beta \mid N, X) = \sum_{i=1}^{m} \left[e^{\boldsymbol{\beta}^T \cdot X_i} \cdot N_i + n_i \cdot (\boldsymbol{\beta}^T \cdot X_i + \log(Ni)) - \log\,(n_i!)\right] \\ - \frac{p}{2}\left[\log\,(S^2) + \log\,(2\pi)\right] - \frac{1}{2S^2}\sum_{k=1}^{p}\beta_k^2.\end{aligned}$$

En el caso de una distribución previa no informativa, se tiene que el valor de S^2 debe ser "grande". En consecuencia, se toma S^2 tal que $S^2 = 1000$.

Al momento de muestrear valores de la distribución condicional completa $p(\boldsymbol{\beta} \mid N, X)$ se implementa un algoritmo de Metrópolis adaptativo haciendo la propuesta de un vector posible $\boldsymbol{\beta}^*$ para la muestra por medio de una distribución normal p variante de manera que:

$$\boldsymbol{\beta}^* \overset{\text{d}}{=} \text{Normal}_p\left(\boldsymbol{\beta}, \delta^2(X^T X)^{-1}\right),$$

donde δ^2 es un real positivo que debe optimizarse hasta obtener tasas de aceptación del algoritmo adaptativo entre 20 % y 50 %. La aceptación de cada vector o candidato propuesto depende del logaritmo de la razón r tal que:

$$\log\,(r) = \log\, p(\boldsymbol{\beta}^* \mid N, X) - \log\, p(\boldsymbol{\beta} \mid N, X).$$

En consecuencia, los pasos para implementar el MCMC bajo este algoritmo de Metrópolis adpatativo son:

1. En la iteración inicial, generar una realización para $\boldsymbol{\beta}$ de manera que:

$$\boldsymbol{\beta}^{(0)} \overset{\text{d}}{=} \text{Normal}_p\left(\mathbf{0}_{p\times 1}, S \cdot I_{p\times p}\right),$$

   donde $S^2 = 1.000$ e $I_{p\times p}$ es la matriz identidad de tamaño $p \times p$.

2. Optimizar el valor de δ^2 de manera que la razón de aceptación se encuentre entre el 20 % y el 50 %.

3. Para la iteración del paso s con $s \geq 1$, realizar el siguiente procedimiento:

   a) Muestrear $\boldsymbol{\beta}^*$ tal que $\boldsymbol{\beta}^* \overset{\text{d}}{=} \text{Normal}_p\left(\boldsymbol{\beta}^{(s-1)}, \delta^2(X^TX)^{-1}\right)$.

   b) Computar el logaritmo de la razón de aceptación r_s en su s-ésima iteración de manera que:

$$\begin{aligned}\log r_s = &\sum_{i=1}^{m}\left[e^{\boldsymbol{\beta^*}^T\cdot X_i}\cdot N_i + n_i\cdot\left(\boldsymbol{\beta^*}^T\cdot X_i + \log(Ni)\right) - \log\left(n_i!\right)\right] - \frac{1}{2S^2}\sum_{k=1}^{p}(\beta_k^*)^2 \\ &- \sum_{i=1}^{m}\left[e^{\boldsymbol{\beta^{(s-1)}}^T\cdot X_i}\cdot N_i + n_i\cdot\left(\boldsymbol{\beta^{(s-1)}}^T\cdot X_i + \log(Ni)\right) - \log\left(n_i!\right)\right] + \frac{1}{2S^2}\sum_{k=1}^{p}\left(\beta_k^{(s-1)}\right)^2.\end{aligned}$$

   c) Muestrear un número aleatorio u_s tal que $u_s \overset{\text{d}}{=} \text{Uniforme}(0,1)$. Calcular el estado actual del vector $\boldsymbol{\beta}^{(s)}$:

$$\boldsymbol{\beta}^{(s)} = \begin{cases}\boldsymbol{\beta^*} & \text{, si } \log(u_s) \leq \log(r_s) \\ \boldsymbol{\beta}^{(s-1)} & \text{, en otro caso.}\end{cases}$$

4. Generar los pasos 2 y 3 en S iteraciones, de manera que $S = 210.000$

5. Quemar las primeras 10.000 iteraciones y muestrear las observaciones de manera sistemática con salto $a = 3$.

Una vez implementado el algoritmo MCMC propuesto anteriormente se tienen los siguientes tamaños efectivos de muestra e intervalos de credibilidad del 95 % para el intercepto β_0 del modelo y el parámetro del tiempo β_1:

Tabla 5: *La tabla reporta en su primera columna los parámetros del modelo de regresión Poisson, la segunda columna reporta el tamaño efectivo de muestra en el algoritmo MCMC, la tercera y quinta columna presentan los límites inferior y superior de los intervalos de credibilidad del* 95 %*, mientras que la cuarta columna proporciona la media posterior.*

Parámetro	Tamaño Efectivo	Límite inferior Intervalo de Credibilidad 95 %	Media Posterior	Límite superior Intervalo de Credibilidad 95 %
β_0	20307,81	-3,694841	-3,62478623	-3,554900
β_1	22392,44	0,030794	0,03276265	0,034743

Con respecto a la bondad del ajuste se puede evidenciar por medio de la siguiente tabla que proporciona los valores p predictivos que solamente se presenta un "buenïndicador para la media, pero no para estadísticas como la varianza, la desviación estándar o el coeficiente de variación:

Tabla 6: *La primera columna muestra el estadístico de bondad de ajuste para el modelo Bayesiano Binomial Negativo para la frecuencia de los siniestros y la segunda columna su respectivo valor p predictivo.*

Estadístico	PPI
Media	0,4953200
Varianza	0,9086491
Coeficiente de variación	0,9228242
Desviación estándar	0,9086491

Finalmente, se presenta la convergencia de las cadenas de Markov obtenidas mediante el algoritmo MCMC para el conjunto de parámetros $\{\beta_0, \beta_1\}$ para su estacionariedad en media después de las etapas de quemado y muestreo sistemático de las observaciones tal y como se puede apreciar en el siguiente gráfico:

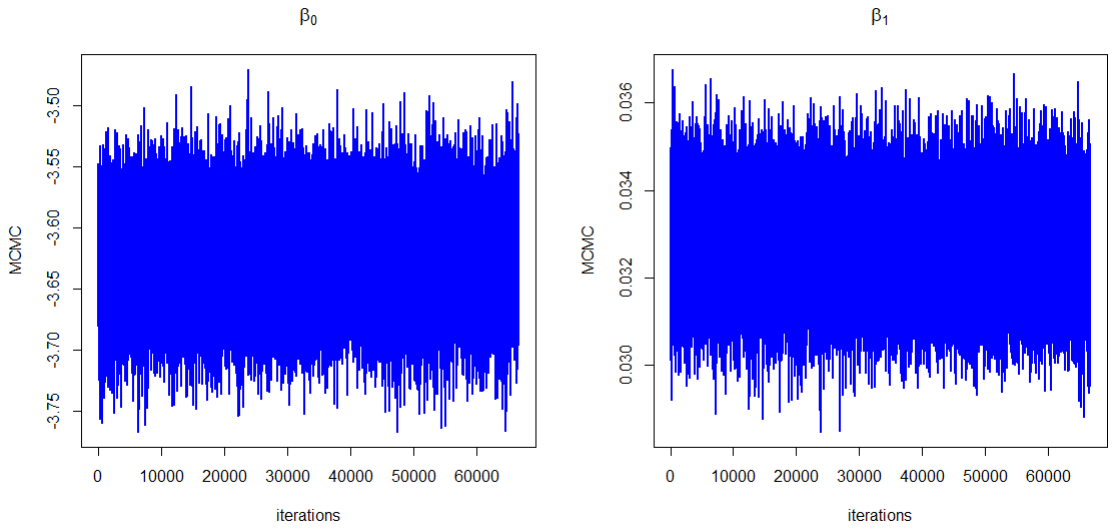

Figura 19: Cadenas de Markov para θ y k en algoritmo MCMC luego de la etapa de quemado y muestreo sistemático de las observaciones.

4.1.3. Modelo Poisson con enlace lineal de parámetros no negativos.

El gráfico presentado por medio de la ilustración 16 permitió evidenciar que la tasa de siniestralidad depende del tiempo de manera que resulta razonable proponer que para todo $i \in \{1, \ldots, n\}$:

$$\theta_i = X_i^T \boldsymbol{\beta},$$

donde cada X_i corresponde a un vector de información de dos componentes que permite observar el intercepto y el tiempo i. Del mismo modo, observar que la media y la varianza tienden al parecer a un mismo valor en cada unidad de tiempo i, permite realizar una propuesta donde el número de siniestros n_i en la semana i es tal que:

$$n_i \overset{\text{d}}{=} \text{Poisson}(\theta_i \cdot N_i),$$

donde N_i es el total de riesgos asegurados para el periodo i. Finalmente, el vector $\boldsymbol{\beta}$ de dimensión $p \times 1$ con $p = 2$, es una realización de una distribución normal p variante con vector de medias $\boldsymbol{\mu}_0$ tal que $\boldsymbol{\mu}_0 = \mathbf{0}_{p\times 1}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma_0} = s \cdot I_{p\times p}$ de manera que $s \in \boldsymbol{R}^+$ y la matriz $I_{p\times p}$ es la identidad de tamaño $p \times p$. En consecuencia:

$$\boldsymbol{\beta} \overset{\mathrm{d}}{=} \text{Normal}_p\left(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}\right).$$

Observe que, bajo una distribución previa no informativa, el valor del escalar s debe ser 'grande'. Para el caso se asume que $s^2 = 100$. Es necesario advertir que bajo la hipótesis anterior donde cada θ_i es tal que $\theta_i = X_t^T\boldsymbol{\beta}$, cada uno de los componentes del vector $\boldsymbol{\beta}$ pueden tomar valores negativos y en dicho caso, no es posible obtener realizaciones de una distribución Poisson de parámetro $\theta_i \cdot N_i$. Por lo tanto, al obtener muestras en la distribución condicional completa de $\boldsymbol{\beta}$, es necesario garantizar que sus componentes sean no negativas. A continuación se presenta la distribución condicional completa de $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta} \mid N, X) \propto \prod_{i=1}^{m} \frac{\exp\left\{\left(\boldsymbol{\beta}^T \cdot X_i\right) \cdot N_i\right\}\left(\left(\boldsymbol{\beta}^T \cdot X_i\right) \cdot N_i\right)^{n_i}}{n_i!} \cdot \prod_{k=1}^{p} \frac{1}{S \cdot \sqrt{2\pi}} \exp\left\{-\frac{1}{2S^2}\beta_k^2\right\}.$$

La anterior distribución no tiene un kernel conocido y esto implica utilizar un algoritmo de Metrópolis adaptativo bajo el enfoque de extensiones de MCMC planteado por Gelman, A. et al. (2014) para muestrear sus valores. En estos términos, resulta conveniente considerar el logaritmo de la anterior distribución condicional completa:

$$\begin{aligned}\log p(\beta \mid N, X) = \sum_{i=1}^{m} \Big[\left(\boldsymbol{\beta}^T \cdot X_i\right) \cdot N_i + n_i \cdot \log\Big(\left(\boldsymbol{\beta}^T \cdot X_i\right) \cdot N_i\Big) - \log\left(n_i!\right)\Big] \\ - \frac{p}{2}\left[\log\left(S^2\right) + \log\left(2\pi\right)\right] - \frac{1}{2S^2}\sum_{k=1}^{p}\beta_k^2.\end{aligned}$$

Bajo la restricción mencionada anteriormente para cada componente del vector $\boldsymbol{\beta}$, es necesario considerar la transformación:

$$\boldsymbol{\Phi} = \left(\log\left(\beta_0\right), \ldots, \log\left(\beta_p\right)\right),$$

de manera que al proponer muestras candidatas $\boldsymbol{\beta^*}$ de la distribución condicional completa $p(\boldsymbol{\beta} \mid N, X)$ se define $\boldsymbol{\Phi^*}$ tal que:

$$\boldsymbol{\Phi^*} \overset{\mathrm{d}}{=} \text{Normal}_p\left(\boldsymbol{\Phi}, \delta^2(X^TX)^{-1}\right).$$

En caso de aceptar la muestra candidata de la distribución $p(\boldsymbol{\beta} \mid N, X)$, se acepta el vector $\boldsymbol{\beta}^{(s)}$ en la s-ésima iteración del algoritmo MCMC con un Metrópolis adapatativo haciendo:

$$\boldsymbol{\beta^{(s)}} = \left(\exp(\Phi_0^*), \ldots, \exp(\Phi_p^*)\right).$$

Implementando la aplicación anterior para generar vectores $\boldsymbol{\beta}$ de componentes no negativos, se computa el logaritmo de la distribución condicional completa de $\boldsymbol{\Phi}$ que en términos de $p(\boldsymbol{\beta} \mid X, N)$ por el teorema de la transformación se expresa de la siguiente forma:

$$\log\left(\boldsymbol{\Phi} \mid X, N\right) = \sum_{k=1}^{p} \exp\{\Phi_p\} + \log p\left(\beta = \left(\exp\{\Phi_0\}, \ldots, \exp\{\Phi_p\}\right) \mid N, X\right).$$

En estos términos del logaritmo de la razón de aceptación r en el algoritmo de Metrópolis adaptativo para un vector candidato de la distribución condicional completa $p(\beta \mid N, X)$ se expresa de la siguiente manera:

$$\begin{aligned}\log\,(r) = &\textstyle\sum_{k=1}^{p} \exp\{\Phi_p^*\} + \log\, p\left(\beta = \left(\exp\{\Phi_0^*\}, \ldots, \exp\{\Phi_p^*\}\right) \mid N, X\right) \\ &- \textstyle\sum_{k=1}^{p} \exp\{\Phi_p\} - \log\, p\left(\beta = (\exp\{\Phi_0\}, \ldots, \exp\{\Phi_p\}) \mid N, X\right).\end{aligned}$$

Los pasos para el algoritmo MCMC que permitan obtener muestras del vector $\boldsymbol{\beta}$ son:

1. En la iteración inicial, guardar $\boldsymbol{\beta}^{(0)}$ como:

$$\boldsymbol{\beta}^{(0)} \leftarrow (0.01, \ldots, 0.01)_{1\times p},$$

2. Optimizar el valor de δ^2 de manera que la razón de aceptación se encuentre entre el 20 % y el 50 %.
3. Para la iteración del paso s con $s \geq 1$, realizar el siguiente procedimiento:
   a) Guardar $\boldsymbol{\Phi^{(s-1)}}$ como:

$$\boldsymbol{\Phi^{(s-1)}} \leftarrow \left(\log\,\left(\beta_0^{(s-1)}\right), \ldots \log\,\left(\beta_p^{(s-1)}\right)\right).$$

   b) Muestrear un vector $\boldsymbol{\Phi}^*$ tal que $\boldsymbol{\Phi}^* \overset{\mathrm{d}}{=} \text{Normal}_p\left(\boldsymbol{\Phi}^{(s-1)}, \delta^2 (X^T X)^{-1}\right)$.
   c) Guardar los vectores $\boldsymbol{E}^*$ y $\boldsymbol{E}^{(s-1)}$ de la siguiente manera:

$$\begin{aligned}\boldsymbol{E}^* &\leftarrow \left(\exp(\Phi_0^*, \ldots, \exp(\Phi_p^*)\right) \\ \boldsymbol{E}^{(s-1)} &\leftarrow \left(\exp(\Phi_0^{(s-1)}), \ldots, \exp(\Phi_p^{(s-1)})\right).\end{aligned}$$

   d) Computar el logaritmo de la razón de aceptación r_s en su s-ésima iteración de manera que:

$$\begin{aligned}\log r_s = &\textstyle\sum_{i=1}^{m} \left[(\boldsymbol{E^{*}}^T \cdot X_i) \cdot N_i + n_i \cdot \log\,\left((\boldsymbol{E^{*}}^T \cdot X_i) \cdot N_i\right) - \log\,(n_i!)\right] \\ &- \textstyle\sum_{i=1}^{m} \left[(\boldsymbol{E^{(s-1)}}^T \cdot X_i) \cdot N_i + n_i \cdot \log\,\left((\boldsymbol{E^{(s-1)}}^T \cdot X_i) \cdot N_i\right) - \log\,(n_i!)\right] \\ &+ \frac{1}{2S^2} \textstyle\sum_{k=1}^{p} \left(E_p^{(s-1)}\right)^2 - \frac{1}{2S^2} \sum_{k=1}^{p} (E_p^*)^2 + \sum_{i=1}^{m} \left((E_p^*) - E_p^{(s-1)}\right).\end{aligned}$$

   e) Muestrear un número aleatorio u_s tal que $u_s \overset{\mathrm{d}}{=} \text{Uniforme}(0, 1)$ y con este condicionar el estado actual del vector $\boldsymbol{\beta}^{(s)}$:

$$\boldsymbol{\beta}^{(s)} \leftarrow \begin{cases} \left(\exp\{\Phi_0^*\}, \ldots, \exp\{\Phi_p^*\}\right) & \text{, si } \log(u_s) \leq \log(r_s) \\ \boldsymbol{\beta}^{(s-1)} & \text{, en otro caso.} \end{cases}$$

4. Generar los pasos 2 y 3 en S iteraciones, de manera que $S = 210.000$
5. Quemar las primeras 10.000 iteraciones y muestrear las observaciones de manera sistemática con salto $a = 3$.

Posterior a la implementación del algoritmo MCMC presentando anteriormente se obtuvieron los siguientes resultados con respecto a tamaños efectivos de muestra, medias posteriores e intervalos de credibilidad del 95 % para los parámetros del modelo:

En término del buen ajuste del modelo se obtuvieron los siguientes resultados de los valores p predictivos para estadísticos como media, varianza, desviación estándar y coeficiente de variación:

Tabla 7: *La tabla reporta en su primera columna los parámetros del modelo de regresión Poisson, la segunda columna reporta el tamaño efectivo de muestra en el algoritmo MCMC, la tercera y quinta columna presentan los límites inferior y superior de los intervalos de credibilidad del* 95 %*, mientras que la cuarta columna proporciona la media posterior.*

Parámetro	Tamaño Efectivo	Límite inferior Intervalo de Credibilidad 95 %	Media Posterior	Límite superior Intervalo de Credibilidad 95 %
β_0	868,3094	0,00916522	0,011638502	0,002150619
β_1	450,3019	0,01418117	0,002269835	0,002392112

Tabla 8: *La primera columna muestra el estadístico de bondad de ajuste para el modelo Bayesiano Binomial Negativo para la frecuencia de los siniestros y la segunda columna su respectivo valor p predictivo.*

Estadístico	PPI
Media	0,4868149
Varianza	0,7764678
Coeficiente de variación	0,8024330
Desviación estándar	0,7764678

Los resultados proporcionados por la tabla anterior muestran el 'buen' ajuste que se tiene con relación a la media, así como mejores indicadores de 'bondad' para varianza, desviación típica y coeficiente de variación en comparación con el modelo lineal generalizado de Poisson presentado previamente.

En este modelo, una vez pasadas las etapas de quemado y muestreo sistemático de las observaciones, se observa estacionariedad en las cadenas de Markov de las muestras posteriores para los parámetros, aunque para β_1 parece requerirse de más observaciones debido al tamaño efectiva de muestra:

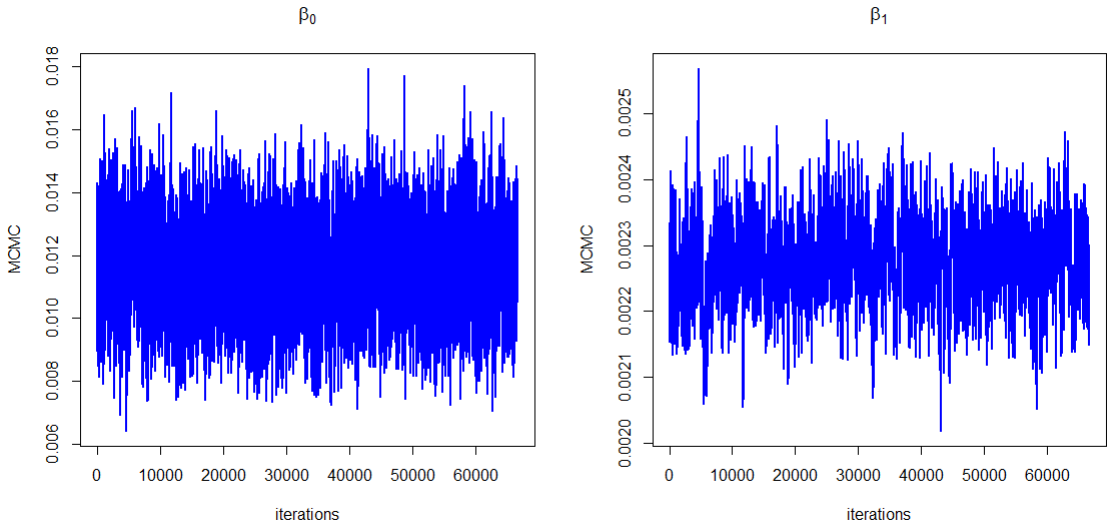

Figura 20: Cadenas de Markov para θ y k en algoritmo MCMC luego de la etapa de quemado y muestreo sistemático de las observaciones.

4.2. Estudio de sensibilidad para modelos Bayesianos de frecuencia de los siniestros.

El estudio de sensibilidad tiene como propósito evaluar algunas métricas de control de calidad en el pronóstico de valores perdidos cuando se propone una variación en los hiperparámetros de la distribución previa. Estos valores perdidos se generan cuando se implementa una validación cruzada conocida como k-fold. Escogiendo un valor para $k \in \mathbb{Z}^+$ con $k \geq 2$, se generan k particiones del "mismo"tamaño mediante un muestreo aleatorio simple sobre el conjunto de observaciones, de manera que al estimar los parámetros de cada modelo Bayesiano en un ensamble de $k-1$ particiones, se realiza la predicción para las observaciones de la partición restante. En consecuencia, por cada partición se genera una predicción y por ende, una métrica para evaluar la calidad del pronóstico. La media de las métricas asociadas a cada una de las particiones resulta ser el indicador que permite evaluar el potencial predictivo para los modelos Bayesianos propuestos. En el presente estudio de sensibilidad, se utilizará una validación cruzada de k-fold como la propuesta por Gelman, A. et al. (2014) con $k = 5$. Las métricas de evaluación en cada partición son: el error porcentual absoluto medio (MAPE), la raíz cuadrada del error cuadrático medio (RMSE), logaritmo de la densidad predictiva puntual (lppd) y la tasa de predicciones que se desvían porcentualmente a lo sumo en un umbral respecto a su valor real. En esta última métrica U, una vez definido un umbral u tan 'pequeño' como se quiera con $u > 0$, de manera que:

$$U(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}, u) = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

donde n es el número de observaciones predichas, $\boldsymbol{Y}$ es el conjunto de observaciones, $\widehat{\boldsymbol{Y}}$ son sus respectivas predicciones y cada X_i con $i \in \{1, \ldots, n\}$ es una función indicadora definida de la siguiente manera:

$$X_i = \begin{cases} 1 & , \text{ si } \left| \frac{\widehat{y}_i - y_i}{y_i} \right| \leq u \\ 0 & , \text{ en otro caso.} \end{cases}$$

Observe que el valor de esta métrica existe en $\mathbb{R}$ si y solamente si, el conjunto de observaciones $\boldsymbol{Y}$ no contiene el valor cero. A continuación se presentan los diferentes resultados del indicador MAPE cuando se realiza la validación cruzada de k-fold con $k = 5$ para el ajuste del modelo binomial negativo sobre la frecuencia del número de siniestros ocurridos por semana cuando los hiperparámetros de la distribución previa son tales que: $\alpha, \beta \in \{0.5, 1.0, 1.5, 2.0\}$ y $\alpha_0 \in \{0.5, 1, 1.5\}$.

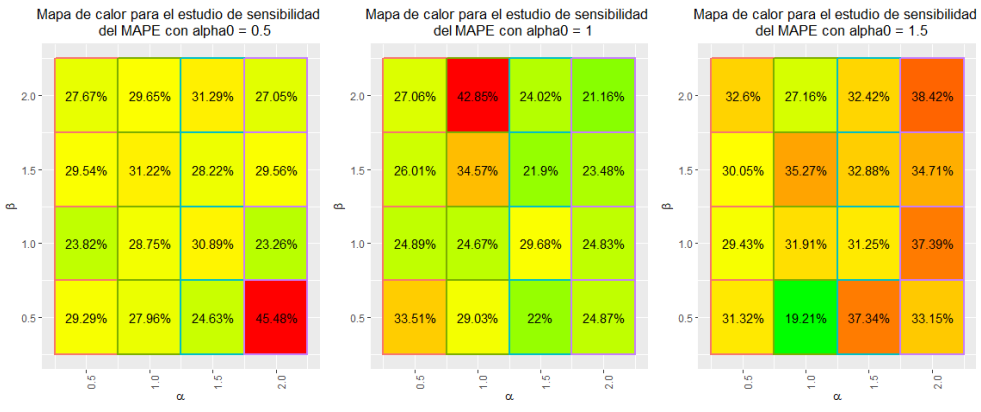

Figura 21: Resultados de la métrica del MAPE en estudio de sensibilidad de hiperparámetros para distribución previa de modelo Binomial Negativo. El algoritmo MCMC fue implementado con 210.000 iteraciones quemando las primeras 10.000 muestras y haciendo posteriormente, un muestreo sistemático con salto igual a 3.

El gráfico anterior permite evidenciar que utilizando la distribución previa no informativa propuesta para el modelo binomial Negativo con un algoritmo MCMC desarrollado bajo un algoritmo adaptativo de Metrópolis con $\alpha_0 = 1$, $\alpha = 1$ y $\beta = 1$, el MAPE alcanza un valor del $24,67\,\%$. Esto indica que en promedio, las predicciones se desvían porcentualmente con respecto a las observaciones en un $24,67\,\%$.

Al establecer una comparación entre el valor en mención y los demás resultados de la métrica cuando se presenta una variación en los valores de los hiperparámetros se observa que las desviaciones de las predicciones con respecto a las observaciones son más altas y oscilan en promedio entre el $30\,\%$ y el $45\,\%$ cuando alguno de los hiperparámetros comienza a ser mayor que 1.

Con respecto a la raíz del error cuadrático medio se observa a continuación que para la distribución previa propuesta, en promedio las predicciones se desvían del valor verdadero de las observaciones en $28,63$ siniestros por semana. En el caso de la métrica RMSE se observa que esta se minimiza cuando $\alpha = 1$, $\beta = 0.5$ y $\alpha_0 = 1.5$.

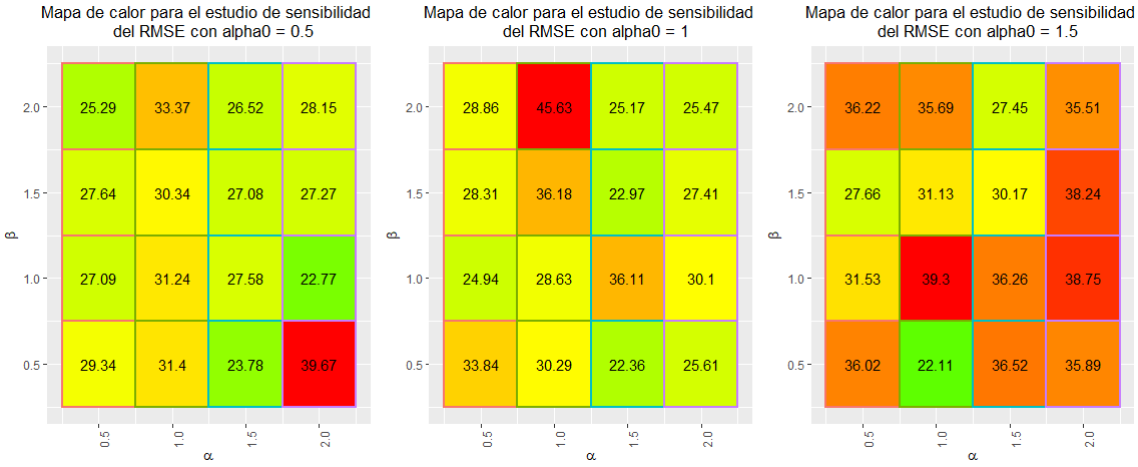

Figura 22: Resultados de la métrica del RMSE en estudio de sensibilidad de hiperparámetros para distribución previa de modelo Binomial Negativo. El algoritmo MCMC fue implementado con 210.000 iteraciones quemando las primeras 10.000 muestras y haciendo posteriormente, un muestreo sistemático con salto igual a 3.

La evaluación de la métrica del logaritmo predictivo en la densidad puntual (lppd) en el modelo Binomial Negativo propuesto mostró un contraste importante con respecto a los resultados de las métricas del MAPE y el RMSE porque para cada una de las distribuciones previas evaluadas, los resultados de la métrica lppd se concentraron entre -52 y -53. La interpretación del resultado expuesto anteriormente sugiere que en términos de la bondad de ajuste, al parecer no existen grandes diferencias entre las distribuciones previas no informativas propuestas para el modelo. A continuación, se presentan los resultados del estudio de sensibilidad para la distribución previa utilizando lppd:

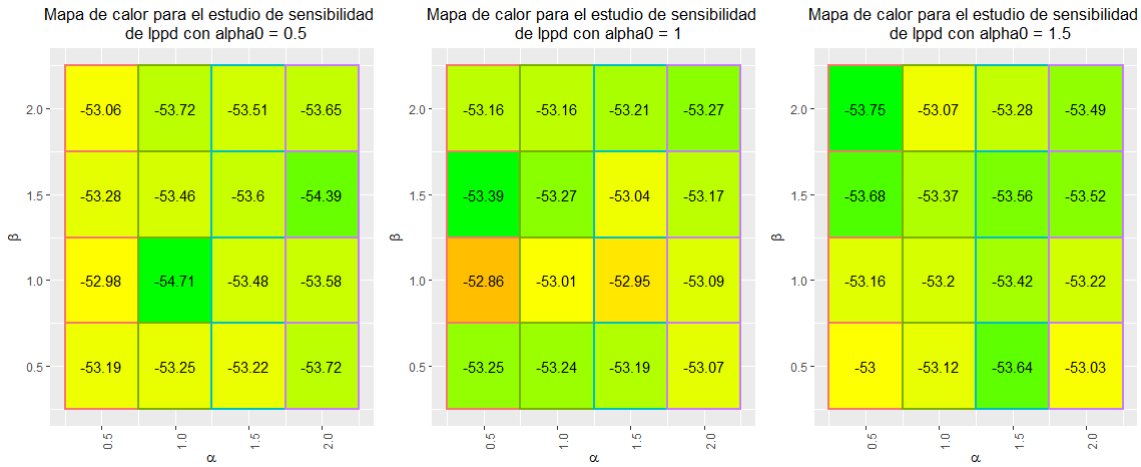

Figura 23: Resultados de la métrica lppd en estudio de sensibilidad de hiperparámetros para distribución previa de modelo Binomial Negativo. El algoritmo MCMC fue implementado con 210.000 iteraciones quemando las primeras 10.000 muestras y haciendo posteriormente, un muestreo sistemático con salto igual a 3.

Finalmente, se reportan los resultados de la métrica de tasa de predicciones que se desvían en a lo sumo un umbral del 20 % con respecto a su valor real, para evaluar la sensibilidad de la distribución previa:

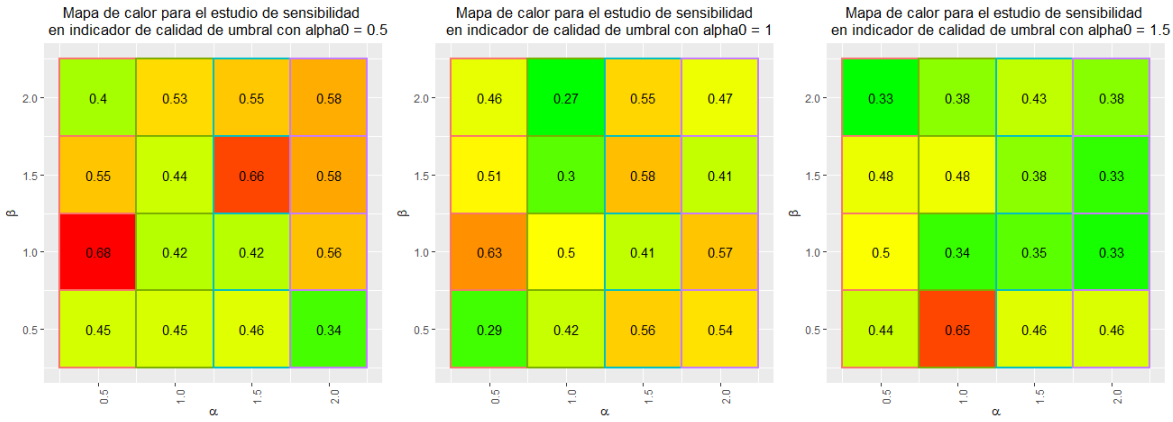

Figura 24: Resultados de la métrica de tasa de predicciones que se desvían en a lo sumo un 20 % de su valor real en estudio de sensibilidad de hiperparámetros para distribución previa de modelo Binomial Negativo. El algoritmo MCMC fue implementado con 210.000 iteraciones quemando las primeras 10.000 muestras y haciendo posteriormente, un muestreo sistemático con salto igual a 3.

Esta métrica evalúa con mayor precisión que el MAPE o el RMSE la predicción puntual, dado que verifica si cada pronóstico se desvía a lo sumo en un 20 % de su valor real y al final, determina el porcentaje de observaciones que cumplen con este criterio. De esta manera, la calidad del pronóstico para esta métrica es bastante informativa porque proporciona una interpretación del potencial predictivo de un modelo, puesto que no se basa únicamente en minimizar un error sino en mostrar en términos porcentuales la tasa de observaciones "bien"predichas con respecto a un umbral de desvío tan pequeño como se quiera.

El resumen del estudio de sensibilidad vía validación cruzada del modelo Binomial Negativo mostró que con la distribución previa de valores puntuales para $\alpha = 1, \beta = 1$ y $\alpha_0 = 1$, se obtuvieron valores promedio para MAPE y RMSE del $24,67\,\%$ y $28,63$ siniestros. Por otra parte, en promedio solamente el $50\,\%$ de las observaciones fueron "bien"predichas porque se desviaron de su valor real en a lo sumo un

20 %. Sin embargo, el mejor modelo predictivo para minimizar el RMSE y el MAPE toma los siguientes valores para la distribución previa: $\alpha = 1, \beta = 0,5$ y $\alpha_0 = 1,5$, de manera que la tasa de observaciones "bien"predichas es del 65 %.

Para los modelos lineal generalizado de Poisson y con link lineal de coeficientes no negativos, a continuación, se presentan los resultados de sensibilidad para las métricas de MAPE, RMSE, lppd y tasa de observaciones que se desvían en a lo sumo un 20 % de su valor real:

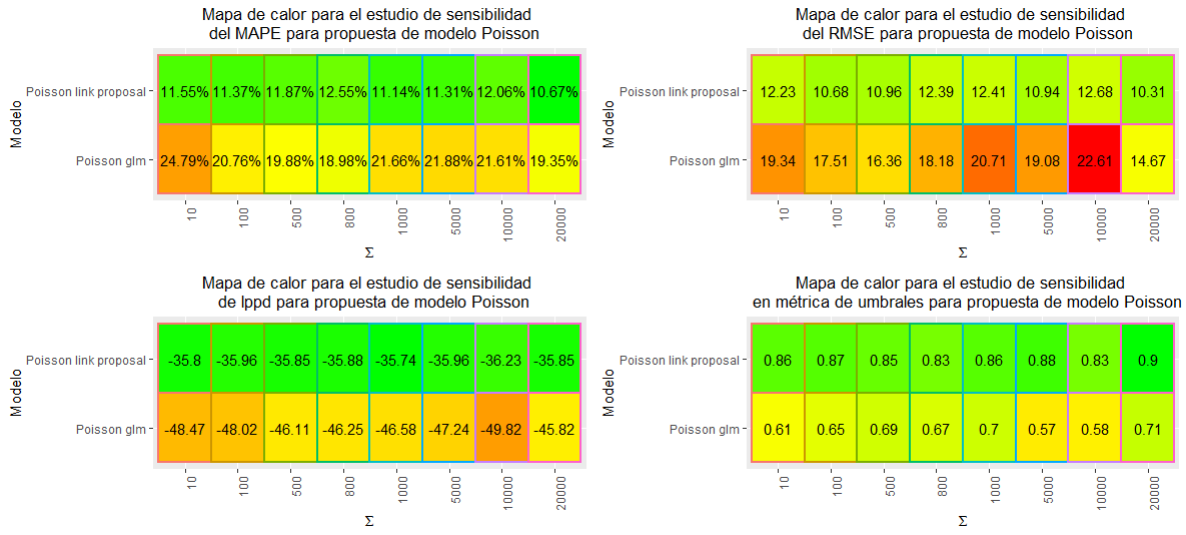

Figura 25: Resultados de estudio de sensibilidad para modelos lineal generalizado de Poisson y con link lineal de coeficientes positivos cuando se establece la variación del hiperparámetro S^2. El algoritmo MCMC fue implementado con 210.000 iteraciones quemando las primeras 10.000 muestras y haciendo posteriormente, un muestreo sistemático con salto igual a 3.

Para la distribución previa del modelo lineal generalizado de Poisson con $S^2 = 1000$, las métricas del MAPE, el RMSE y lppd son respectivamente del $21,66$ %, $20,71$ siniestros y $-46,58$. Es este sentido, las métricas de calidad del pronóstico son mejores en comparación el modelo binomial negativo. Adicionalmente, en promedio el 70 % de las observaciones son bien predichas porque se desvían en a lo sumo un 20 % de su valor real. Sin embargo, el modelo de enlace lineal con coeficientes no negativos con distribución previa de hiperparámetro $S^2 = 100$ mostró mejores métricas que los dos modelos mencionados anteriormente alcanzando un 87 % de observaciones "bien"predichas que se desvían en a lo sumo un 20 % de su valor real. Con respecto a la sensibilidad cuando se establece la variación de S^2, se observaron resultados estables muy concentrados alrededor de un mismo punto en cada métrica para cada uno de los modelos.

Pese a que los modelos lineales generalizado de Poisson y de enlace lineal con coeficientes no negativos mostraron un mejor desempeño con respecto a la calidad del pronóstico, el modelo lineal generalizado de Poisson presenta problemas en los valores p predictivos para estadísticos como la varianza, la desviación estándar y el coeficiente de variación, mientras que el modelo de enlace lineal tiende a resolver en cierto modo esta problemática. En términos del valor p predictivo para la media, ambos modelos mostraron un ajuste óptimo.

Finalmente, la comparación entre los modelos se establece por medio de los criterios WAIC y DIC. A medida que los valores de estos indicadores son menores, mejor es el modelo para explicar la frecuencia de los siniestros ocurridos. Teniendo en cuenta los criterios Bayesianos en mención, el mejor modelo es la propuesta de enlace lineal con coeficientes no negativos según el criterio DIC dado que DIC $= 359,178$; mientras que en términos del criterio WAIC, el mejor modelo es el binomial negativo de manera que WAIC $= -265.9869$. Los resultados expuestos se presentan por medio de la siguiente tabla:

Tabla 9: *La primera columna reporta el modelo propuesto para la frecuencia de los siniestros ocurridos por semana, y la segunda y tercera columna muestran el valor del WAIC y DIC, respectivamente.*

Modelo	WAIC	DIC
Binomial Negativo	-265,9869	531,6273
Poisson glm	-232,4517	461,3098
Enlace lineal de coeficientes no negativos	-179,1122	359,178

5. Modelo Bayesiano para la severidad de los siniestros ocurridos.

El análisis descriptivo de la distribución de la cuantía de los siniestros mostró que estas son asimétrica y posiblemente presenta al menos dos modas. Debido a lo anterior, la propuesta de modelos bayesianos de mezcla y una distribución log normal parece ser apropiada no solo por la asimetría sino por las recomendaciones y supuestos mencionados en el marco teórico con respecto a la Directiva de Solvencia II para la cuantía de los siniestros en entidades aseguradoras.

5.1. Propuesta de modelos Bayesianos para la severidad de siniestros.

5.1.1. Modelo Bayesiano de mezcla con distribución lognormal con número de componentes fijo.

La cuantía de los siniestros en una distribución asimétrica con al menos dos modos presenta un 'alto' nivel de heterogeneidad si se considera su coeficiente de variación. Sin embargo, al construir clases latentes, segmentos o clústeres de manera que la cuantía de los siniestros tiende a cierto nivel de homogeneidad dentro de estos grupos, resulta razonable una propuesta para modelar la variable aleatoria X de severidad de los siniestros según su clase latente z_h con $h \in \{1, \ldots, H\}$, $H \in \mathbb{Z}^+$ y $H \geq 2$:

$$x_i \mid z_i \overset{\text{iid}}{=} \text{lognormal}(x_i \mid \theta_h, \sigma_h^2).$$

Este modelo de verosimilitud indica que la cuantía de los siniestros según su clase latente z_h, es realización de una distribución log normal de vector de parámetros $\boldsymbol{\Theta} = (\theta_h, \sigma_h^2)$. En este caso, la distribución lognormal es una propuesta hecha por la Directiva de Solvencia II para la pérdida generada por causa de los siniestros ocurridos en compañías aseguradoras al momento de determinar el riesgo de suscripción y su respectivo requerimiento de capital en compañías aseguradoras. En conclusión, esta propuesta permite estudiar la severidad o cuantía de los siniestros como una 'composición' de funciones de densidad de distribuciones log normales que consideran segmentos o clases latentes para las cuantías, de manera que las clases tienden a agrupar los siniestros según algún tipo de homogeneidad en su cuantía. Por otra parte, las clases latentes se encuentran en función de proporciones que en la práctica son aleatorias dado que existe incertidumbre en el negocio asegurador sobre las cuantías que tendrán los siniestros una vez se presente su ocurrencia. De este modo, no es posible determinar si el siguiente año los siniestros de 'menor' cuantía se presentarán con menos frecuencia que siniestros de 'mayor' cuantía. En consecuencia, si las clases latentes se asocian con proporciones o pesos, la propuesta de modelo de verosimilitud para estas considera una distribución multinomial en caso de tener al menos tres segmentos:

$$z_i \mid w \overset{\text{iid}}{=} \text{Multinomial}(1; \omega_1, \omega_2, \ldots, \omega_H).$$

Por tanto, las clases latentes asociadas a cada siniestro son realización de una distribución multinomial cuyo vector de parámetros es un vector de pesos o proporciones $\boldsymbol{\omega}$ tal que $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_h)$. Observe que

en caso de tener dos clases o segmentos $\{Z_1, Z_2\}$ para agrupar las cuantías de los siniestros, estas son realización de una distribución binomial; en particular, una distribución Bernouilli de parámetro p. De ese modo, un siniestro se asocia a la clase Z_1 con probabilidad p y a la clase Z_2 con probabilidad $1-p$.

Para concluir la propuesta de un modelo Bayesiano de mezcla sobre la cuantía de los siniestros, es necesario explicar como obtener realizaciones de $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_h)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_h)$ y $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_h^2)$. En este caso, la ciencia estadística propone una distribución Dirichlet como previa para un conjunto de parámetros de una distribución multinomial:

$$\boldsymbol{\omega} \overset{\mathrm{d}}{=} \text{Dirichlet}(\alpha, \alpha, \ldots, \alpha).$$

Note que el vector de pesos de las clases latentes de los siniestros es realización de una distribución Dirichlet en caso de que existan al menos tres clústeres de agrupación. Sin embargo, en caso de contar únicamente con dos clases, la probabilidad p de que un siniestro se encuentre en una de las clases será realización de una distribución Beta.

Con respecto a los parámetros de media θ_h y varianza σ_h^2 de cada distribución lognormal por clase latente z_h, se observa que el soporte θ_h es $\mathbb{R}$ y de σ_h^2 es $\mathbb{R}^+$. Por lo anterior, una distribución normal es posible candidata para obtener realizaciones de θ_h; mientas que una distribución inversa gamma es posible candidata para obtener realizaciones de σ_h^2:

$$\begin{aligned}\theta_h &\overset{\mathrm{d}}{=} \text{Normal}(\mu_h, \tau_h^2)\\ \sigma_h^2 &\overset{\mathrm{d}}{=} \text{Inv-Gamma}(a_h, b_h).\end{aligned}$$

Con base en la propuesta de modelo Bayesiano de mezcla con distribución log normal presentado anteriormente con número de componentes fijo H y $H \geq 2$, el conjunto de parámetros está dado por: $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_H)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H)$ y $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_H^2)$. Por otra parte, el conjunto de hiperparámetros está dado por: α, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_H)$, $\boldsymbol{\tau} = (\tau_1^2, \ldots, \tau_H^2)$, $\boldsymbol{a} = (a_1, \ldots, a_H)$ y $\boldsymbol{b} = (b_1, \ldots, b_H)$.

Distribución previa no informativa para el modelo.

Dado que no se cuenta con información previa sobre el conjunto de hiperparámetros α, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_H)$, $\boldsymbol{\tau} = (\tau_1^2, \ldots, \tau_H^2)$, $\boldsymbol{a} = (a_1, \ldots, a_H)$ y $\boldsymbol{b} = (b_1, \ldots, b_H)$, se presentan a continuación algunas propuestas no informativas para el modelo Bayesiano de mezcla con número de componentes fijo:

1. Como no se conoce información sobre las proporciones de las clases que agrupan los siniestros, se considera α tal que $\alpha = 1$. El vector $\boldsymbol{\mu}$ considera la media muestral del logaritmo de la cuantía de los siniestros por cada clase. Por otra parte, $\boldsymbol{\tau}$ es un vector donde cada una de sus componentes toma el valor de 1. En este caso, la distribución es no informativa dado que 1 es una desviación 'grande' a nivel de escala logarítmica de manera que los montos reales de los siniestros son 'altamente' sensibles a esta variabilidad.

   Finalmente, los vectores de $\boldsymbol{a}$ y $\boldsymbol{b}$ de componentes a_h y b_h respectivamente, se asocian con los parámetros de una distribución inversa gamma donde:

$$\begin{aligned}\mathbb{E}(\sigma_h^2) &= \frac{b_h}{a_h - 1} = E_{0,h}\\ \mathbb{V}\text{ar}(\sigma_h^2) &= \frac{b_h^2}{(a_h-1)^2 \cdot (a_h - 2)}.\end{aligned}$$

   De esta manera, un valor no informativo para a_h es tal que $a_h = 2$ dado que si $a_h \to 2$ entonces $\mathbb{V}\text{ar}(\sigma_h^2) \to \infty$. En virtud de lo anterior, $b_h = E_{0,h}$. El valor esperado de la varianza de los siniestros

en cada grupo $E_{0,h}$ tiende a ser 'pequeño' si se considera que cada clase latente agrupa siniestros de cuantías homogéneas. Sin embargo, no necesariamente se tiene homogeneidad de varianza porque los grupos o clases latentes con mayores cuantías respecto a una unidad monetaria presentan en el negocio asegurador mayor variabilidad que grupos con menores cuantías. De esta manera, se propone un valor tan 'pequeño' como se quiera para $E_{0,h=1}$, en este caso $E_{0,h=1}$ es tal que $E_{0,h=1} = 0,09$. En los demás grupos, se tiene una varianza proporcional a $0,09$ de manera que:

$$E_{0,h} = 0,09 \cdot h.$$

En consecuencia, la distribución previa no informativa para el modelo es tal que:

$$\begin{aligned}
&\alpha = 1\\
&\boldsymbol{\mu} = (\mu_1 = \bar{y_{z_1}}, \ldots, \mu_H = \bar{y_{z_h}})\\
&\boldsymbol{\tau} = (\tau_1^2 = 1, \ldots, \tau_H^2 = 1)\\
&\boldsymbol{a} = (a_1 = 0.09, \ldots, a_h = 0.09h, \ldots, a_H = 0.09H)\\
&\boldsymbol{b} = (b_1 = 2, \ldots, b_H = 2),
\end{aligned}$$

donde $\bar{y_{z_h}}$ se define de la siguiente manera para todo $h \in \{1, \ldots, H\}$:

$$\bar{y_{z_h}} = \frac{1}{n_h} \textstyle\sum_{i=1}^{n_h} \log(x_{i,h}).$$

Cada $x_{i,h}$ satisface que $x_{i,h} \in Z_h$ y $n_h = \#\{x : x \in Z_h\}$.

2. En esta propuesta, α, $\boldsymbol{\mu}$ y $\boldsymbol{\tau}$ son equivalentes a la propuesta del numeral 1. La variante se presenta para los vectores $\boldsymbol{a}$ y $\boldsymbol{b}$. De esta manera, un escenario de alta variabilidad para cada $\boldsymbol{\sigma_h^2}$ se tiene cuando el coeficiente de variación $\mathbb{C}\mathrm{V}(\sigma_h^2)$ es 'grande' o igual al $100\,\%$. El coeficiente de variación para σ_h^2 cuando $\sigma_h^2 \overset{\mathrm{d}}{=} \text{Inv-Gamma}(a_h, b_h)$ es tal que:

$$\mathbb{C}\mathrm{V}(\sigma_h^2) = (a_h - 2)^{-\frac{1}{2}}.$$

Por tanto, otra propuesta no informativa para cada a_h con $h \in \{1, \ldots, H\}$ se tiene cuando se hace $a_h = 3$. El valor de cada b_h se obtiene de la siguiente manera, bajo el supuesto de valor esperado de σ_h^2 hecho en el numeral anterior:

$$b_h = h \cdot (a_h - 1) \cdot E_{0,h} = 2h \cdot E_{0,h}.$$

Teniendo en cuenta la posible homogeneidad de cada clase latente y la heterocedasticidad mencionada anteriormente sobre segmentos que agrupan de forma ordinal la cuantía de los siniestros en el negocio del sector asegurador, el valor de cada b_h es tal que $b_h = 2 \cdot h \cdot E_{0,h}$, donde $E_{0,h}$ es un valor tan 'pequeño' como se quiera, en este caso $E_{0,h} = 0,09$. Así pues, $b_h = 0,18 \cdot h$.

3. La tercera propuesta conserva los mismos valores definidos de las propuestas 1 y 2 para α y los vectores $\boldsymbol{\mu}, \boldsymbol{\tau}$. Nuevamente, en otro escenario de 'alta' variabilidad para σ_h^2, un coeficiente de variación 'grande' se tiene cuando $\mathbb{C}\mathrm{V}(\sigma_h^2) = 70\,\%$. De este modo α_h es tal que α_h tal que $(\alpha_h - 2)^{-\frac{1}{2}} = 0,7$. En consecuencia $\alpha_h = 4,040816$. Sin embargo, por practicidad se escoge $a_h = 4$ para cada $h \in \{1, \ldots, H\}$. Considerando el valor esperado $E_{0,h}$ de σ_h^2 cuando se distribuye inversa gamma de parámetros a_h, b_h, se tiene que $b_h = 3 \cdot E_{0,h}$. Con un valor 'pequeño' para $E_{0,h}$ como $E_{0,h} = 0,09$ y considerando la heterogeneidad de varianza de grupos proporcional a h se tiene que:

$$b_h = 0,27 \cdot h.$$

El objetivo para la modelación es implementar las tres propuestas no informativas mencionadas anteriormente.

Algoritmo MCMC.

Para cada una de las propuestas hechas anteriormente sobre la distribución previa no informativa en el modelo Bayesiano de mezcla de distribución log normal con número de componentes fijo, se implementa un algoritmo MCMC con los siguientes pasos:

1. En la iteración inicial, guardar una realización de θ_h, σ_h^2 para todo $h \in \{1, \ldots, H\}$ considerando sus distribuciones previas:

$$\begin{aligned} \theta_h^{(0)} &\leftarrow \text{Normal}(\mu_h, \tau_h^2) \\ (\sigma_h^2)^{(0)} &\leftarrow \text{Inv-Gamma}(a_h, b_h) \\ \omega_h^{(0)} &\leftarrow \frac{n_h}{\sum_{j=1}^{H} n_j}, \end{aligned}$$

   donde n_h es el total de elementos de la clase h para todo $h \in \{1, \ldots, H\}$. Observe que las clases son obtenidas inicialmente mediante la implementación del algoritmo EM.

2. Para la iteración del paso s con $s \geq 1$, realizar el siguiente procedimiento:

   - Computar las probabilidades:

$$p\left(Z_i^{(s)} = h \mid \text{resto}\right) = \frac{\omega_h^{(s-1)} \text{lognormal}\left(x_i \mid \theta_h^{(s-1)}, (\sigma_h^2)^{(s-1)}\right)}{\sum_{l=1}^{H} \omega_l^{(s-1)} \text{lognormal}\left(x_i \mid \theta_l^{(s-1)}, (\sigma_l^2)^{(s-1)}\right)},$$

     para todo $i \in \{1, \ldots, n\}$ y $h \in \{1, \ldots, H\}$.

   - Guardar m_i tal que:

$$m_i \leftarrow \text{máx}\left\{p\left(Z_i^{(s)} = l \mid \text{resto}\right) : l \in \{1, \ldots, H\}\right\}$$

     para todo $i \in \{1, \ldots, n\}$.

   - Para optimizar computacionalmente las probabilidades de pertenecer a cada clúster al momento de muestrear la clase latente de cada observación, calcular:

$$p\left(Z_{i,*}^{(s)} = h \mid \text{resto}\right) = \frac{\exp\left\{p\left(Z_i^{(s)} = h \mid \text{resto}\right) - m_i\right\}}{\sum_{h=1}^{H} \exp\left\{p\left(Z_i^{(s)} = h \mid \text{resto}\right) - m_i\right\}},$$

     para todo $i \in \{1, \ldots, n\}$ y $h \in \{1, \ldots, H\}$.

   - Muestrear la clase latente de cada x_i con el vector de probabilidades $\boldsymbol{p}_i$:

$$\boldsymbol{p}_i = \left(p\left(Z_{i,*}^{(s)} = 1 \mid \text{resto}\right), \ldots, p\left(Z_{i,*}^{(s)} = H \mid \text{resto}\right)\right).$$

     para todo $i \in \{1, \ldots, n\}$. Paso seguido al anterior, se guarda $\mathbb{N}_h^{(k)}$ tal que:

$$\mathbb{N}_h^{(s)} \leftarrow \left(n_1^{(s)}, \ldots, n_H^{(s)}\right),$$

     donde $n_h^{(s)}$ es el tamaño o cardinal del grupo h en la iteración s, para todo $h \in \{1, \ldots, H\}$.

- Muestrear un vector $\boldsymbol{\omega}^{(s)}$ | resto tal que:

$$\boldsymbol{\omega}^{(s)} \mid \text{resto} \leftarrow \text{Dirichlet}(n_1 + \alpha, \ldots, n_H + \alpha).$$

- Muestrar una observación $\theta_h^{(s)}$ | resto tal que:

$$\theta_h^{(s)} \mid \text{resto} \leftarrow \text{Normal}\left(\widehat{m}_{h,(s-1)}\widehat{V}_{h,(s-1)}^{-1}, \widehat{V}_{h,(s-1)}^{-1}\right),$$

donde

$$\widehat{m}_{h,(s-1)} = \frac{\sum_{i:Z_i \in h} \log(x_i)}{(\sigma_h^2)^{(s-1)}} + \frac{\mu_h}{\tau_h^2}$$
$$\widehat{V}_{h,(s-1)} = \frac{n_h^{(s)}}{(\sigma_h^2)^{(s-1)}} + \frac{1}{\tau_h^2},$$

para todo $h \in \{1, \ldots, H\}$.

- Muestrear una observación para $(\sigma_h^2)^{(s)}$ | resto tal que:

$$(\sigma_h^2)^{(s)} \mid \text{resto} \leftarrow \text{Inv-Gamma}\left(\frac{n_h^{(s)}}{2} + a_h, \frac{1}{2}\sum_{i:Z_i \in h}\left(\log(x_i) - \theta_h^{(s)}\right)^2 + b_h\right),$$

para todo $h \in \{1, \ldots, H\}$.

3. Guardar el vector de parámetros $\boldsymbol{\Theta}^{(s)}$ tal que:

$$\boldsymbol{\Theta}^{(s)} \leftarrow \left(\theta_1^{(s)}, \ldots, \theta_H^{(s)}, (\sigma_1^2)^{(s)}, \ldots, (\sigma_H^2)^{(s)}, \omega^{(s)}\right).$$

El paso 2 del algoritmo anterior se hace en S iteraciones. Para el caso, se asume $S = 210.000$. Posteriormente se realiza una etapa de quemada de las primeras 10.000 observaciones seguido de un muestreo sistemático con salto $a = 2$. Una vez observadas las cadenas de Markov de los parámetros es necesario verificar su estacionariedad, así como los tamaños efectivos de las muestras.

Estudio de simulación.

Previo a la implementación computacional del modelo Bayesiano propuesto para una mezcla con distribución log normal y número de componentes fijo, se realiza una simulación que permite conocer de antemano el número de clases latentes que agrupan los siniestros ocurridos según sus cuantías, así como la media y la varianza de cada distribución log normal que genera las observaciones en cada segmento.

El objetivo principal en la modelación es comprobar si la propuesta realizada anteriormente tiene la capacidad de reconstruir las clases latentes iniciales asignadas a cada una de las observaciones. Este hecho será comprobado mediante la construcción de una matriz de adyacencia. Adicionalmente, es necesario verificar si se tiene una 'buena' estimación de los parámetros evaluando si su valor se encuentra contenido o no dentro de los intervalos de credibilidad obtenidos, así como una óptima bondad del ajuste explicada por los valores p predictivos sobre estadísticas como la media, la varianza, la desviación estándar y el coeficiente de variación. Para el estudio, solamente se implementará la primera propuesta de distribución previa cuando $\mathbb{V}\text{ar}(\sigma_h^2) \to \infty$ dado que $a_h = 2$.

Para el algoritmo MCMC, solamente se utilizarán 110.000 iteraciones, haciendo una etapa de quemado de las primeras 10.000 observaciones seguido de un posterior muestreo sistemático con salto de $a = 2$. En total se simulan 1.500 cuantías de siniestros ocurridos agrupados en seis clases latentes. El valor de los parámetros de media y varianza de cada clase generados mediante una distribución log normal, así como sus respectivos pesos se reportan en la siguiente tabla:

Tabla 10: *La primera columna presenta la clase latente, la segunda y tercera presentan la media y varianza respectivamente de la distribución log normal que genera las observaciones y la última columna muestra el peso o proporción de cada clase.*

Clase latente h	Media θ_h	Varianza σ_h^2	Peso ω_h
Z_1	2,0	0,22	0,25
Z_2	3,0	0,01	0,35
Z_3	3,5	0,05	0,15
Z_4	5,0	0,02	0,12
Z_5	6,0	0,08	0,10
Z_6	7,0	0,10	0,03

Las cuantías simuladas para los 1.500 siniestros se pueden visualizar mediante un histograma o función de densidad para cada una de las seis clases construidas:

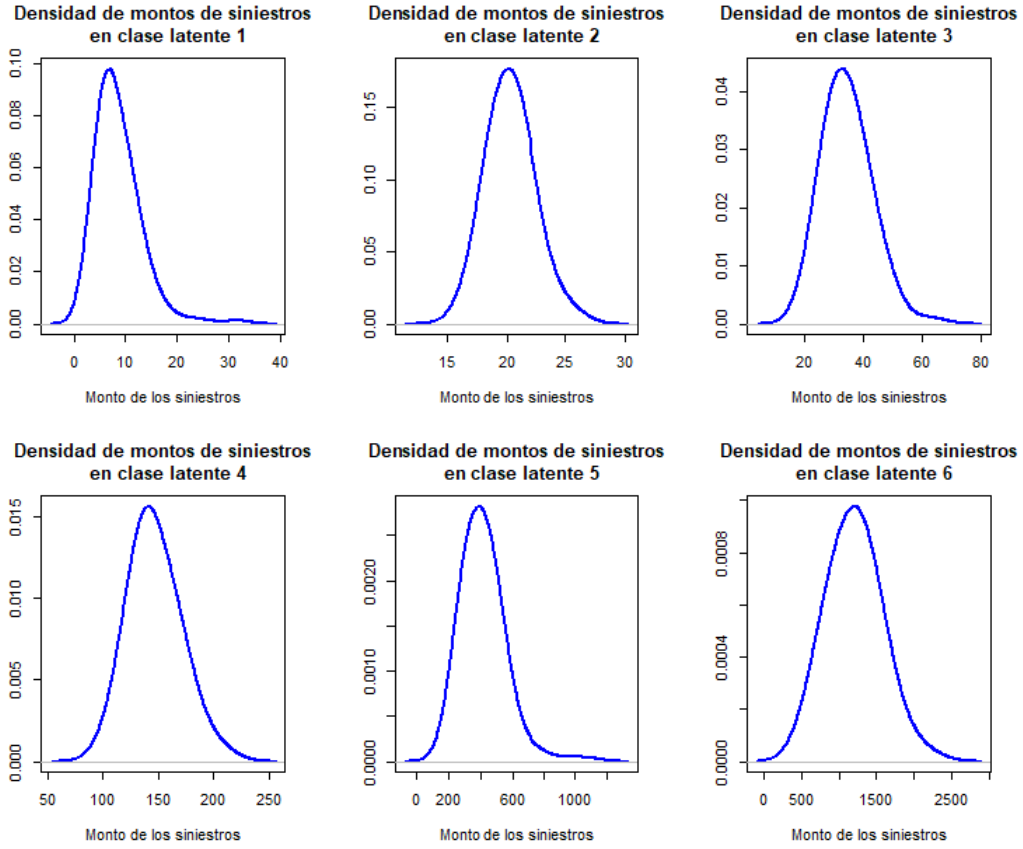

Figura 26: Densidad la cuantía de los siniestros por cada clase latente definida en la simulación.

El gráfico presentado anteriormente permite evidenciar distribuciones asimétricas a derecha dentro de cada clase porque estas observaciones son realización de distribuciones log normales. Por otra parte, las clases presentan atributos de una variable categórica ordinal dado que a medida que se ïncrementa"la modalidad de cada segmento, mayor es el valor de la media de la cuantía de los siniestros simulados.

En la simulación, se conoce como se asignan las observaciones a cada uno de los seis clústeres propuestos. Sin embargo, en la práctica no necesariamente estos grupos o clases latentes se encuentran bien definidas, de manera que se hace necesario realizar una propuesta inicial en las asignaciones, de manera que, con implementación computacional del modelo Bayesiano de mezcla con número de componentes fijo, la convergencia permita identificar las observaciones que se encuentran en un mismo grupo o no. La propuesta inicial de asignaciones para las observaciones se basa en identificar un número óptimo de clústeres entre 2 y 10 con el criterio Bayesiano BIC, dado que según Gelman, A. et al. (2014), el algoritmo EM se puede utilizar para este tipo de problemáticas en modelos de mezcla donde no se conocen los grupos. A continuación, se presentan los resultados del criterio Bayesiano BIC para cada número de grupos propuesto:

Tabla 11: *La primera columna presenta el número de clústeres propuestos y la segunda columna muestra el valor del criterio Bayesiano BIC utilizando el algoritmo EM con 115 iteraciones.*

Número de clústeres	Criterio BIC
2	4715,855
3	4409,238
4	4247,788
5	3940,568
6	3914,788
7	3936,672
8	3958,484
9	3972,414
10	3992,230

La información proporcionada anteriormente permite evidenciar que el criterio BIC se minimiza para 6 clústeres. Por tanto, se implementa el modelo Bayesiano de componentes fijo con este número de clústeres considerando una asignación inicial de grupos para cada una de las observaciones de las cuantías de los 1.500 siniestros simulados basados en el vector de probabilidades estimado para cada observación con en el algoritmo EM. Paso seguido a la implementación computacional, se determina el desempeño de la bondad de ajuste de estadísticos como la media, la varianza, la desviación estándar y el coeficiente de variación por medio de los valores p predictivos:

Tabla 12: *La primera columna muestra el estadístico de bondad de ajuste para el modelo Bayesiano Binomial Negativo para la frecuencia de los siniestros y la segunda columna su respectivo valor p predictivo.*

Estadístico	PPI
Media	0,46882
Varianza	0,45996
Coeficiente de variación	0,46932
Desviación estándar	0,45996

En conclusión, el modelo Bayesiano de mezcla con distribución log normal y número de componentes fijo presenta una 'alta' calidad en su bondad de ajuste sobre las observaciones de las cuantías simuladas para 1.500 siniestros dado que cada uno de los valores p predictivos por estadística se encuentran alrededor de $0,5$. A continuación, se presentan las cadenas de Markov de los parámetros del modelo, posteriores a la etapa de quemado y muestreo sistemático:

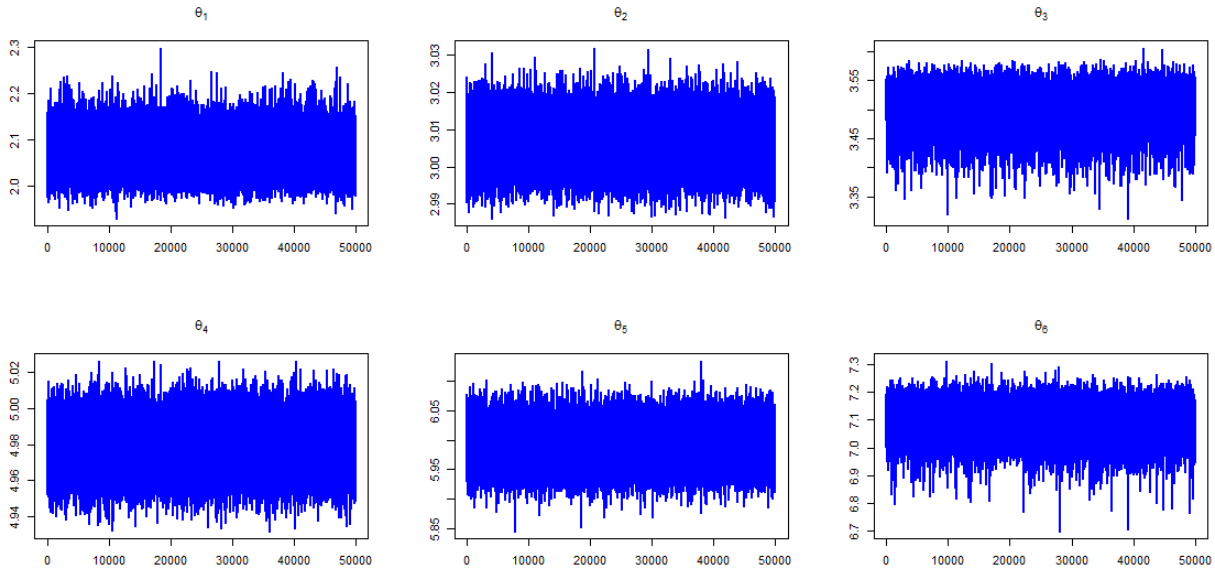

Figura 27: Cadenas de Markov posteriores a etapa de quemado y muestreo sistemático para vector de medias en distribuciones log normales de cada clase latente construida a partir de los montos de cuantía de los siniestros.

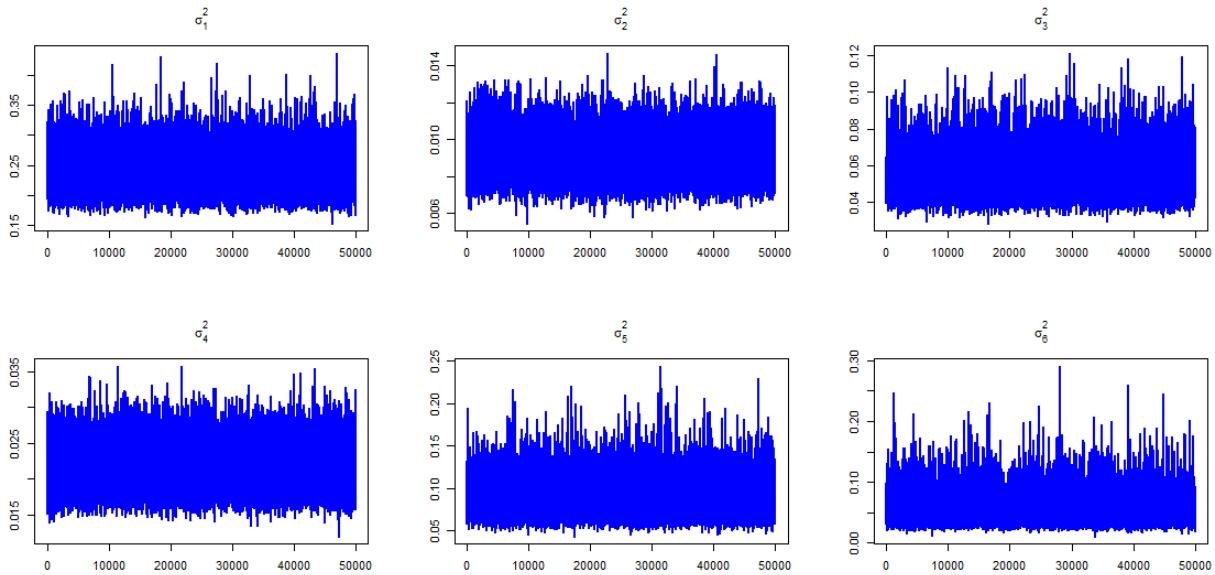

Figura 28: Cadenas de Markov posteriores a etapa de quemado y muestreo sistemático para vector de varianzas en distribuciones log normales de cada clase latente construida a partir de los montos de cuantía de los siniestros.

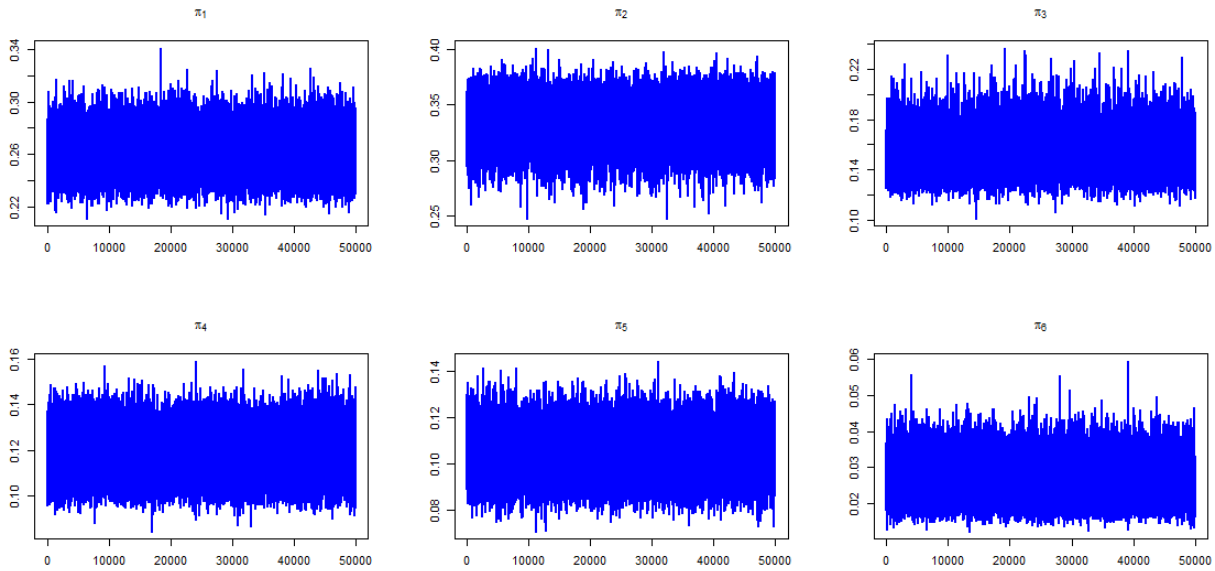

Figura 29: Cadenas de Markov posteriores a etapa de quemado y muestreo sistemático para vector de pesos en clases latentes construidas según los montos de cuantía de los siniestros.

Los gráficos presentados anteriormente proporcionan una información descriptiva con respecto a la estacionariedad las cadenas de Markov. Para cada parámetro estimado se aprecia de forma exploratoria, la convergencia de las observaciones muestreadas a un valor medio constante. En términos del tamaño efectivo, parece existir suficiencia de las muestras observadas para cada parámetro. Sin embargo, este hecho se confirma mediante el cálculo del tamaño efectivo de muestra:

Tabla 13: *La primera columna presenta la clase latente, mientras que las segunda, tercera y cuarta columna presentan los tamaños efectivos de muestra por clase laente para media, varianza y peso, respectivamente.*

Clase latente h	Media θ_h	Varianza σ_h^2	Peso ω_h
Z_1	13646,289	12945,240	20585,938
Z_2	18437,282	10835,992	9809,460
Z_3	6061,063	6531,898	7603,711
Z_4	38689,958	37683,435	43972,598
Z_5	29093,588	12450,765	27284,236
Z_6	13480,901	12970,266	22008,167

El modelo Bayesiano propuesto no solamente mostró un óptimo desempeño para la bondad del ajuste, sino propiedades de convergencia de las observaciones muestreadas por parámetro a un valor medio constante, así como suficiencia en los tamaños de muestra. Sin embargo, es necesario verificar si las estimaciones de los parámetros obtenidas mediante sus respectivas medias posteriores son razonables respecto a su valor real propuesto en el estudio de simulación. Para el proceso de medición de calidad de las estimaciones de los parámetros, se comprueba si el parámetro está contenido o no en un intervalo de credibilidad del 95 % para cada muestra de observaciones:

Tabla 14: *Las columnas uno y tres de la tabla presentan los límites inferior y superior del intervalo de credibilidad de la estimación del parámetro de media en cada clase latente. Por otra parte, las columnas dos y cuatro reportan la media posterior y el valor real del parámetro. Finalmente, la columna cinco es una función indicadora que toma el valor de uno cuando el valor real del parámetro se encuentra contenido en el intervalo de credibilidad y cero en caso contrario.*

Límite inferior	Media Posterior	Límite superior	Media θ_h	Función indicadora
2,00241	2,07029	2,15069	2,0	0
2,99592	3,00674	3,01767	3,0	1
3,42641	3,49672	3,55094	3,5	1
4,95594	4,97902	5,00217	5,0	1
5,93485	5,99121	6,04873	6,0	1
6,97931	7,09395	7,19028	7,0	1

Tabla 15: *Las columnas uno y tres de la tabla presentan los límites inferior y superior del intervalo de credibilidad de la estimación del parámetro de varianza en cada clase latente. Por otra parte, las columnas dos y cuatro reportan la media posterior y el valor real del parámetro. Finalmente, la columna cinco es una función indicadora que toma el valor de uno cuando el valor real del parámetro se encuentra contenido en el intervalo de credibilidad y cero en caso contrario.*

Límite inferior	Media Posterior	Límite superior	Varianza σ_h^2	Función indicadora
0,193159	0,241226	0,303657	0,22	1
0,007356	0,009161	0,011199	0,01	1
0,039949	0,056261	0,079895	0,05	1
0,017063	0,021598	0,027161	0,02	1
0,061423	0,089842	0,132607	0,08	1
0,028748	0,055444	0,103505	0,10	1

Tabla 16: *Las columnas uno y tres de la tabla presentan los límites inferior y superior del intervalo de credibilidad de la estimación del parámetro de peso en cada clase latente. Por otra parte, las columnas dos y cuatro reportan la media posterior y el valor real del parámetro. Finalmente, la columna cinco es una función indicadora que toma el valor de uno cuando el valor real del parámetro se encuentra contenido en el intervalo de credibilidad y cero en caso contrario.*

Límite inferior	Media Posterior	Límite superior	Peso ω_h	Función indicadora
0,23607	0,26187	0,29010	0,25	1
0,29569	0,33136	0,36469	0,35	1
0,13075	0,15658	0,18784	0,15	1
0,10254	0,11886	0,13611	0,12	1
0,08798	0,10431	0,12182	0,10	1
0,01831	0,02701	0,03725	0,03	1

Los resultados proporcionados por cada estimación puntual de los parámetros del modelo, así como de los intervalos de credibilidad del 95 % permiten concluir que las medias posteriores en promedio (RMSE) se desvían en $0,03116$ unidades respecto a su valor real. En términos porcentuales, la desviación promedio en mención corresponde a un valor del $7,27\,\%$. Finalmente, en el $94,73\,\%$ de los casos, el intervalo de credibilidad estimado contiene el valor del parámetro.

Uno de los principales objetivos del modelo Bayesiano de mezcla con número de componentes fijo es mostrar si este tiene una 'alta' capacidad de reconstruir los grupos o clases latentes en su versión original

al momento de realizar la simulación de las cuantías de los siniestros con seis clases latentes. Una manera de verificarlo es estableciendo una comparación entre las matrices de adyacencia original y generada por las probabilidades posteriores que tienen dos observaciones diferentes cualesquiera de tener la misma asignación en su clase latente. Para esta segunda matriz en mención, es necesario observar por cada iteración s del algoritmo MCMC una matriz de funciones indicadoras que toman el valor de 1 cuando dos observaciones tienen asignada la misma clase latente o 0 en caso contrario. Estas asignaciones se obtienen una vez se muestrean del vector $\boldsymbol{p}_i$ tal que:

$$\boldsymbol{p}_i = \left(p\left(Z_{i,*}^{(s)} = 1 \mid \text{resto}\right), \ldots, p\left(Z_{i,*}^{(s)} = H \mid \text{resto}\right)\right),$$

para todo $i \in \{1, \ldots, n\}$. De esta manera, por cada iteración s se genera una matriz $\boldsymbol{A}_{s,n\times n}$ de tamaño $n \times n$ con las siguientes propiedades:

- El componente $a_{ii,s}$ de $\boldsymbol{A}_s$ para todo $i \in \{1, \ldots, n\}$ es tal que $a_{ii,s} = 1$. Lo anterior es consecuencia de que un elemento siempre se encuentra consigo mismo en alguna asignación de las clases latentes.
- Por el numeral anterior, la diagonal de la matriz $\boldsymbol{A}_s$ está compuesta por unos. De este modo:

$$\text{Tr}(\boldsymbol{A}_s) = n.$$

- La matriz $\boldsymbol{A}_s$ es simétrica.
- El componente $a_{ij,s}$ con $i \neq j$ para todo $i, j \in \{1, \ldots, n\}$, es tal que:

$$a_{ij,s} = a_{ji,s} = \begin{cases} 1 & \text{, si } x_i, x_j \in Z_h^{(s)} \text{ para algún } h \in \{1, \ldots, H\} \\ 0 & \text{, en otro caso.} \end{cases}$$

En virtud de la simetría de $\boldsymbol{A}_{s,n\times n}$, para calcular la probabilidad posterior de que dos observaciones cualesquiera tengan la misma asignación, es suficiente con guardar en cada iteración del algoritmo MCMC, la matriz triangular inferior o la matriz triangular superior. En consecuencia, la matriz de adyacencia $\boldsymbol{A}_{\text{post}}$ de tamaño $n \times n$ con las probabilidades posteriores que tienen dos observaciones cualesquiera de tener una misma asignación de clase latente en un algoritmo MCMC de S iteraciones, se define de la siguiente manera componente a componente:

$$a_{ij,\text{post}} = \begin{cases} 1 & \text{, si } i = j \text{ para todo } i, j \in \{1, \ldots, n\} \\ \frac{1}{S}\sum_{s=1}^{S} a_{ij,s} & \text{, si } i \neq j \text{ para todo } i, j \in \{1, \ldots, n\} \end{cases}$$

De esta manera, las componentes de la matriz de adyacencia de las asignaciones originales controladas en la simulación de las observaciones de las cuantías de 1.500 siniestros, únicamente valores toman valores de 1 cuando las observaciones se encuentran en el mismo clúster o 0 en caso contrario. Por otra parte, los componentes de la matriz de adyacencia de probabilidades posteriores toman valores que se encuentran entre 0 y 1. Por la razón anterior, al visualizar un contraste entre ambas matrices se observan colores rojos o blancos en la matriz original, mientras que en la matriz de probabilidades posteriores se aprecian además de los colores en mención, tonos amarillos de transición entre 0 y 1. A continuación se presentan las matrices de adyacencia de las asignaciones originales controladas mediante la simulación y las estimaciones de las probabilidades de que dos observaciones se encuentren con la misma asignación mediante modelo Bayesiano de mezcla con distribución lognormal y número de componentes fijo:

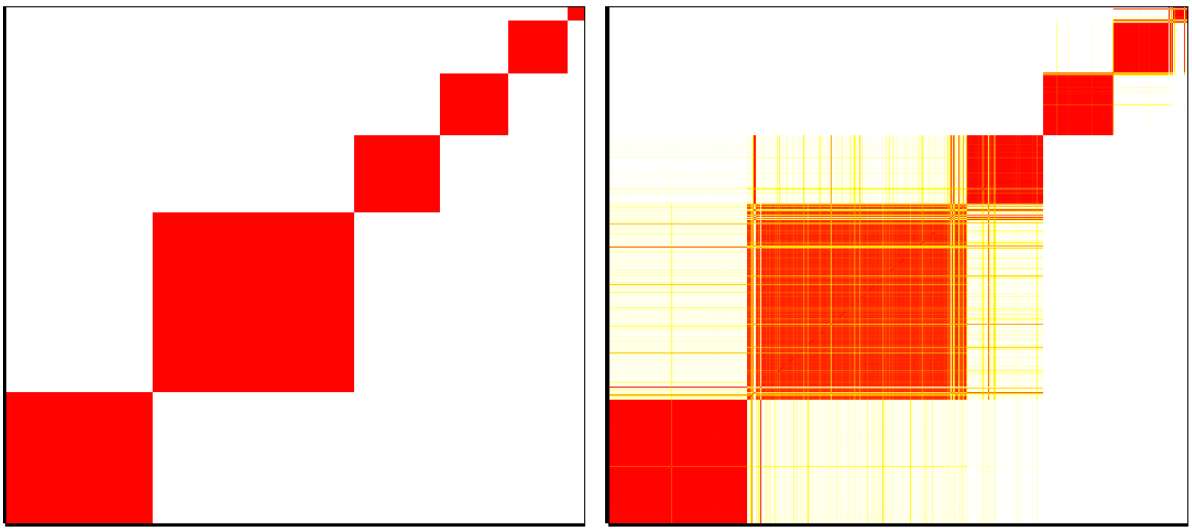

Figura 30: La matriz del panel izquierdo de la visualización muestra las asignaciones originales controladas en la simulación y la matriz del panel derecho presenta la estimación de las probabilidades posteriores de que dos observaciones cualesquiera tengan la misma asignación.

En términos descriptivos, se observa que mediante la implementación computacional del modelo Bayesiano propuesto existe una 'buena' tendencia para reconstruir y recuperar los clústeres originales propuestos en la simulación.

Implementación computacional del modelo para la severidad de los siniestros.

Previo a la implementación computacional del modelo se requiere de los grupos o clases latentes para las cuantías de los siniestros. Como estos grupos se desconocen, por medio de la propuesta del algoritmo EM se determinará el número inicial de clústeres utilizando el criterio Bayesiano BIC:

Tabla 17: *La tabla muestra el número de clústeres y su respectivo valor de BIC cuando se implementa el algoritmo EM con 115 iteraciones.*

Número de clústeres	Criterio Bayesiano BIC
2	13826,90
3	13765,96
4	13788,03
5	-26900,66
6	-26875,27

Los resultados presentados en la tabla anterior permiten concluir que el criterio Bayesiano BIC se minimiza para 5 clústeres. De este modo, la propuesta inicial de clases latentes previa a la implementación del modelo Bayesiano con número de componentes fijo se hace con cinco segmentos utilizando las probabilidades para las asignaciones obtenidas mediante su respectivo algoritmo EM. A continuación, se observa de manera descriptiva la función de densidad de cada clase latente propuesta para la implementación del modelo Bayesiano con número de componentes fijo:

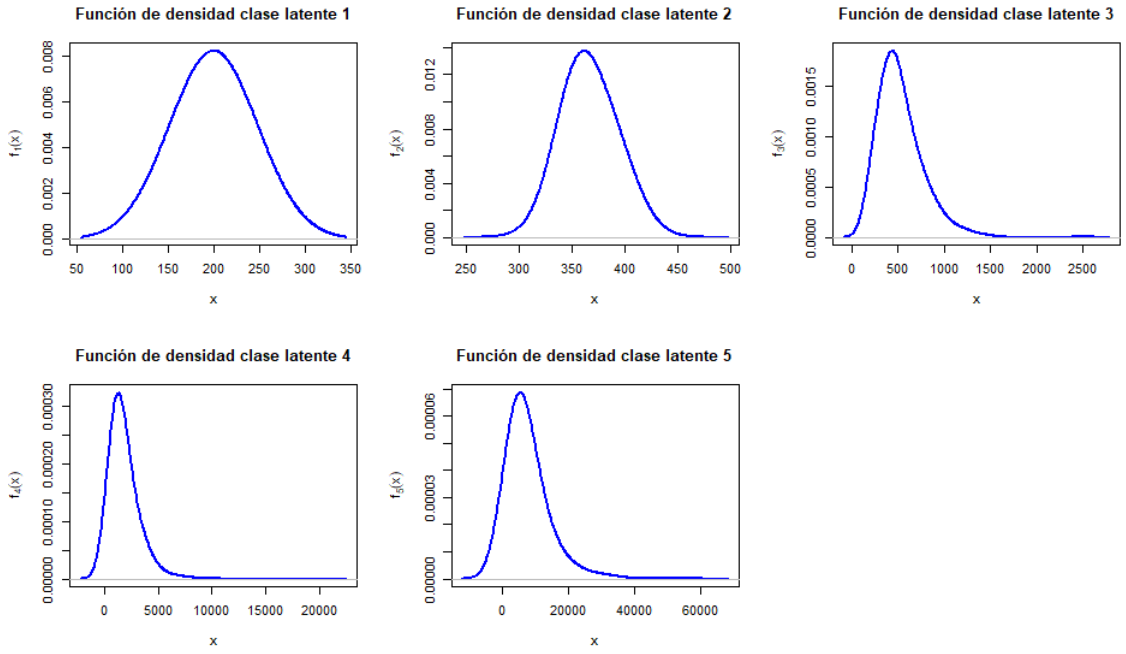

Figura 31: Gráfico de funciones de densidad de las cuantías de los sinietros según su clase latente asignada mediante las probabilidades obtenidas en el algoritmo EM.

Las funciones de densidad presentadas anteriormente siguen una distribución lognormal en cada clase latente para la cuantía de los siniestros. De manera descriptiva y exploratoria se aprecia como a medida que se 'incrementa' de manera ordinal el número de clase latente crece la varianza de la cuantía de los siniestros. Este resultado implica que hay una posible 'tendencia' de heterogeneidad de varianza para la cuantía de los grupos. Adicionalmente, se evidencia que con el 'incremento' ordinal de cada clase, que cada distribución pasa de una forma 'simétrica' a una forma asimétrica a derecha como suele ser característico en las distribuciones lognormales.

Construir los clústeres con el algoritmo EM para la propuesta inicial de asignaciones en la implementación computacional implica realizar una descripción de los atributos de cada uno dado que, al momento de hacer la tarifación en términos de la prima pura de riesgo y la prima comercial, es necesario caracterizar cada segmento. De esta manera si las clases latentes dependen por ejemplo de: el tipo de vehículo, la zona donde se moviliza el conductor del automóvil asegurado, o el género del conductor, las tarifas tendrán variaciones porque las cuantías de los siniestros en la última clase latente son 'mayores' que las cuantías de la primera clase latente. Esta metodología de realizar una tarifación por segmentos resulta más apropiada que una única tarifa para todo el portafolio de asegurados porque en el sector de la industria aseguradora, las tasas no deben ser excesivamente 'altas' porque deben ser 'justas' según el riesgo asegurado para responder a las necesidades y el perfil del cliente.

Complementando los resultados de análisis exploratorio expuestos anteriormente, resulta natural definir cuantías 'bajas' cuando los siniestros se asocian a las clases latentes de los grupos 1 y 2, cuantías 'medias' cuando los siniestros se asocian a la clase latente 3 y cuantías 'altas' cuando los siniestros se asocian a las clases latentes 4 y 5. De cara a describir cada grupo en forma exploratoria, se presentan algunos gráficos de perfil como por ejemplo: la clase latente dado el tipo de vehículo y la clase latente dada la zona de movilidad del vehículo asegurado. El primer gráfico de perfil en mención es importante porque dentro del portafolio de asegurados existen automóviles 'livianos' y 'pesados'. Dentro del grupo de automóviles livianos, las cuantías de los siniestros suelen ser diferentes porque existen diferencias entre vehículos de alta gama y vehículos más sencillos. Al final el objetivo del gráfico de perfil en mención es mostrar si dentro del análisis descriptivo se aprecian estas diferencias entre las cuantías según el tipo de vehículo:

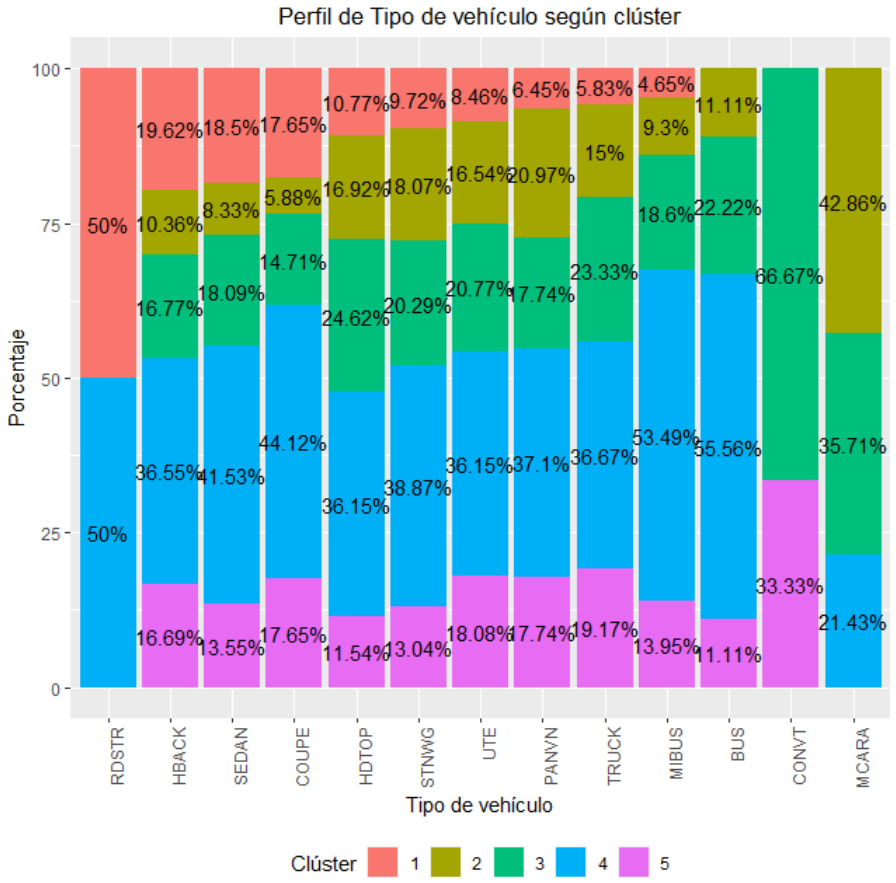

Figura 32: Gráfico de perfil de los vehículos asegurados según las clases latentes de sus cuantías construidas bajo el algoritmo EM.

El gráfico de perfil permite evidenciar que las cuantías más bajas de los siniestros se presentan para el auto RDSTR. Sin embargo, para este mismo vehículo se presentan cuantías 'altas' porque se encuentran en la clase latente del grupo 4. Esto puede atribuirse a factores no tenidos en cuenta aún como la zona de movilidad del vehículo asegurado o el género del conductor. Por otra parte, para vehículos como HBACK, SEDAN, COUPE, HDTOP, STWNG, UTE, PANVN y TRUCK las cuantías 'bajas' y 'altas' se encuentran aproximadamente en un 45 % y 55 % respectivamente, lo cual puede estar atribuido a factores aún no considerados aún en este análisis. Para el caso especial de microbuses y buses, se observa que las cuantías tienden a ser 'altas' porque aproximadamente el 66 % de las cuantías de los siniestros se encuentran en las clases latentes de los grupos 4 y 5. En términos de un automóvil lujoso y de alta gama como el convertible se aprecia que las cuantías tienden a ser de un nivel 'medio-alto' porque se encuentran en las clases latentes de los grupos 3 y 5. Finalmente, el auto MCARA reporta cuantías de nivel 'bajo-medio' porque se encuentran en los grupos 2, 3 y 4, con mayor prevalencia en los dos primeros.

El análisis descriptivo presentado anteriormente permite observar en su modo exploratorio que efectivamente el tipo de vehículo asegurado según categorías como 'liviano', 'pesado', 'lujo' o 'sencillo' puede ser un factor influyente para la cuantía de los siniestros ocurridos de manera que la consideración de segmentos o clases latentes resulta apropiada para la tarifación de los seguros porque esto implica tasas 'justas' dependiendo de cada riesgo en vez de una tasa global 'media' para todos los vehículos influenciada por las cuantías más altas como es característico en distribuciones asimétricas a derecha. Una tasa de seguros basada en la media de todas las cuantías genera desde luego un coste 'alto' que a su promueve bajas posibilidades de competencia en el sector asegurador con los productos de automóviles ofertados

por otras compañías.

El gráfico de perfil de las zonas de movilidad de los vehículos dadas las clases latentes ordinales para la cuantía de los siniestros ocurridos se presenta a continuación:

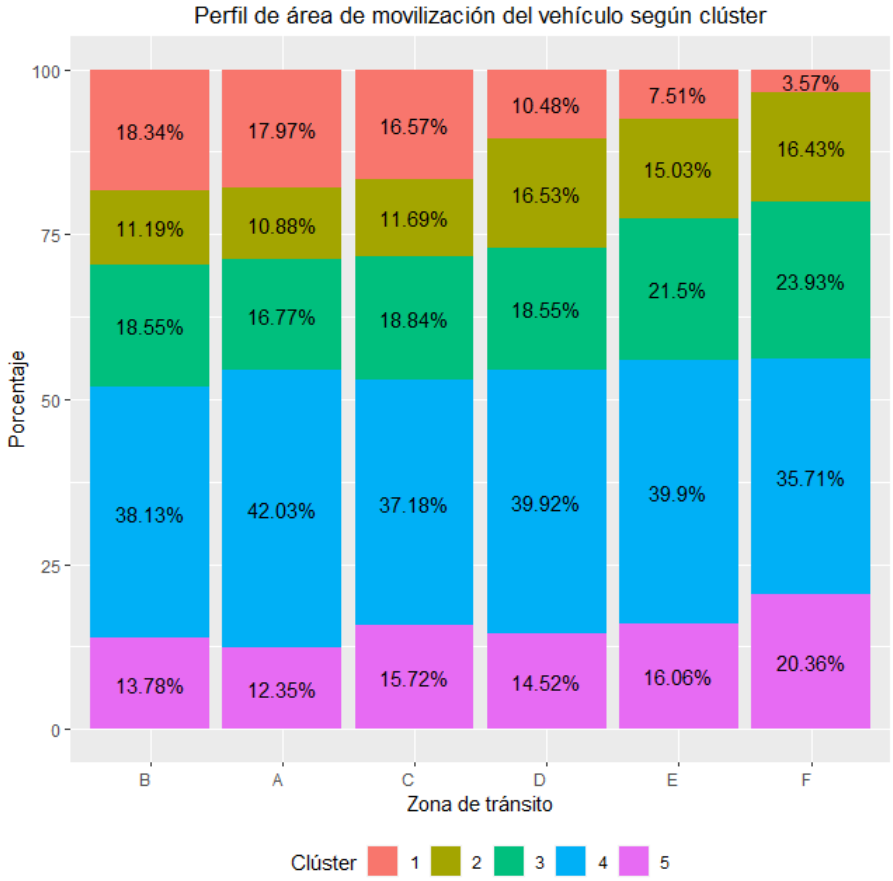

Figura 33: Gráfico de perfil de las zonas de movilidad de los vehículos asegurados según las clases latentes de sus cuantías construidas bajo el algoritmo EM.

El gráfico permite evidenciar que en la zona B, las cuantías de los siniestros de cuantía 'baja-media' (clases latentes 1, 2 y 3) y de cuantía 'alta' (clases latentes 4 y 5) se encuentran en una proporción aproximada de 50 % y 50 % respectivamente. Sin embargo, si las zonas de movilidad de los vehículos asegurados se ordenan de la siguiente forma: A, C, D, E y F, se aprecia que las cuantías 'bajas' (clases latentes 1 y 2) comienzan tienen una menor participación a medida que se 'incrementa' la zona, de manera que se observa un fenómeno donde las cuantías que prevalecen por ejemplo en la última zona (F) son 'medias' (clase latente 3) y 'altas' (clases latentes 4 y 5). De esta manera, la posible interacción entre el tipo de vehículo y la zona de movilidad por ejemplo debe mostrar como los segmentos o clases latentes son 'apropiados' para definir una tarifa por segmento según sus atributos y no de forma global sesgando el costo de la póliza por las cuantías de los siniestros más altas. Para observar una caracterización multivariante de los grupos construidos anteriormente y de acuerdo con las ideas expuestas por Asan y Ercan (2012) se puede utilizar una red neuronal de mapas autoorganizados de kohonen de manera que cada observación se asocia a un nodo específico en la capa de salida de la red tal. Previo a la implementación computacional del mapa autoorganizado se construyó una categoría ordinal para los vehículos según la media del logaritmo de sus cuantías tal y como se presenta en la siguiente tabla:

La red neuronal con capas de kohonen fue implementada computacionalmente con 10.000 épocas al

Tabla 18: *La primera columna presenta el tipo de vehículo asegurado y la segunda columna reporta la variable categórica ordinal asociada a cada tipo de vehículo según la media del logaritmo de la cuantía de los siniestros.*

Vehículo	Categórica ordinal	Media del logaritmo de la cuantía
RDSTR	0-RDSTR	6,181307
MCARA	1-MCARA	6,323063
CONVT	2-CONVT	6,814727
SEDAN	3-SEDAN	6,755895
HDTOP	4-HDTOP	6,762193
HBACK	5-HBACK	6,806451
STNWG	6-STNWG	6,819561
PANVN	7-PANVN	6,882605
BUS	8-BUS	6,912843
UTE	9-UTE	6,920123
TRUCK	10-TRUCK	6,997404
COUPE	11-COUPE	7,067228
MIBUS	12-MIBUS	7,129499

momento de ejecutar el algoritmo de propagación hacia atrás. El error cuadrático medio asociado a la función de costo fue de 0.81651. La segmentación por nodos para las observaciones de las cuantías de los siniestros en un mapa de características de 5 se presenta a continuación:

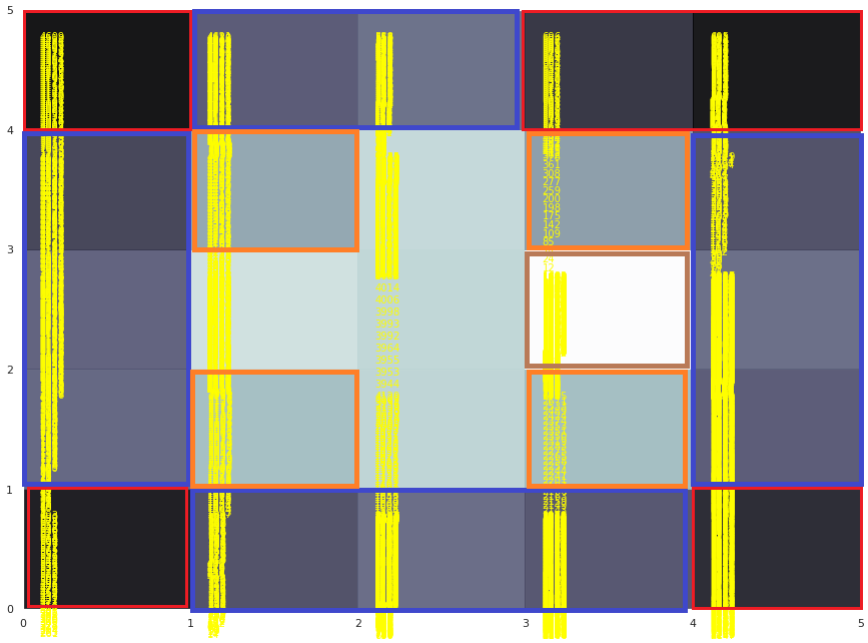

Figura 34: Mapa autoorganizado de redes neuronales para segmentación de las cuantías de los siniestros.

De forma descriptiva, los recuadros de color rojo, naranja, azul y café permiten apreciar que al parecer

se puede realizar una segmentación con 5 grupos según las tonalidades apreciadas en la visualización del mapa autoorganizado. Sin embargo, no necesariamente estos mismos cinco grupos se corresponden con las clases latentes del algoritmo EM. Para observar las clases en mención con sus respectivos atributos en el mapa autoorganizado se tienen las siguientes visualizaciones de datos:

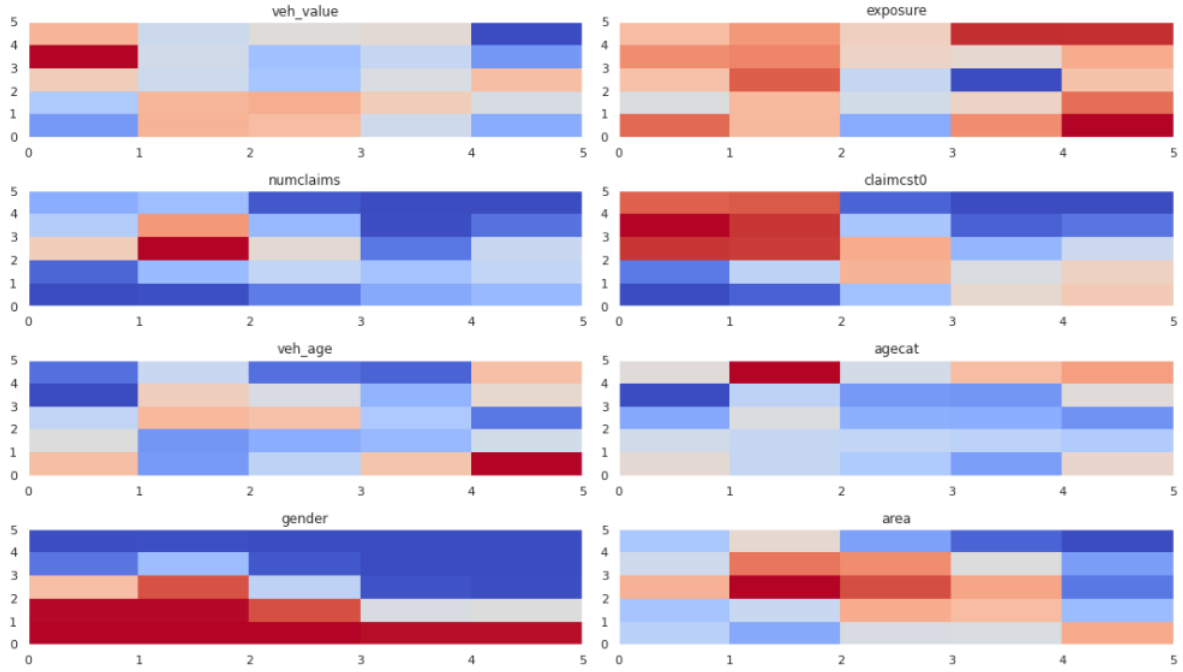

Figura 35: Atributos de los vehículos asegurados en mapa autoorganizado de kohonen.

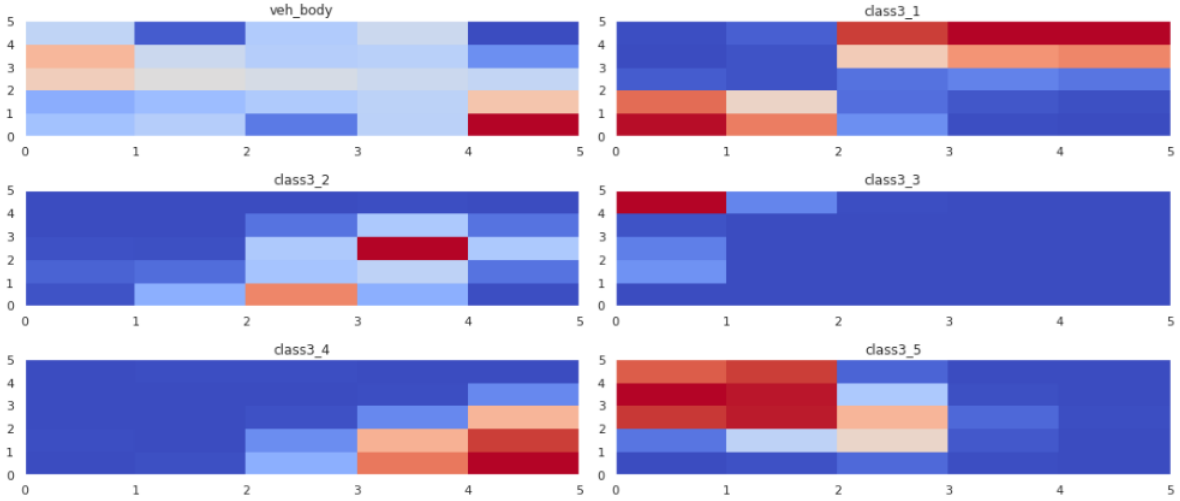

Figura 36: Atributos de los vehículos asegurados y clases latentes en mapa autoorganizado de kohonen.

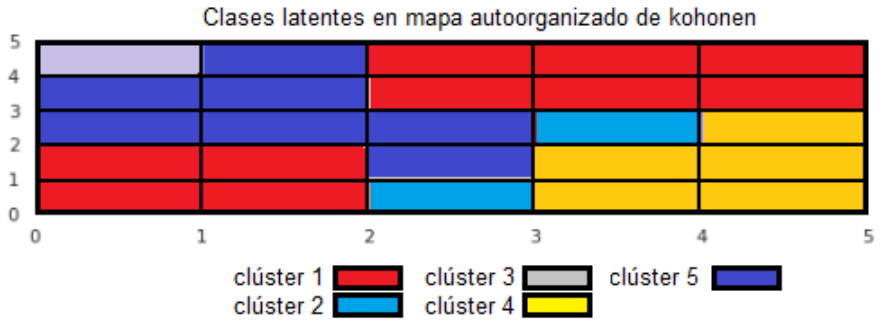

Figura 37: Agrupación de clases latentes de algoritmo EM en mapa autoorganizado de kohonen.

Tanto los colores rojos como azules en las figuras 35 y 36 muestran una 'excitación' de las neuronas por el volumen de sus pesos de manera que proporcionan información sobre la magnitud del atributo, de manera que para tonalidades de color rojo, las observaciones de las cuantías para los individuos del nodo tienden a tener el atributo, mientras que para tonalidades de color azul, las observaciones de las cuantías para los individuos del nodo tienden a no tener el atributo. El atributo se asocia por ejemplo con: el valor asegurado del vehículo, su tiempo de exposición o cobertura en el contrato de la póliza de seguros, el número de siniestros de la póliza, la cuantía de los siniestros, la edad del vehículo, el tiempo de uso del vehículo, el género del conductor del vehículo asegurado, el área o zona de tránsito usual del vehículo asegurado, el tipo de vehículo asegurado y las clases latentes de las cuantías construidas con el algoritmo EM. Al realizar una 'fusión' entre los mapas autoorganizados de las clases latentes observadas en la figura 36 resulta la figura 37. En esta fusión las clases son excluyentes excepto un nodo (0,5) de la clase 5 que tiende a traslaparse con la clase 3, pero que por 'excitación' de las neuronas por el volumen de sus pesos, tiene a estar más en la clase 3 que en la clase 5.

En términos del valor asegurado se aprecia que tiende a ser mayor para la clase latente 5 y menor en la clase latente 1, lo cual implica que la cuantía depende de este valor con el cual se asegura el vehículo en la póliza de seguros, lo cual resulta natural desde el negocio asegurador dado que, a mayor valor asegurado, en caso de siniestro mayor tiende a ser la cuantía; y entre menor es el valor asegurado, menor tiende a ser la cuantía en caso de siniestro. Note que el valor asegurado está asociado en términos prácticos con el tipo de vehículo porque se encuentra en función de su valor comercial, así pues, este valor depende de si el vehículo es 'liviano', 'pesado', de 'lujo' o no.

Con respecto a la protección del vehículo en función del tiempo de cobertura de la póliza de seguros se aprecia que entre mayor es el tiempo de exposición del riesgo, las cuantías pueden estar en las clases latentes 1 o 4, y así pueden ser muy 'bajas' o muy 'altas'; mientras que, a menor tiempo de exposición, las cuantías tienden a ser 'bajas' porque se encuentran en la clase latente 2. Para el número de siniestros que presenta la póliza se observa que, a mayor número de siniestros, mayor es la cuantía porque las observaciones se agrupan en nodos de la clase latente 5. Sin embargo, cuando la póliza presenta menos siniestros, las cuantías tienden a ser 'bajas' porque se agrupan en nodos de la clase latente 1.

La edad del vehículo permite evidenciar de manera descriptiva que cuando esta es 'grande', las cuantías tienden a ser 'altas' porque se encuentran en nodos de la clase 4, pero cuando esta edad del vehículo es 'pequeña', las cuantías tienden a concentrarse en las clases 1, 3 y 5. En comparación con lo anterior, cuando el tiempo de uso del vehículo, la concentración se asocia con un segmento de cuantías altas en la clase latente 5.

Para la segmentación del mapa autoorganizado, el género fue codificado de manera que: 1 se corresponde con la modalidad de hombres y 0 con la modalidad de mujeres. De esta manera, el color rojo en la visualización de género se asocia con hombres y el color azul con mujeres. Por tanto, las cuantías de los siniestros para vehículos asegurados de conductores hombres tiende a estar en las clases 1, 2 y 4; mientras

que las cuantías de los siniestros para vehículos asegurados conducidos por mujeres se encuentran en todas las clases latente. De este modo, al parecer las cuantías de los siniestros en mujeres son más 'grandes' porque tienen participación en clases latentes de los grupos 3 y 5; mientras que los hombres pueden tener una participación casi nula o nula en estas.

Las áreas de tránsito fueron categorizadas en la segmentación de la siguiente manera: 1 para A, 2 para B, 3 para C, 4 para D, 5 para E y 6 para F. De este modo el color rojo en la visualización se asocia a áreas como D, E y F, y el color azul a áreas como A, B y C. De este modo los resultados permiten apreciar que las mayores cuantías que se asocian a la clase 5 por supuesto, se encuentran áreas como D, E y F; mientras que las menores cuantías de clases como por ejemplo 1, se asocian con áreas como A, B y C. Finalmente con respecto al tipo de vehículo, se evidencia que de acuerdo con la categorización ordinal presentada en la tabla 18 según la media del logaritmo de la cuantía de los siniestros, los autos de 'mayor' categoría según su media de siniestralidad tienden a estar localizados en los segmentos de clases latentes 4 y 5; mientras que los demás autos tienden a estar el resto de clases latentes. Este resultado implica que en efecto el tipo de vehículo asegurado se encuentra asociado con las cuantías de los siniestros de manera que las más altas se presentan para pesados o de lujo como camiones, buses, microbuses o convertibles.

Una vez descritos los segmentos mediante las técnicas descriptivas presentadas anteriormente resaltando la importancia de estos para el negocio asegurador y para la tarifa, se implementa computacionalmente el modelo Bayesiano con distribución lognormal para la cuantía de los siniestros bajo cada una de las tres propuestas de distribuciones previas no informativas. Los resultados de tamaño efectivo de muestra para cada propuesta con 210.000 iteraciones en una etapa de quemado de las primeras 10.000 observaciones y un salto de muestreo sistemático de $a = 2$, se presentan a continuación:

Tabla 19: *Tabla de tamaños efectivo bajo propuesta no informativa 1. La primera columna presenta la clase latente, la segunda columna reporta los tamaños efectivos de muestra para la media de cada clase, la tercera columna proporciona el tamaño efectivo de muestra para la varianza de cada clase y la última columna, el tamaño efectivo para el peso de cada clase.*

Clase	θ_h	σ_h^2	w_h
Z_1	100000,00	98558,43	100000,00
Z_2	63513,11	48670,50	28419,70
Z_3	1096,33	1328,21	984,66
Z_4	815,46	814,18	605,00
Z_5	532,02	752,57	550,45

Tabla 20: *Tabla de tamaños efectivo bajo propuesta no informativa 2. La primera columna presenta la clase latente, la segunda columna reporta los tamaños efectivos de muestra para la media de cada clase, la tercera columna proporciona el tamaño efectivo de muestra para la varianza de cada clase y la última columna, el tamaño efectivo para el peso de cada clase.*

Clase	θ_h	σ_h^2	w_h
Z_1	100000,00	97738,73	100000,00
Z_2	64017,88	43314,43	21158,01
Z_3	922,42	1154,14	839,38
Z_4	501,95	683,76	512,39
Z_5	649,07	845,93	616,65

Tabla 21: *Tabla de tamaños efectivo bajo propuesta no informativa 3. La primera columna presenta la clase latente, la segunda columna reporta los tamaños efectivos de muestra para la media de cada clase, la tercera columna proporciona el tamaño efectivo de muestra para la varianza de cada clase y la última columna, el tamaño efectivo para el peso de cada clase.*

Clase	θ_h	σ_h^2	w_h
Z_1	97884,08	96804,34	100000,00
Z_2	55991,13	37487,49	15082,52
Z_3	632,81	807,87	680,55
Z_4	495,25	764,33	553,29
Z_5	426,05	577,82	459,82

Los resultados anteriores bajo las tres propuestas de distribuciones previas no informativas muestran que los tamaños efectivos de muestra luego de las etapas de quemado y muestreo sistemático de las observaciones tienden a ser 'grandes' para los parámetros de media, varianza y peso de las dos primeras clases; mientras que tienden a ser 'pequeños' para los parámetros en mención dentro de las clases 3, 4 y 5. El paso seguido al anterior, es presentar los resultados de la estimación de parámetros bajo cada una de las propuestas de distribuciones previas no informativas teniendo en cuenta las medias posteriores con sus respectivos intervalos de credibilidad del 95 %. En los tres casos, La primera columna presenta el parámetro de interés, la segunda el límite inferior del intervalo de credibilidad del 95 %, la tercera columna presenta la media posterior y la cuarta columna presenta el límite superior del intervalo de credibilidad del 95 %:

Tabla 22: *Tabla de estimación de parámetros bajo la propuesta no informativa 1.*

Parámetro	Límite inferior	Media Posterior	Límite superior
θ_1	5,29752	5,29871	5,29990
θ_2	5,90008	5,90654	5,91304
θ_3	5,99324	6,11205	6,25924
θ_4	6,82366	7,01804	7,24934
θ_5	7,90503	8,17455	8,57191
σ_1^2	0,00024	0,00026	0,00029
σ_2^2	0,00339	0,00399	0,00469
σ_3^2	0,09877	0,14106	0,19778
σ_4^2	0,17380	0,31789	0,45943
σ_5^2	0,54008	0,74613	0,90937
w_1	0,14526	0,15562	0,16621
w_2	0,11729	0,13015	0,14327
w_3	0,08813	0,14404	0,21423
w_4	0,15459	0,27505	0,39627
w_5	0,17458	0,29514	0,39269

Tabla 23: *Tabla de estimación de parámetros bajo la propuesta no informativa 2.*

Parámetro	Límite inferior	Media Posterior	Límite superior
θ_1	5,29733	5,29903	5,30073
θ_2	5,90175	5,90890	5,91616
θ_3	6,00836	6,14261	6,31144
θ_4	7,91470	8,20141	8,61938
θ_5	6,83630	7,04699	7,30280
σ_1^2	0,00048	0,00053	0,00059
σ_2^2	0,00442	0,00521	0,00616
σ_3^2	0,10582	0,15359	0,21994
σ_4^2	0,52229	0,73635	0,90881
σ_5^2	0,19182	0,32978	0,47430
w_1	0,14570	0,15613	0,16678
w_2	0,12193	0,13521	0,14867
w_3	0,08476	0,14609	0,22333
w_4	0,16228	0,28484	0,38680
w_5	0,15986	0,27774	0,39681

Tabla 24: *Tabla de estimación de parámetros bajo la propuesta no informativa 3.*

Parámetro	Límite inferior	Media Posterior	Límite superior
θ_1	5,29728	5,29938	5,30150
θ_2	5,90330	5,91126	5,91941
θ_3	6,01610	6,16818	6,36230
θ_4	6,83721	7,07059	7,36223
θ_5	7,92970	8,23805	8,68180
σ_1^2	0,00072	0,00080	0,00090
σ_2^2	0,00550	0,00654	0,00780
σ_3^2	0,11085	0,16402	0,24047
σ_4^2	0,19521	0,33769	0,49050
σ_5^2	0,49904	0,71803	0,90010
w_1	0,14624	0,15663	0,16732
w_2	0,12590	0,13960	0,15362
w_3	0,08142	0,14758	0,23604
w_4	0,15138	0,28246	0,40618
w_5	0,14750	0,27372	0,38117

Los resultados presentados anteriormente permiten concluir que las estimaciones de los parámetros del

modelo Bayesiano con número de componentes fijo en términos de las medias posteriores tienden a ser 'similares'. Sin embargo, para la propuesta 2, los clústeres de las clases latentes 4 y 5 observados en las propuestas 1 y 3 se intercambian en el proceso de mezcla y muestreo de las asignaciones del algoritmo MCMC. Haciendo claridad en esta salvedad, las estimaciones de los parámetros bajo las distribuciones no informativas propuestas para el modelo, tienden a ser las 'mismas'. Respecto a la bondad de ajuste del modelo Bayesiano haciendo uso de las tres propuestas de previas no informativas, se aprecia una importante 'estabilidad' en los valores p predictivos asociados a estadísticos tales como: media, varianza, desviación estándar y coeficiente de variación puesto que se encuentran alrededor de $0,5$ tal y como se aprecia en la siguiente visualización de resultados:

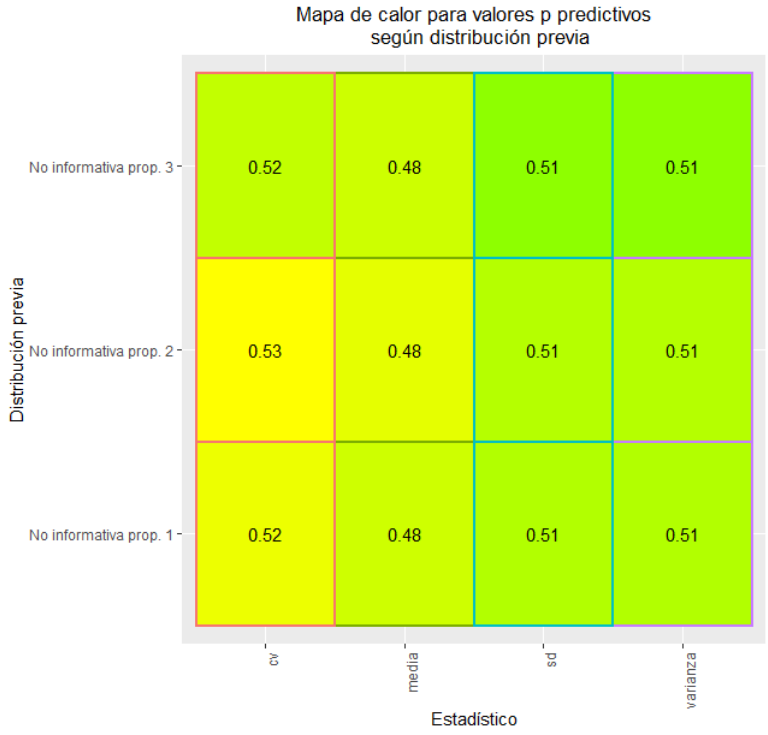

Figura 38: Mapa de calor de valores p predictivos de bondad de ajuste para modelo Bayesiano con número de componentes fijo bajo las tres propuestas de distribuciones previas no informativas.

Estudio de sensibilidad.

El estudio de sensibilidad del modelo Bayesiano con número de componentes fijo se realiza teniendo en cuenta la variabilidad de los hiperparámetros b_h y τ_h^2. En cada una de las distribuciones previas no informativas se consideraron valores de a_h que garantizaran una 'alta' variabilidad de $\mathbb{V}\text{ar}(\sigma_h^2)$ cuando $\sigma_h^2 \overset{\text{d}}{=} \text{Inversa Gamma}(a_h, b_h)$. Así mismo es importante considerar la variabilidad de τ_h^2 cuando $\theta_h \overset{\text{d}}{=} \text{Normal}(\mu_h, \tau_h^2)$ para todo $h \in \{1, \ldots, H = 5\}$. En consecuencia, se considera un estudio de sensibilidad tal que $a_h \in \{2, 3, 4\}$, $b_h \in \{0.01, 0.05, 0.09\}$, $\tau_h^2 \in \{0.5, 1, 1.5\}$ y cada μ_h es la media muestral del logaritmo de la cuantía de los siniestros de la clase latente h, donde $h \in \{1, \ldots, H = 5\}$.

El objetivo fundamental del estudio de sensibilidad es realizar una validación cruzada bajo la metodología k- fold con $k = 5$ de manera que el fusionar $k-1$ particiones se estime el modelo Bayesiano de componentes fijo con estas para muestrear de la distribución posterior tantas observaciones como número de estas se tenga en la partición restante. Note que esta metodología asume que las observaciones de la última partición son faltantes de manera que se pueda establecer una comparación entre los valores esperados y los valores reales de las cuantías de los siniestros por medio de métricas como la raíz del error cuadrático medio (RMSE) y el error porcentual absoluto medio (MAPE). Para complementar estos resultados se

estima la densidad puntual del logaritmo de la predicción (lpdd) tal y como se aprecia en el siguiente mapa de calor:

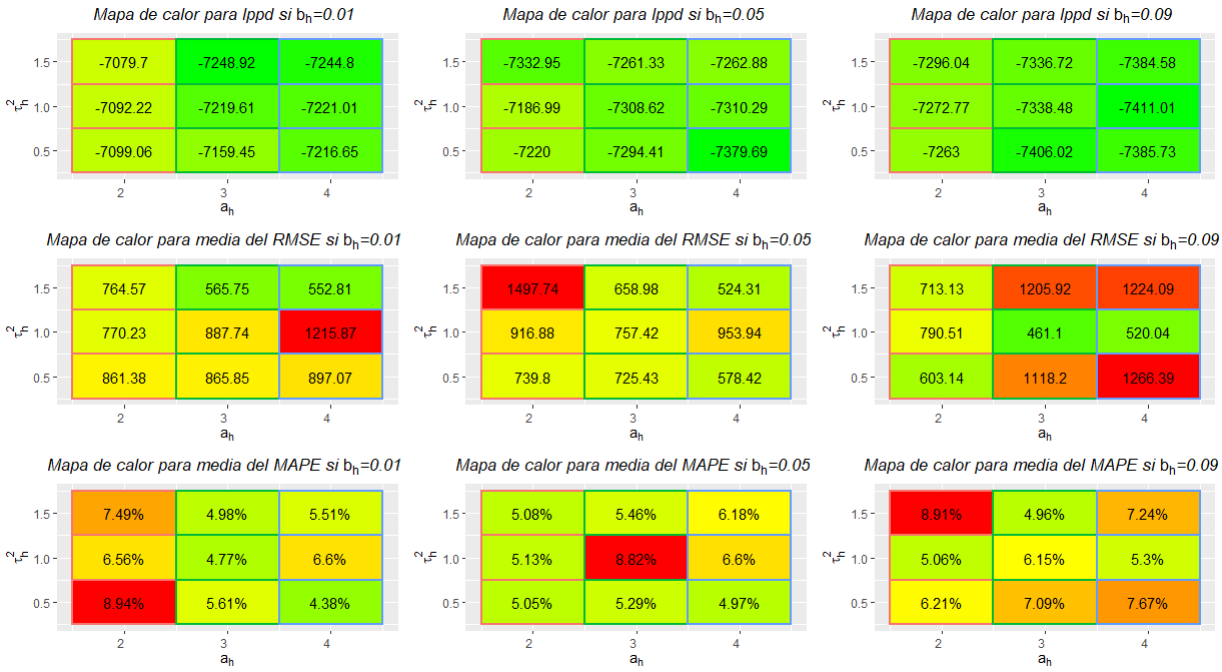

Figura 39: Mapa de calor para métricas de MAPE, RMSE y lppd en validación cruzada k-fold de modelo Bayesiano con número de componentes fijo.

En la visualización anterior se aprecia que para las tres propuestas de distribución previa, el modelo que minimiza el indicador lppd con un valor de $-7.411,01$ es aquel donde se estableció que $a_h = 4, b_h = 0.09$ y $\tau_h^2 = 1$. Sin embargo, este modelo en mención no minimiza el RMSE o el MAPE puesto que la métrica RMSE es mínima cuando $a_h = 3, b_h = 0.09$ y $\tau_h^2 = 1$ con un valor de $461,1$US y la métrica MAPE toma el valor mínimo considerando las tres propuestas de modelo cuando $a_h = 2, b_h = 0.09$ y $\tau_h^2 = 1$ con un valor del $5,06\,\%$. Analizando la sensibilidad del modelo Bayesiano con número de componentes fijo bajo el estudio propuesto previamente se observa que la media del logaritmo de densidad puntual predictiva se encuentra entre -7411.01 y -7099.7. Desde un punto de vista perceptivo esta oscilación no parece ser importante, sin embargo, al evaluar las métricas de calidad de pronóstico como RMSE y MAPE, si se observan importantes diferencias explicadas por los cambios de b_h y τ_h^2. Complementado el análisis de resultados con respecto a la métrica lppd, se observa que a medida que se incrementa el valor de b_h cuando τ_h^2 y a_h se mantienen constantes, esta tiende a ser menor. En consecuencia, a medida que aumenta b_h el modelo tiende a un mejor ajuste respecto a las predicciones del modelo Bayesiano con número de componentes fijo. Del mismo modo, el aumento de b_h bajo los demás hiperparámetros constantes tiende a 'minimizar' simultáneamente los valores de las métricas para la raíz cuadrada del error cuadrático medio y el error porcentual absoluto medio.

5.1.2. Modelo Bayesiano de mezcla con distribución lognormal con número de componentes aleatorio.

Retomando la propuesta del modelo Bayesiano de mezcla con distribución lognormal y número de componentes fijo, es posible plantear un número de componentes aleatorio entre 2 y un máximo H. La escogencia del número óptimo de clústeres depende de la moda posterior entre 2 y H con respecto al número de clases latentes que quedan con al menos una observación asignada.

En principio, la propuesta de número de componentes aleatorio no se diferencia de la propuesta de número de componentes fijo porque se establece una segmentación de la cuantía de los siniestros en H clases latentes, de manera que se busca homogeneidad dentro de cada clase y sus observaciones son realizaciones de una distribución lognormal con parámetros θ_h y σ_h^2:

$$x_i \mid z_i \overset{\text{iid}}{=} \text{lognormal}(x_i \mid \theta_h, \sigma_h^2).$$

Por otra parte, la proporción de las clases latentes tiene una medida de incertidumbre en la práctica porque en el negocio asegurador cuando se tiene un portafolio de contratos de seguros, no es posible determinar con certeza si ocurrirán más siniestros de severidad baja que de severidad media o severidad alta en caso de contar únicamente con tres grupos o segmentos para agrupar las cuantías. En consecuencia, si se dispone de al menos tres clases latentes, las proporciones aleatorias de los grupos son realización de una distribución multinomial:

$$z_i \mid w \overset{\text{iid}}{=} \text{Multinomial}(1; \omega_1, \omega_2, \ldots, \omega_H).$$

Bajo el enfoque del modelo de número de componentes fijo, el vector $\boldsymbol{\omega}$ es realización de una distribución Dirichlet. Sin embargo, bajo la propuesta de número de componentes aleatorio, este hecho presentará una variación. En la teoría de la probabilidad, la generación de un vector con distribución Dirichlet con H componentes se obtiene mediante H números aleatorios independientes $\{y_1, \ldots, y_H\}$ con distribución gamma cuyos parámetros de forma y escala son respectivamente α y 1. De esta manera, la generación de un vector $\boldsymbol{w}$ con distribución Dirichlet se obtiene de la siguiente manera:

$$\boldsymbol{w} = \left(\frac{y_1}{\sum_{h=1}^{H} y_h}, \ldots, \frac{y_H}{\sum_{h=1}^{H} y_h} \right).$$

En consecuencia, se propone una distribución previa para el vector aleatorio $\boldsymbol{\omega}$ desde la generación de números aleatorios Dirichlet:

$$\begin{aligned} \lambda_h &\overset{\text{d}}{=} \text{Gamma}\left(\frac{n_0}{H}, 1\right) \\ \boldsymbol{\omega} &= \left(\frac{\lambda_1}{\sum_{h=1}^{H} \lambda_h}, \ldots, \frac{\lambda_H}{\sum_{h=1}^{H} \lambda_h} \right) \end{aligned} \tag{21}$$

Observe que los componentes del vector de medias $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H)$ se encuentran en los números reales, mientras que los componentes del vector de varianzas $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_H^2)$ se encuentran en los números reales positivos. Por lo anterior, es razonable proponer para todo $h \in \{1, \ldots, H\}$ que cada θ_h sea realización de una distribución normal dado que su soporte se encuentra en $\mathbb{R}$, y que cada σ_h sea realización de una distribución inversa gamma dado que soporte se encuentra en $\mathbb{R}^+$:

$$\begin{aligned} \theta_h &\overset{\text{d}}{=} \text{Normal}(\mu_h, \tau_h^2) \\ \sigma_h^2 &\overset{\text{d}}{=} \text{Inv-Gamma}(a_h, b_h). \end{aligned}$$

Por cada iteración del algoritmo MCMC que será propuesto para el modelo Bayesiano de mezcla descrito anteriormente se genera un tamaño n_h para el clúster Z_h con $h \in \{1, \ldots, H\}$, de manera que puede existir algún h para el cual $n_h = 0$ en caso de que su peso sea muy pequeño o de que dos o más grupos tiendan a un proceso de fusión. En consecuencia, por cada iteración s se determina el número de clústeres no vacíos $H_n^{(s)}$, de manera que el número de componentes óptimo del modelo de mezcla es la moda posterior de H_n y es aleatorio porque si bien puede ser H, también puede estar entre 1 y H:

$$1 \leq H_h \leq H.$$

Observe que en caso de que la moda posterior de H_n sea 1, no existe un modelo de mezcla sino una distribución lognormal que genera una muestra aleatoria X compuesta por observaciones $\{x_1, \ldots, x_{N(t)}\}$ y que corresponden a la cuantía de $N(t)$ siniestros ocurridos en el intervalo $(0, t]$. Por otra parte el conjunto de parámetros del modelo está compuesto por: H_n, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_H)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H)$ y $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_H^2)$, mientras que el conjunto de hiperparámetros está dado por: n_0, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_H)$, $\boldsymbol{\tau} = (\tau_1^2, \ldots, \tau_H^2)$, $\boldsymbol{a} = (a_1, \ldots, a_H)$ y $\boldsymbol{b} = (b_1, \ldots, b_H)$.

Distribución previa no informativa para el modelo.

A continuación se presenta un conjunto de propuestas para distribuciones previas no informativas análogas a las presentadas para el modelo Bayesiano de mezcla con número de componentes fijo:

1. El valor de n_0 es tal que n_0 dado que no se conoce información previa sobre las clases latentes y sus proporciones. Los componentes del vector $\boldsymbol{\mu}$ son las medias muestrales del logaritmo de la cuantía de los siniestros por cada clase. Los componentes de $\boldsymbol{\tau}$ toman el valor de 1 y se considera este valor como una 'alta' desviación en escala logarítmica que en caso de tomar valores más grandes genera importantes variaciones que afectan la convergencia y la mezcla de asignaciones en cada grupo. Por último, si no se dispone de información sobre σ_h^2, es natural asumir que cada a_h es tal que $a_h = 2$ porque en consecuencia $\mathbb{V}\text{ar}(\sigma_h^2) \to \infty$. Ante la posible heterogeneidad de varianza inherente al negocio asegurador y presente en cada clase latente, además de la homogeneidad dentro de cada grupo, se propone que cada a_h sea el producto entre h y un valor 'pequeño' E_0 para representar la homogeneidad dentro de los grupos y h:

   $$a_h = h \cdot E_0,$$

   donde $E_0 = 0,09$. Observe que el valor esperado de la varianza para cada grupo es proporcional a h con $h \in \{1, \ldots, H\}$, y de ese modo se garantiza la heterocedasticidad.

2. De forma análoga a la primera propuesta se tiene que: $n_0 = 1$, las componentes de $\boldsymbol{\mu}$ son las medias muestrales del logaritmo de las cuantías de cada clase y las componentes de $\boldsymbol{\tau}$ toman el valor de 1. El desconocimiento de σ_h^2 permite proponer de manera razonable que es realización de una variable aleatoria inversa gamma con coeficiente de variación del 100 %. En tal caso, cada a_h es tal que $a_h = 3$. Bajo el mismo principio de heterocedasticidad de varianza y homogeneidad entre grupos, cada b_h es tal que:

   $$b_h = h \cdot (a_h - 1) \cdot E_0,$$

   donde $E_0 = 0,09$.

3. En esta propuesta se conservan los valores de n_0, $\boldsymbol{\mu}$ y $\boldsymbol{\tau}$ de las dos propuestas anteriores. Si no se dispone de información sobre cada σ_h^2, es razonable proponer que es una realización de una variable aleatoria con distribución inversa gamma de coeficiente de variación del 70 %. Por tanto, el valor aproximado de a_h es tal que $a_h = 4$. Ante la heterogeneidad de varianza y homogeneidad de los grupos para las cuantías de los siniestros, cada b_h para todo $h \in \{1, \ldots, H\}$ es tal que:

   $$b_h = h \cdot (a_h - 1) \cdot E_0,$$

   donde $E_0 = 0,09$.

Algoritmo MCMC.

A continuación, se presentan los pasos a seguir para implementar computacionalmente el algoritmo MCMC para el modelo Bayesiano de mezcla con distribución lognormal y número de componentes aleatorio. Este algoritmo requiere de un Metrópolis adaptativo para cada λ_h dado que su distribución condicional completa no tiene el kernel de una forma conocida:

1. En la iteración inicial, muestrear una observación para θ_h y σ_h^2 para todo $h \in \{1, \ldots, H\}$ de manera que:

$$\begin{aligned} \theta_h^{(0)} &\leftarrow \text{Normal}(\mu_h, \tau_h^2) \\ (\sigma_h^2)^{(0)} &\leftarrow \text{Inv-Gamma}(a_h, b_h). \end{aligned}$$

   Adicionalmente, guardar $\boldsymbol{\omega}^{(0)}$ tal que:

$$\boldsymbol{\omega}^{(0)} \leftarrow \left(\frac{n_1^{(0)}}{\sum_{h=1}^{H} n_h^{(0)}}, \ldots, \frac{n_H^{(0)}}{\sum_{h=1}^{H} n_h^{(0)}} \right),$$

   donde $n_h^{(0)}$ es el tamaño inicial de la clase latente h para todo $h \in \{1, \ldots, H\}$.

2. Seleccionar valores para $\delta_1^2, \ldots, \delta_H^2$ en $\mathbb{R}^+$ para optimizar el algoritmo de Metrópolis con tasas de aceptación entre el 20 % y el 50 %.

3. Para la iteración del paso k con $k \geq 1$, realizar el siguiente procedimiento:

   - Guardar una matriz de probabilidades $\boldsymbol{P}_{n \times H}$ de tamaño $n \times H$, de manera que cada componente p_{ih} es tal que:

$$p_{ih} \leftarrow p\left(Z_i^{(k)} = h \mid \text{resto}\right) = \frac{\omega_h^{(k-1)} \text{lognormal}\left(x_i \mid \theta_h^{(k-1)}, (\sigma_h^2)^{(k-1)}\right)}{\sum_{l=1}^{H} \omega_l^{(k-1)} \text{lognormal}\left(x_i \mid \theta_l^{(k-1)}, (\sigma_l^2)^{(k-1)}\right)}.$$

   - Guardar un vector $\boldsymbol{m}_{n \times 1}$, donde cada componente m_i es el máximo de cada fila de la matriz $P_{n \times H}$:

$$\boldsymbol{m}^T \leftarrow (m_1, \ldots, m_n),$$

     donde m_i es tal que:

$$m_i = \text{máx}\left(p\left(Z_i^{(k)} = 1 \mid \text{resto}\right), \ldots, p\left(Z_i^{(k)} = H \mid \text{resto}\right)\right).$$

   - Construir la matriz de probabilidades $\boldsymbol{P^*}_{n \times H}$ de tamaño $n \times H$ donde cada componente p_{ih}^* es tal que:

$$p_{ih}^* \leftarrow \frac{\exp\{p_{ih} - mi\}}{\sum_{l=1}^{H} \exp\{p_{il} - mi\}}.$$

   - Muestrear las asignaciones de clúster para cada observación x_i con el vector de probabilidades $\boldsymbol{p}^*$:

$$(\boldsymbol{p}^*)^T = (p_{i1}^*, \ldots, p_{iH}^*).$$

     Note que este vector de probabilidades generar mayor estabilidad computacional al momento de muestrear la clase de cada observación. Paso seguido a lo anterior, guardar $\boldsymbol{N}_h^{(k)}$ tal que:

$$\boldsymbol{N}_h^{(k)} \leftarrow \left(n_1^{(k)}, \ldots, n_H^{(k)}\right),$$

donde $n_h^{(k)}$ es tal que $n_h^{(k)} = \#\{x_i : x_i \in Z_h\}$, para todo $h \in \{1, \ldots, H\}$.

- Obtener una muestra para $\theta_h^{(k)}$ | resto de manera que:

$$\theta_h^{(k)} \mid \text{resto} \overset{\text{d}}{=} \text{Normal}\left(\widehat{m}_{h,(k-1)}\widehat{V}_{h,(k-1)}^{-1}, \widehat{V}_{h,(k-1)}^{-1}\right),$$

donde

$$\widehat{m}_{h,(k-1)} = \frac{\sum_{i:Z_i \in h} \log(x_i)}{(\sigma_h^2)^{(k-1)}} + \frac{\mu_h}{\tau_h^2}$$
$$\widehat{V}_{h,(k-1)} = \frac{n_h^{(k)}}{(\sigma_h^2)^{(k-1)}} + \frac{1}{\tau_h^2},$$

para todo $h \in \{1, \ldots, H\}$.

- Obtener una muestra para $(\sigma_h^2)^{(k)}$ | resto a partir de:

$$(\sigma_h^2)^{(k)} \mid \text{resto} \overset{\text{d}}{=} \text{Inv-Gamma}\left(\frac{n_h^{(k)}}{2} + a_h, \frac{1}{2}\sum_{i:Z_i \in h}\left(\log(x_i) - \theta_h^{(k)}\right)^2 + b_h\right),$$

para todo $h \in \{1, \ldots, H\}$.

- Obtener una muestra para $\boldsymbol{\omega}^{(k)}$ | resto con $\boldsymbol{\omega}^{(k)} = \left(\lambda_1^{(k)}, \ldots, \lambda_H^{(k)}\right)$ utilizando las pasos del algoritmo del Metrópolis.

  a) Guardar $\boldsymbol{\Phi}_h^{(k-1)}$ como:

$$\boldsymbol{\Phi}^{(k-1)} \leftarrow \left(\Phi_1^{(k-1)}, \ldots, \Phi_H^{(k-1)}\right),$$

  donde $\Phi_h^{(k-1)} = \log\left(\lambda_h^{(k-1)}\right)$ para todo $h \in \{1, \ldots, H\}$.

  b) Para todo $h \in \{1, \ldots, H\}$, generar una muestra para Φ_h^* tal que $\Phi_h^* \overset{\text{d}}{=} \text{Normal}\left(\Phi_h^{(k-1)}, \delta_h^2\right)$.

  c) Computar el logaritmo de la razón de aceptación $r_{k,h}$ para todo $h \in \{1, \ldots, H\}$ de la siguiente manera:

$$\log(r_{k,h}) = \left(\frac{n_0}{H} + n_h^{(k)}\right)\left(\Phi_h^* - \Phi_h^{(k-1)}\right) - n_h^{(k)}\log\left(\frac{m_h + \Phi_h^*}{m_h + \Phi_h^{(k-1)}}\right) - \left(e^{\Phi_h^*} - e^{\Phi_h^{(k-1)}}\right),$$

  donde m_h es tal que:

$$m_h = \begin{cases} \sum_{l=2}^{H} \lambda_l^{(k-1)} & \text{, si } h = 1 \\ \sum_{l=1}^{h-1} \lambda_l^{(k)} + \sum_{j=h+1}^{H} \lambda_j^{(k-1)} & \text{, si } 2 \leq h \leq H-1 \\ \sum_{l=1}^{H-1} \lambda_l^{(k)} & \text{, si } h = H. \end{cases}$$

  d) Para todo $h \in \{1, \ldots, H\}$, generar un número aleatorio $u_{k,h}$ tal que $u_{k,h} \overset{\text{d}}{=} \text{Uniforme}(0,1)$. Calcular el estado actual para $\lambda_h^{(k)}$ de la siguiente manera:

$$\lambda_h^{(k)} = \begin{cases} e^{\Phi^*} & \text{, si } \log(u_{k,h}) \leq \log(r_{k,h}) \\ \lambda_h^{(k-1)} & \text{, en otro caso.} \end{cases}$$

e) Guardar $\boldsymbol{\omega}^{(k)}$ como:

$$\boldsymbol{\omega}^{(k)} \leftarrow \left(\frac{\lambda_1^{(k)}}{\sum_{h=1}^{H} \lambda_h^{(k)}}, \ldots, \frac{\lambda_H^{(k)}}{\sum_{h=1}^{H} \lambda_h^{(k)}} \right).$$

4. Guardar el vector de parámetros $\boldsymbol{\theta}^{(k)}$ tal que:

$$\boldsymbol{\theta}^{(k)} \leftarrow \left(\theta_1^{(k)}, \ldots, \theta_H^{(k)}, (\sigma_1^2)^{(k)}, \ldots, (\sigma_H^2)^{(k)}, \omega^{(k)} \right).$$

5. Guardar $H_n^{(k)}$ tal que:

$$H_n^{(k)} \leftarrow \sum_{h=1}^{H} 1_{n_h>0}.$$

Repetir los pasos 3, 4 y 5 del proceso anterior en un número de iteraciones $S = 210.000$ calibrando los valores $\{\delta_1^2, \ldots, \delta_H^2\}$ para todo $h \in \{1, \ldots, H\}$ que optimicen las tasas de aceptación del algoritmo de Metrópolis. En la siguiente etapa, quemar las primeras 10.000 observaciones para muestrear sistemáticamente las restantes con salto $a = 2$. El número total de componentes o grupos de riesgos asegurados es la moda de H_n. Observe que en cada iteración del algoritmo MCMC se realiza la suma H variables indicadoras que toman el valor de uno solamente si el h-ésimo grupo no se encuentra vacío.

Estudio de simulación

Para el estudio de simulación se utilizará el mismo conjunto de datos generado a partir de los valores de los parámetros reportados en la tabla 10. De acuerdo con la tabla 11 donde se reporta el criterio Bayesiano BIC posterior a la implementación computacional del algoritmo EM, se proponen 6 clases latentes para las asignaciones iniciales de las 1.500 observaciones simuladas para las cuantías de los siniestros.

Una vez implementado computacionalmente el MCMC bajo un algoritmo de Metrópolis adaptativo con la primera propuesta de distribución previa no informativa presentada anteriormente, se obtuvieron los siguientes resultados de tasas de aceptación para cada λ_h al momento de optimizar su respectivo δ_h con $h \in \{1, \ldots, 6\}$:

Tabla 25: *La primera columna el parámetro del modelo Bayesiano para el cual se utiliza un algoritmo de Metrópolis adaptativo, la segunda columna presenta el respectivo valor de δ_h que permite optimizar el algoritmo de Metrópolis y la última columna muestra la tasa de aceptación.*

Parámetro h	δ_h	Tasa de aceptación
λ_1	0,03	$33,75545\,\%$
λ_2	0,02	$35,86091\,\%$
λ_3	0,05	$33,63545\,\%$
λ_4	0,06	$35,02636\,\%$
λ_5	0,08	$32,74182\,\%$
λ_6	0,2	$38,85455\,\%$

En este caso, las tasas de aceptación presentan valores entre el $32,74\,\%$ y el $38,85455\,\%$ de manera que optimizan el desempeño del algoritmo de Metrópolis al momento de obtener muestras de una distribución no conocida para cada parámetro λ_h. Por otra parte, la bondad del ajuste del modelo mostró 'optimalidad' al momento de verificar los valores p predictivos en la media, la varianza, la desviación estándar y el coeficiente de variación tal y como se observa en la siguiente tabla:

Tabla 26: *La primera columna muestra el estadístico de bondad de ajuste para el modelo Bayesiano Binomial Negativo para la frecuencia de los siniestros y la segunda columna su respectivo valor p predictivo.*

Estadístico	PPI
Media	0,46718
Varianza	0,46552
Coeficiente de variación	0,47296
Desviación estándar	0,46552

Estos resultados para los valores p predictivos se encuentran alrededor de $0,5$ y demuestran que el modelo Bayesiano presenta un 'buen' ajuste al evaluar cada una de las estadísticas de la tabla. Las cadenas de Markov para cada una de las medias, varianzas y pesos del modelo de mezcla una vez se utilizan 110.000 iteraciones con una etapa de quemado para las primeras 10.000 observaciones y un muestreo sistemático con salto $a = 2$:

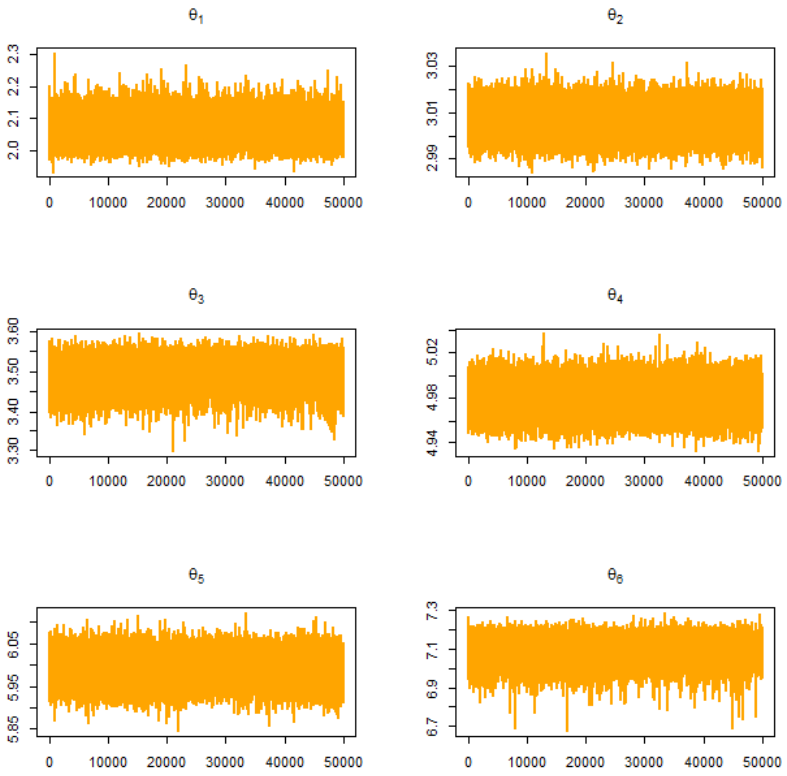

Figura 40: Cadenas de Markov posteriores a etapa de quemado y muestreo sistemático para vector de medias en clases latentes construidas según los montos de cuantía de los siniestros.

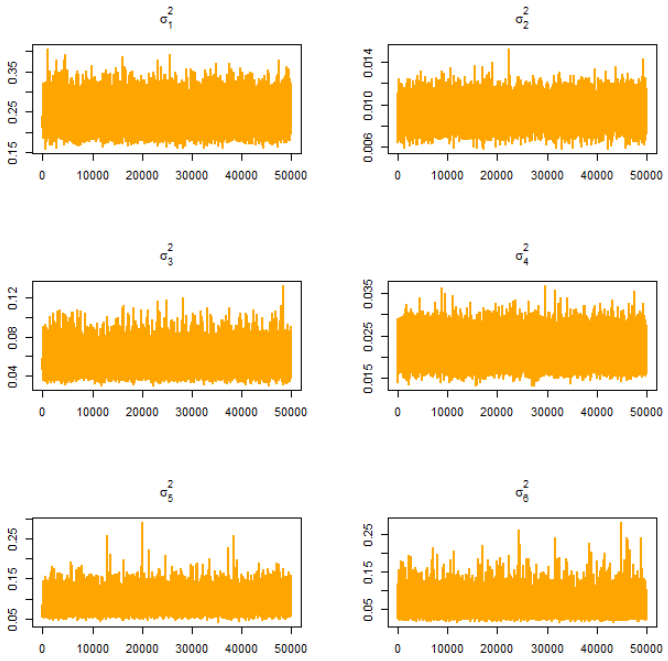

Figura 41: Cadenas de Markov posteriores a etapa de quemado y muestreo sistemático para vector de varianzas en clases latentes construidas según los montos de cuantía de los siniestros.

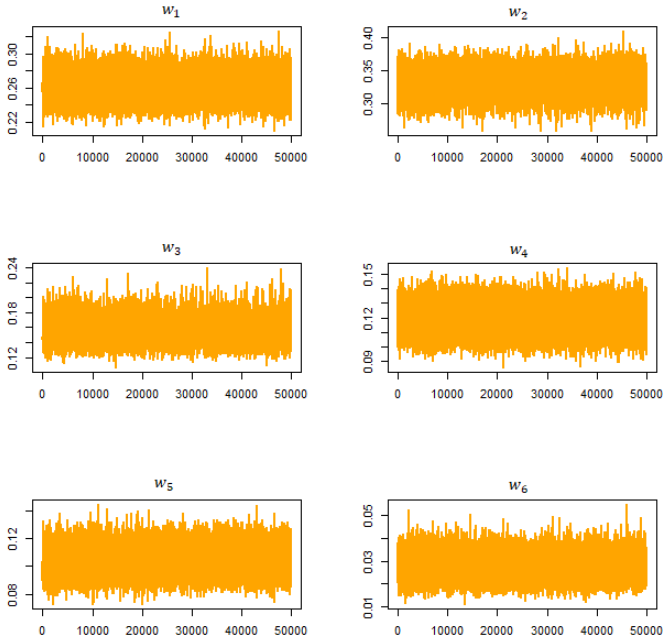

Figura 42: Cadenas de Markov posteriores a etapa de quemado y muestreo sistemático para vector de pesos en clases latentes construidas según los montos de cuantía de los siniestros.

Los gráficos mostrados anteriormente permiten apreciar de manera descriptiva la estacionariedad en

las cadenas de Markov de cada uno de los parámetros del modelo Bayesiano. Al parecer no existen problemas de suficiencia de muestras por parámetro, pero esto se confirma por medio de la siguiente tabla que reporta los tamaños efectivos de muestra:

Tabla 27: *La primera columna presenta la clase latente, mientras que la columnas dos, tres y cuatro reportan los tamaños efectivos de muestra para media, varianza y peso por cada clase latente.*

Clase latente h	Media θ_h	Varianza σ_h^2	Peso ω_h
Z_1	12103,502	12161,285	10332,311
Z_2	15551,666	8117,72	5163,7
Z_3	4350,965	4350,965	4335,228
Z_4	41299,572	37818,753	18990,852
Z_5	27901,39	11176,42	12690,829
Z_6	11954,355	11845,976	11531,464

En términos de los tamaños efectivos de muestra se aprecia que, en efecto, las muestras de cada parámetro son suficientes para inferencia posterior y la convergencia de las cadenas de Markov a sus respectivas medias posteriores. En comparación con el modelo Bayesiano de mezcla con número de componentes fijo presentado en la sección anterior, se tienen menores tamaños efectivos de muestra para el modelo Bayesiano de mezcla con número de componentes aleatorio y esto se atribuye al hecho de implementar un algoritmo de Metrópolis adaptativo optimizado para generar tasas de aceptación en los parámetros de distribución desconocida alrededor del 35 %.

En términos del número componentes óptimo para el modelo de mezcla entre 1 y $H = 6$, se presenta a continuación el gráfico de moda posterior para H_n:

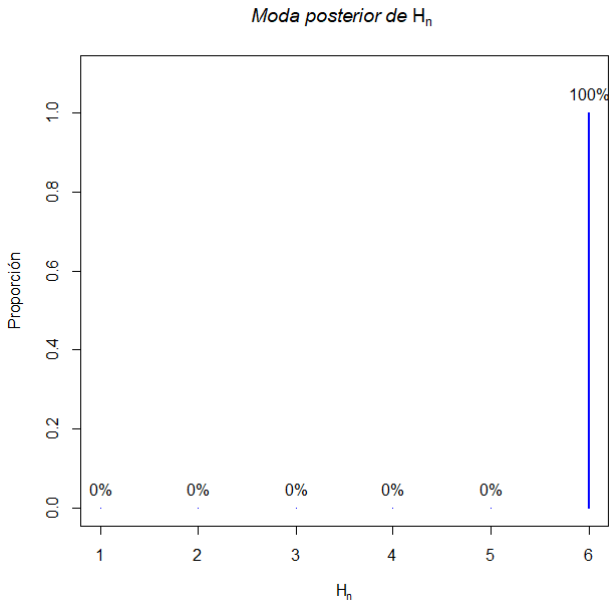

Figura 43: Distribución discreta para H_n.

El gráfico presentado anteriormente permite concluir que el número óptimo de clústeres entre 1 y $H = 6$ es 6 puesto que es la moda posterior de clústeres no vacíos en la implementación de las iteraciones

computacionales del algoritmo MCMC de Metrópolis adaptativo para modelo Bayesiano de mezcla con número de componentes aleatorio. Note que en el 100 % de las iteraciones, ninguno de los clústeres estuvo vacío.

Finalmente, a continuación, se presentan las estimaciones de parámetros para medias, varianzas y pesos con sus respectivos intervalos de credibilidad del 95 % en cada una de las clases latentes propuestas desde la asignación inicial para la implementación computacional del modelo Bayesiano de mezcla con número de componentes aleatorio:

Tabla 28: *La primera columna muestra la clase latente. La segunda y cuarta columna muestran los límites inferior y superior respectivamente para el intervalo de credibilidad del* 95 % *para la media posterior. Las columnas tres y cinco muestran la media posterior y la media real. La última columna es una función indicadora que toma el valor de 1 en caso de que la media real se encuentre dentro del intervalo de credibilidad propuesto y 0 en caso contrario.*

Clase	límite inferior	Media Posterior	Límite superior	Media θ_h	Indicadora
Z_1	2,001169	2,070084	2,148853	2,0	0
Z_2	2,995792	3,006683	3,017685	3,0	1
Z_3	3,422956	3,495377	3,550917	3,5	1
Z_4	4,955878	4,979045	5,002231	5,0	1
Z_5	5,933789	5,990892	6,047153	6,0	1
Z_6	6,976659	7,092774	7,187513	7,0	1

Tabla 29: *La primera columna muestra la clase latente. La segunda y cuarta columna muestran los límites inferior y superior respectivamente para el intervalo de credibilidad del* 95 % *para la varianza posterior. Las columnas tres y cinco muestran la varianza posterior y la varianza real. La última columna es una función indicadora que toma el valor de 1 en caso de que la varianza real se encuentre dentro del intervalo de credibilidad propuesto y 0 en caso contrario.*

Clase	límite inferior	Varianza Posterior	Límite superior	Varianza σ_h^2	Indicadora
Z_1	0,193575	0,241143	0,302615	0,22	1
Z_2	0,007347	0,009140	0,011153	0,01	1
Z_3	0,040095	0,056697	0,080828	0,05	1
Z_4	0,017077	0,021600	0,027122	0,02	1
Z_5	0,061117	0,089587	0,130564	0,08	1
Z_6	0,028787	0,055888	0,104685	0,10	1

Los resultados de las estimaciones en este modelo presentan grandes 'similitudes' con respecto a las estimaciones del modelo Bayesiano de mezcla con número de componentes fijo. Análogamente, la media real de la primera clase tampoco se encuentra contenido dentro del intervalo de credibilidad del 95 % propuesto para su respectiva media posterior.

Al momento de evaluar el potencial que tiene el modelo Bayesiano de mezcla con distribución lognormal y número de componentes aleatorio para recuperar y reconstruir los clústeres iniciales propuestos en la simulación, se establece un paralelo comparativo entre la matriz de adyacencia con las asignaciones iniciales y la matriz de adyacencia de probabilidades posteriores que tienen dos observaciones de cuantías de siniestros cualesquiera en tener una misma asignación de clúster.

Debido a lo anterior la matriz de adyacencia de asignaciones iniciales y originales de la simulación presenta solamente dos tonalidades: rojo si las observaciones están juntas o blanco en caso contrario. Para la matriz de probabilidades posteriores estimadas por el modelo Bayesiano de número de componentes aleatorios, se observarán las tonalidades mencionadas bajo el mismo criterio, pero en adición, existe una tonalidad

Tabla 30: *La primera columna muestra la clase latente. La segunda y cuarta columna muestran los límites inferior y superior respectivamente para el intervalo de credibilidad del* 95 % *para el peso posterior. Las columnas tres y cinco muestran el peso posterior y el peso real por clase latente. La última columna es una función indicadora que toma el valor de 1 en caso de que el peso se encuentre dentro del intervalo de credibilidad propuesto y 0 en caso contrario.*

Clase	límite inferior	Peso Posterior	Límite superior	Peso ω_h	Indicadora
Z_1	0,236276	0,261819	0,289887	0,25	1
Z_2	0,295081	0,330910	0,364393	0,35	1
Z_3	0,131030	0,157191	0,189818	0,15	1
Z_4	0,102418	0,118777	0,135908	0,12	1
Z_5	0,088065	0,104213	0,121671	0,10	1
Z_6	0,018375	0,027090	0,037431	0,03	1

de color amarillo que representa un valor entre cero y uno dado que corresponde a una probabilidad. En tal sentido, en la matriz de probabilidades el color rojo muestra que las observaciones permanecen juntas en una misma asignación (no necesariamente constante) en el proceso iterativo del algoritmo MCMC, mientras que observaciones de tonalidad amarilla no permanecieron siempre en con la misma asignación en las iteraciones. A continuación, los resultados de las matrices de adyacencia:

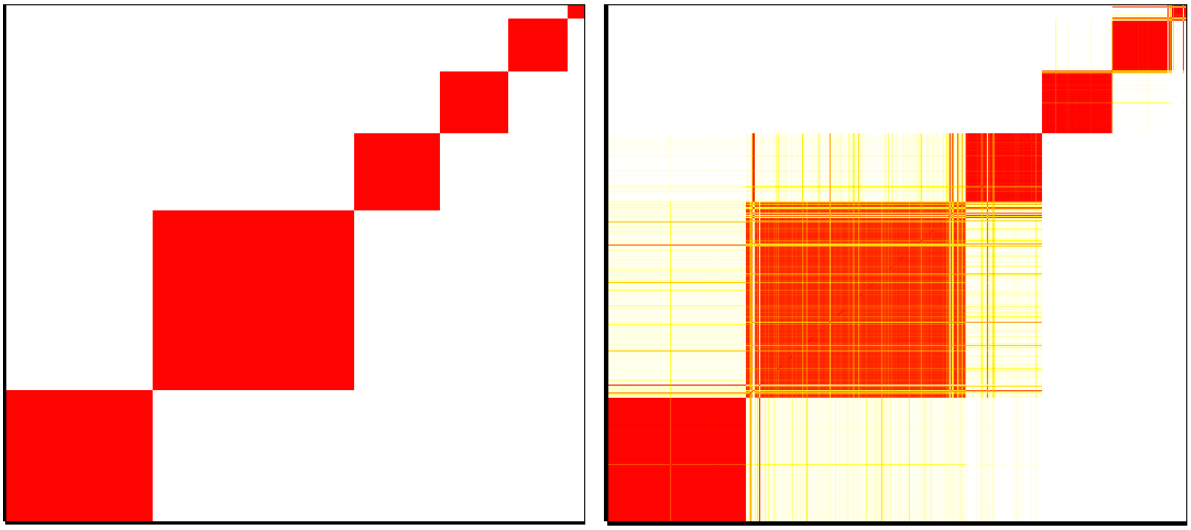

Figura 44: La matriz del panel izquierdo de la visualización muestra las asignaciones originales controladas en la simulación y la matriz del panel derecho presenta la estimación de las probabilidades posteriores de que dos observaciones cualesquiera tengan la misma asignación.

En términos descriptivos, las visualizaciones permiten concluir que el modelo Bayesiano propuesto con número de componentes aleatorio para la mezcla permite reproducir en 'gran medida' los clústeres originales propuestos en la simulación de las observaciones de las cuantías de los siniestros.

Implementación computacional del modelo para la severidad de los siniestros.

La implementación computacional del modelo Bayesiano con número de componentes aleatorio tendrá en cuenta la misma propuesta inicial de asignaciones de clúster para las observaciones de las cuantías de los siniestros que se utilizó para el modelo Bayesiano de número de componentes fijo con el algoritmo EM. De esta manera, el modelo de componentes con número de componentes aleatorio indicará por medio de la moda posterior de clústeres que no se encuentran vacíos en cada iteración del algoritmo

MCMC, si existe un número óptimo de clústeres que no necesariamente es $H = 5$ puesto que puede estar asociado a un h tal que $1 \leq h \leq H = 5$. El modelo será implementado con 210.000 iteraciones realizando una etapa de quemado con las 10.000 primeras observaciones y posteriormente un muestreo sistemático con salto de $a = 2$. Los valores óptimos de δ_h^2 para todo $h \in \{1, \ldots, H\}$ para tener tasas de aceptación entre el 20 % y el 50 % cuando se requiere del algoritmo de Metrópolis adaptativo son tales que: $\{\delta_1^2 = 0.01, \delta_2^2 = 0.02, \delta_3^2 = 0.01, \delta_4^2 = 0.01, \delta_5^2 = 0.01\}$. Finalmente, se consideran las tres propuestas de distribuciones previas no informativas descritas anteriormente. Los resultados de los tamaños efectivos de muestra al implementar el algoritmo MCMC con cada propuesta se presentan a continuación. En cada tabla, la primera columna presenta la clase latente, la segunda columna los tamaños efectivos para θ_h, la tercera columna los tamaños efectivos de σ_h^2 y la última columna los tamaños efectivos de w_h:

Tabla 31: *Tabla de tamaños efectivo bajo propuesta no informativa 1.*

Clase	θ_h	σ_h^2	w_h
Z_1	95425,55	100000,00	42459,54
Z_2	62019,47	42832,97	14795,90
Z_3	664,04	931,97	538,35
Z_4	498,79	390,31	275,29
Z_5	287,06	394,56	279,30

Tabla 32: *Tabla de tamaños efectivo bajo propuesta no informativa 2.*

Clase	θ_h	σ_h^2	w_h
Z_1	98989,90	92330,34	41334,57
Z_2	59137,28	41308,06	14527,05
Z_3	540,61	689,58	457,35
Z_4	249,52	394,22	283,74
Z_5	398,78	497,23	284,55

Tabla 33: *Tabla de tamaños efectivo bajo propuesta no informativa 3.*

Clase	θ_h	σ_h^2	w_h
Z_1	96275,45	96285,01	42174,36
Z_2	47530,05	29537,36	8191,30
Z_3	326,72	417,58	333,49
Z_4	251,69	371,78	230,06
Z_5	182,17	251,01	185,49

Los resultados de los tamaños efectivos de muestra de los parámetros bajo cada propuesta en comparación con los tamaños efectivos del modelo Bayesiano con número de componentes fijo tienden a ser aproximadamente la mitad de estos dado que en este caso existe un efecto de muestreo de las observaciones explicado por las tasas de aceptación del algoritmo adaptativo de Metrópolis. Las estimaciones de las medias posteriores de cada parámetro con sus respectivos límites del intervalo de credibilidad del 95 % se presentan a continuación:

Tabla 34: *Tabla de estimación de parámetros bajo la propuesta no informativa 1.*

Parámetro	Límite inferior	Media Posterior	Límite superior
θ_1	5,29752	5,29871	5,29990
θ_2	5,90018	5,90657	5,91307
θ_3	5,99107	6,11492	6,26446
θ_4	6,81667	7,02313	7,25270
θ_5	7,91581	8,18169	8,58485
σ_1^2	0,00024	0,00026	0,00029
σ_2^2	0,00339	0,00399	0,00469
σ_3^2	0,09859	0,14209	0,20024
σ_4^2	0,17138	0,31877	0,46767
σ_5^2	0,53485	0,74250	0,90375
w_1	0,14538	0,15567	0,16630
w_2	0,11762	0,13033	0,14343
w_3	0,08547	0,14521	0,21632
w_4	0,15336	0,27598	0,39994
w_5	0,17185	0,29281	0,38778

Tabla 35: *Tabla de estimación de parámetros bajo la propuesta no informativa 2.*

Parámetro	Límite inferior	Media Posterior	Límite superior
θ_1	5,29734	5,29903	5,30073
θ_2	5,90170	5,90891	5,91625
θ_3	6,00606	6,14443	6,31722
θ_4	7,91980	8,19794	8,59836
θ_5	6,83643	7,04775	7,30752
σ_1^2	0,00048	0,00053	0,00059
σ_2^2	0,00442	0,00521	0,00616
σ_3^2	0,10533	0,15436	0,22220
σ_4^2	0,53253	0,73816	0,90778
σ_5^2	0,19447	0,32891	0,46783
w_1	0,14570	0,15612	0,16676
w_2	0,12202	0,13529	0,14856
w_3	0,08520	0,14686	0,22558
w_4	0,16748	0,28555	0,38445
w_5	0,15636	0,27618	0,39071

Tabla 36: *Tabla de estimación de parámetros bajo la propuesta no informativa 3.*

Parámetro	Límite inferior	Media Posterior	Límite superior
θ_1	5,29730	5,29939	5,30148
θ_2	5,90332	5,91133	5,91953
θ_3	6,00997	6,17285	6,39556
θ_4	6,83608	7,08235	7,44483
θ_5	7,93451	8,26573	8,83307
σ_1^2	0,00072	0,00080	0,00090
σ_2^2	0,00550	0,00655	0,00780
σ_3^2	0,10924	0,16598	0,25337
σ_4^2	0,20130	0,34709	0,52424
σ_5^2	0,45185	0,70606	0,89568
w_1	0,14619	0,15659	0,16719
w_2	0,12590	0,13987	0,15405
w_3	0,07817	0,14842	0,24362
w_4	0,16061	0,28900	0,41688
w_5	0,11267	0,26612	0,37891

Los resultados presentados anteriormente para la estimación de parámetros del modelo Bayesiano con número de componentes aleatorio resultan ser similares a las estimaciones obtenidas para el modelo Bayesiano con número de componentes fijo. Por supuesto esto es un indicio de que el número óptimo de clústeres es 5. A continuación se presentan las tasas de aceptación para los pesos de cada clase considerando los valores δ_h^2 utilizados anteriormente para el algoritmo adaptativo de Metrópolis:

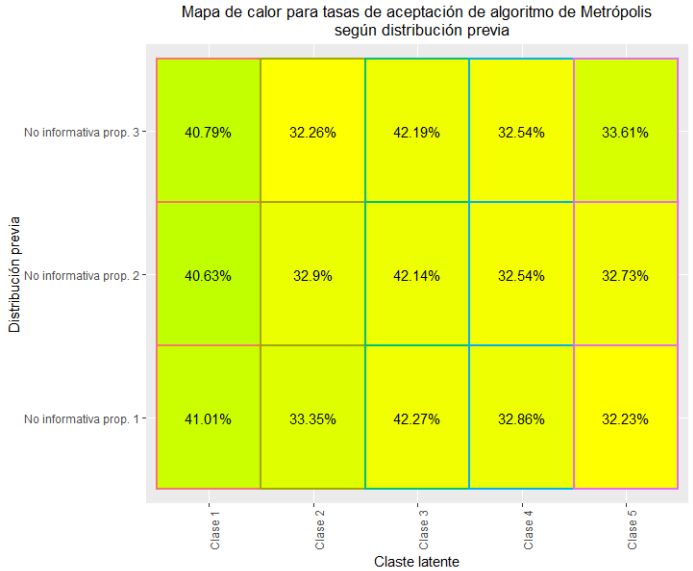

Figura 45: Matriz de tasas de aceptación del algoritmo de Metrópolis según los pesos de cada clase y la distribución previa no informativa implementada

La matriz presentada anteriormente permite concluir que las distribuciones previas no informativas utilizadas para implementar computacionalmente el modelo Bayesiano de componentes fijo reportan tasas de aceptación para los pesos de las clases con el algoritmo de Metrópolis entre $32,23\,\%$ y $42,27\,\%$. Por otra parte, la bondad del ajuste del modelo considerando los valores p predictivos en estadísticas como la media, la varianza, la desviación estándar y el coeficiente de variación permitió apreciar un ajuste 'óptimo' bajo las tres distribuciones previas no informativas propuestas porque los valores encuentran alrededor de $0,5$ tal y como se evidencia a continuación:

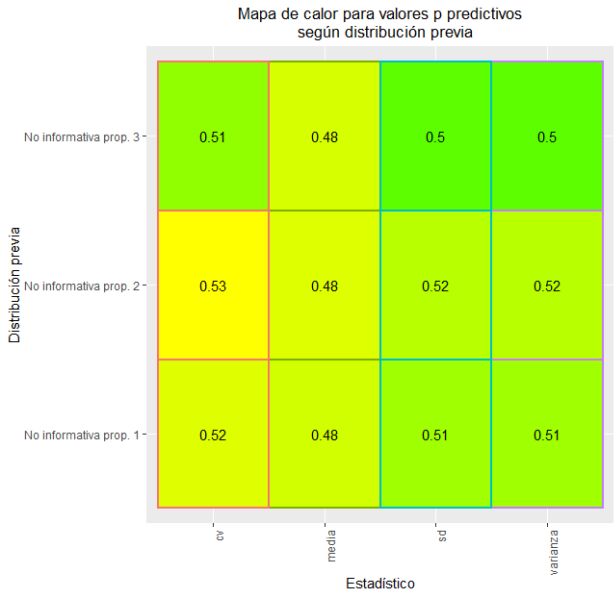

Figura 46: Matriz de tasas de aceptación del algoritmo de Metrópolis según los pesos de cada clase y la distribución previa no informativa implementada.

Las modas posteriores para encontrar el número óptimo de clústeres entre 1 y $H = 5$ por cada modelo Bayesiano se aprecian por medio de las siguientes distribuciones discretas:

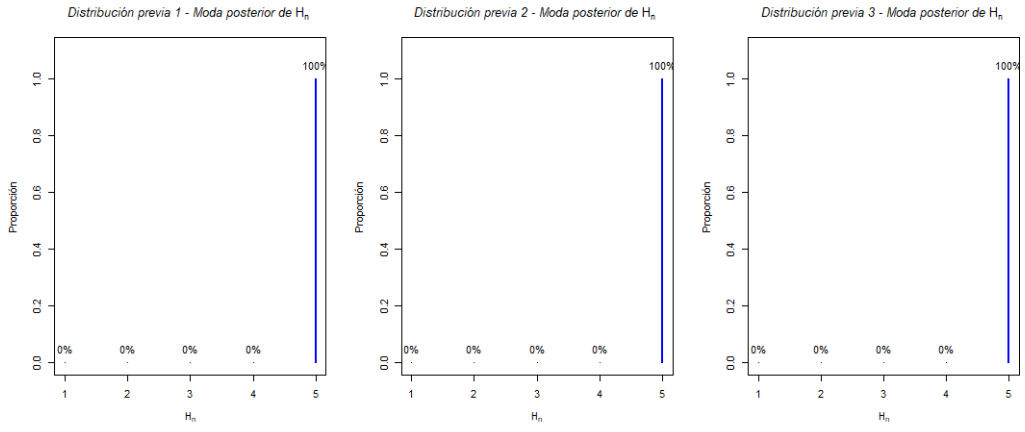

Figura 47: Moda posterior de H_n para modelo Bayesiano con número de componentes aleatorio bajo cada distribución previa no informativa.

Este resultado permite confirmar la hipótesis que se tenía previamente sobre el número óptimo de clústeres para el modelo Bayesiano cuando se observó que existen 'grandes' similitudes entre los resultados

de las estimaciones de parámetros de los modelos con número de componentes fijo y aleatorio. De esta manera, los gráficos permiten observar que las modas posteriores de H_h son iguales a 5 bajo cada distribución no informativa propuesta, de manera que, el número de componentes óptimo para el modelo es $H = 5$.

Estudio de sensibilidad.

De manera análoga a como se estudió la sensibilidad del modelo Bayesiano de mezcla con número de componentes fijo, se estudiará la sensibilidad del modelo Bayesiano con número de componentes aleatorio. Esto es, analizar las métricas de lpdd, RMSE y MAPE en una validación cruzada de k-fold tal que $k = 5$ para comparar las predicciones de las observaciones muestreadas de la distribución posterior con su valor real para una partición cuando se ensamblan todas las posibles combinaciones de $k - 1$ particiones. El estudio consiste en considerar estas la sensibilidad de estas métricas cuando se genera una variación en los valores de los hiperparámetros de las distribuciones previas no informativas para τ_h^2 y b_h. De este modo $\tau_h^2 \in \{0.5, 1, 1.5\}$ y $b_h \in \{0.01, 0.05, 0.09\}$. En primera instancia se presentarán los valores de las tasas de aceptación del algoritmo de Metrópolis cuando se escogen los mismos valores δ_h^2 utilizados para implementar el modelo de número de componentes aleatorio en la sección anterior:

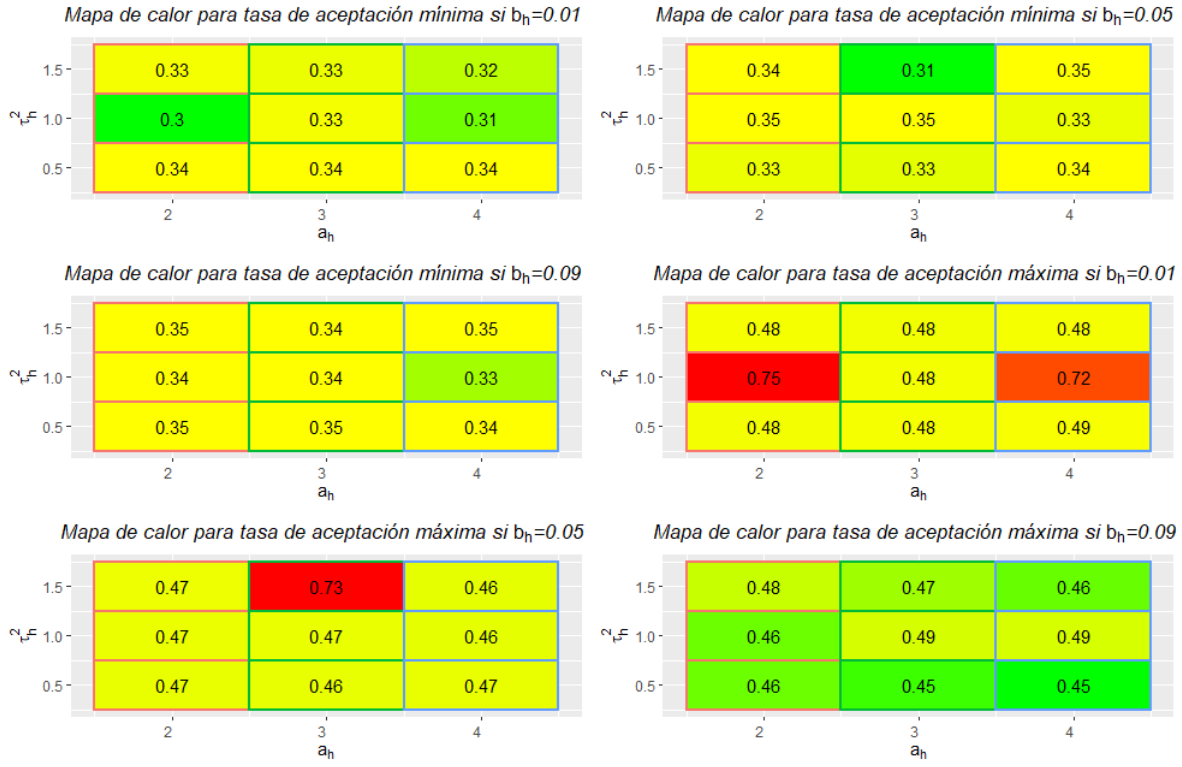

Figura 48: Mapa de calor de tasas de aceptación del algoritmo adaptativo de Metrópolis en validación cruzada de k-fold bajo estudio de sensibilidad para cada distribución previa no informativa.

Los resultados presentados anteriormente permiten evidenciar que cuando $b_h = 0.09$, las tasas de aceptación mínimas y máximas de cada uno de los 5 'split' generados en el proceso de validación cruzada $k-$fold se encuentran aproximadamente entre el 30 % y el 50 %. Sin embargo para valores de b_h tales que $b_h \in \{0.01, 0.05\}$ algunas de las tasas máximas se incrementan un poco más del 70 %.

Respecto a las métricas de lppd, RMSE y MAPE, las siguientes matrices presentan los resultados del estudio de sensibilidad para cada distribución previa no informativa. En términos de los hiperparámetros de las tres propuestas hechas para este modelo se aprecia que el valor de lpdd se minimiza en $-7383,85$ cuando $a_h = 4, \tau_h^2 = 1$ y $b_h = 0.09$. Sin embargo, para la media de RMSE y MAPE en la validación

cruzada, los valores de estas métricas se minimizan en $641,13$US y $5,11\,\%$, respectivamente cuando $a_h = 3, \tau_h^2 = 1$ y $b_h = 0.09$:

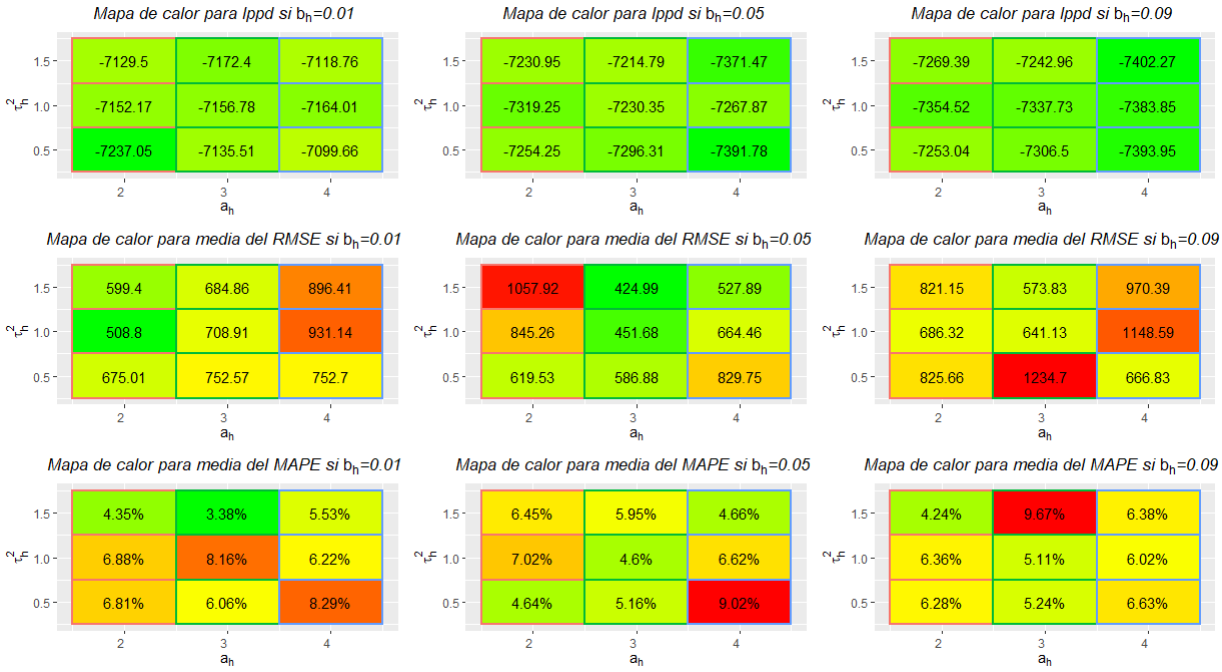

Figura 49: Mapa de calor para métricas de MAPE, RMSE y lppd en validación cruzada k-fold de modelo Bayesiano con número de componentes fijo.

Respecto a la sensibilidad de los hiperparámetros del modelo Bayesiano con número de componentes aleatorio se observa que a medida que aumenta el valor de b_h el valor de la métrica lppd tiende a minimizarse. Sin embargo, para la comparación de las predicciones y las observaciones de las particiones generadas en la validación cruzada de $k-$ fold, los valores tienden a minimizarse cuando $b_h = 0.05$ y $a_h = 3$ independiente del valor de $\tau_h^2 \in \{0.5, 1, 1.5\}$.

5.2. Comparación entre modelos Bayesianos propuestos y modelo clásicos para el estudio de frecuencia y severidad

Al implementar la metodología de k-fold con $k = 5$ tanto para el modelo Bayesiano de mezcla con número de componentes fijo como el modelo de número de componentes aleatorio bajo las tres propuestas de distribución previa no informativa, es importante establecer un paralelo comparativo entre el valor medio del RMSE y MAPE en las muestras de test de cada modelo, así como de los valores que toman los criterios bayesianos de WAIC y DIC.

En estos términos, el mejor modelo bajo el criterio Bayesiano WAIC corresponde al modelo de número de componentes aleatorio bajo la propuesta no informativa 3; mientras que según el criterio Bayesiano DIC, el mejor modelo resulta ser el modelo de número de componentes aleatorio bajo la propuesta 1. Sin embargo, al momento de minimizar las métricas de RMSE y MAPE con k-fold escogiendo $k = 5$, el mejor modelo según el RMSE es el modelo de número de componentes fijo bajo la propuesta no informativa 2; y en contraste con lo anterior, el mejor modelo para la métrica del MAPE es el de número de componentes fijo bajo la propuesta 1. Es necesario anotar que las estimaciones entre los modelos pese a tener un número de componentes fijo o aleatorio fueron 'similares' dado que el número óptimo de clústeres o grupos según el modelo de componentes aleatorio fue de $H = 5$, y con este valor de número de clases latentes, se realizaron las estimaciones de parámetros para los modelos Bayesianos de número

de componentes fijo. Los resultados comparativos en mención se resumen mediante la siguiente tabla:

Tabla 37: *La tabla presenta en su orden de izquierda a derecha en la siguiente información en cada una de sus columnas: Modelo, propuesta de distribución previa no informativa, valor del criterio Bayesiano WAIC, valor del criterio Bayesiano DIC, y medias del RMSE Y MAPE en muestras de test bajo metodología k-fold con k=5.*

Modelo	Distribución Previa	WAIC	DIC	RMSE k-fold	MAPE k-fold
Componentes aleatorio	Propuesta 1	-36256,9414	72514,5649	686,3177025	$6,36263\,\%$
Componentes aleatorio	Propuesta 2	-36510,8432	73021,8728	641,1260072	$5,10612\,\%$
Componentes aleatorio	Propuesta 3	-36667,5522	73334,3305	1148,591381	$6,01919\,\%$
Componentes fijo	Propuesta 1	-36257,0841	72514,8769	790,5117938	$5,05504\,\%$
Componentes fijo	Propuesta 2	-36510,8563	73021,8855	461,0955719	$6,15218\,\%$
Componentes fijo	Propuesta 3	-36667,2123	73333,9618	520,0429362	$5,29817\,\%$

Uno de los objetivos del presente trabajo es establecer una comparación entre los resultados de pérdida agregada por siniestros al utilizar modelos de enfoque de estadística clásica y Bayesiana. Para ello, es necesario en el enfoque clásico, determinar mediante pruebas de hipótesis, las distribuciones que tengan el 'mejor' ajuste sobre los conjuntos de datos de número de siniestros ocurridos y cuantía o severidad de los siniestros.

El sistema de hipótesis para la distribución que se ajuste estadísticamente al conjunto de datos de número de siniestros ocurridos en las 52 semanas tiene las siguientes hipótesis:

- Hipótesis nula: El conjunto de datos de número de siniestros siguen la distribución.
- Hipótesis alternativa: El conjunto de datos de número de siniestros no siguen la distribución.

El criterio de rechazo para la hipótesis nula consiste en hacerlo cuando el p valor es menor que el nivel de significancia α del $5\,\%$. Para realizar la prueba de hipótesis se implementa la prueba chi cuadrado contrastando los valores de las frecuencias del número de siniestros por póliza y las probabilidades teóricas para el número de siniestros bajo una distribución Poisson o Binomial Negativa. Los resultados se presentan a continuación:

Tabla 38: *La tabla presenta en sus columnas de izquierda a derecha la siguiente información: distribución candidata para el ajuste de los datos, valor del estadístico de prueba, número de grados de libertad en prueba chi cuadrado y p valor asociado.*

Distribución	Estadístico	Grados de libertad	p valor
Poisson	$177,15$	4	$p < 2,2 \times 10^{-16}$
Binomial negativa	$0,81273$	4	$0,9367$

Los resultados muestran que, según el criterio de rechazo definido anteriormente, la distribución binomial negativa se ajusta adecuadamente al conjunto de datos de número de siniestros por póliza dado que el p valor asociado es mayor que el nivel de significancia del $5\,\%$. Observe que el número de siniestros ocurridos agregados N en 52 semanas se obtiene a partir del número de siniestros por póliza de la siguiente manera:

$$N = \sum_{i=1}^{M} I_i,$$

donde I_i es una función indicadora para riesgo asegurado i que toma el valor de 1 si la póliza tuvo al menos un siniestro y 0 cero en caso contrario. Observe que M es el total de riesgos asegurados en el portafolio.

Para las cuantías de siniestros se realizan pruebas de bondad de ajuste por medio de distribuciones de colas ligeras y pesadas. En este caso la función de exceso de pérdida permite observar de manera descriptiva si la distribución de cuantías es candidata a ser de cola ligera o cola pesada. A continuación se presenta la función de exceso de pérdida para la cuantía de los siniestros:

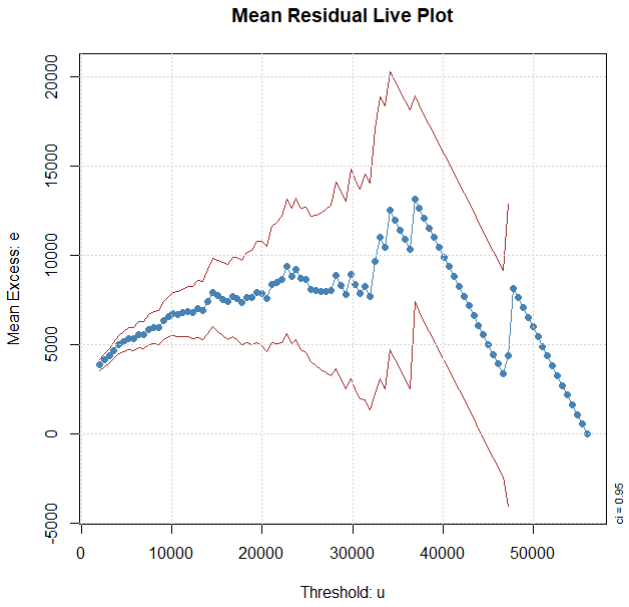

Figura 50: Función de exceso de pérdida para la cuantía de los siniestros con fronteras correspondientes a los límites de un intervalo de confianza del 95 %.

El comportamiento de la función de exceso de pérdida para la cuantía de los siniestros es creciente hasta 30.000US pero a partir de allí comienza a decrecer. Por lo anterior no existe un claro comportamiento que permita diferenciar si la distribución de severidad de los siniestros es de cola ligera o cola pesada puesto que las distribuciones de cola pesada tienen una función de exceso de pérdida creciente y las distribuciones de cola ligera tienen una función de exceso de pérdida decreciente.

De este modo, resulta pertinente realizar pruebas de bondad de ajuste considerando tanto distribuciones de cola ligera como cola pesada, para establecer si alguna se ajusta de forma univariada al conjunto de datos de las cuantías de los siniestros. Sin embargo, muy posiblemente no se tendrá un buen ajuste con alguna de las distribuciones de cola ligera o cola pesada porque el análisis descriptivo y exploratorio mostró que al parecer el conjunto de datos presenta una distribución multimodal.

Por lo anterior, es necesario realizar, por ejemplo, una prueba de bondad de ajuste por clases latentes bajo una distribución lognormal. Estas clases latentes son las mismas que se obtuvieron mediante el algoritmo EM como propuesta inicial para el modelo Bayesiano de mezcla con número de componentes fijo y aleatorio. Adicionalmente, las pruebas de bondad de ajuste se realizarán para las siguientes distribuciones candidatas: Weibull, lognormal, gamma, exponencial, paralogística, loglogística, inversa Pareto, inversa exponencial, Burr e inversa Burr. El sistema de hipótesis de bondad de ajuste tiene las siguientes hipótesis:

- Hipótesis nula: El conjunto de datos de número de siniestros siguen la distribución.
- Hipótesis alternativa: El conjunto de datos de número de siniestros no siguen la distribución.

El criterio de rechazo para la hipótesis nula consiste en hacerlo cuando el p valor es menor que el nivel de significancia α del 5 %. Los test de bondad de ajuste se implementan con las pruebas de Anderson-Darling, Crammer Von Misses y Kolmogorov-Smirnov. Los resultados de los p valores asociados a estas pruebas para cada distribución candidata y para distribuciones logrnormales según clases latentes identificadas bajo el algoritmo EM se presentan a continuación:

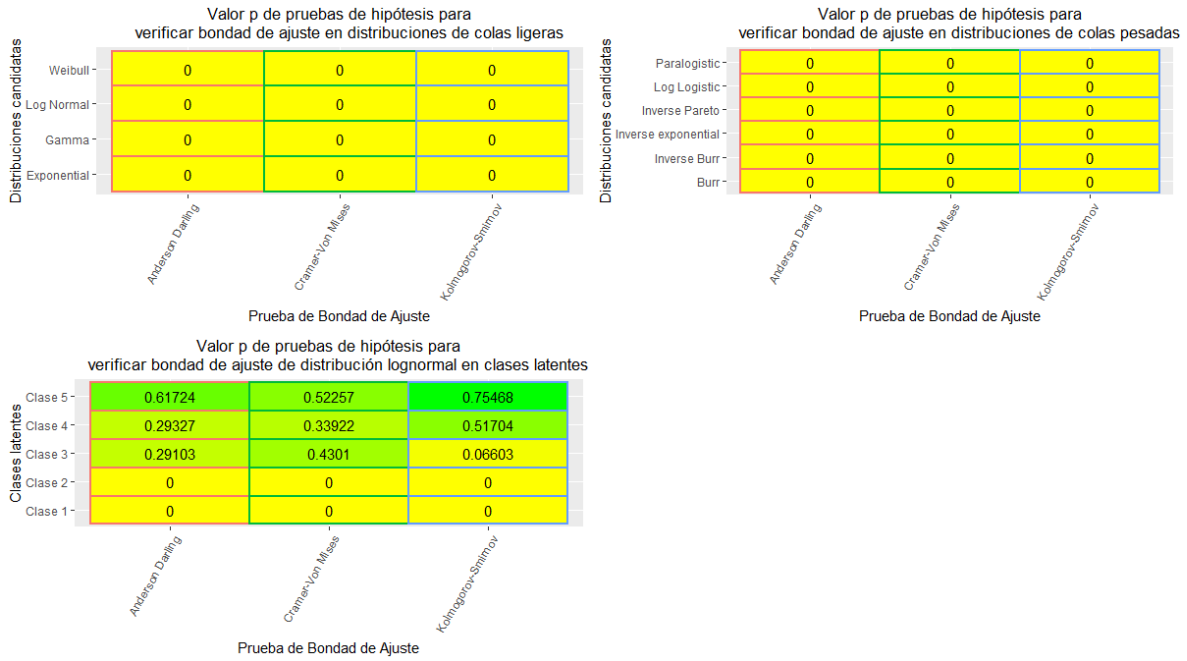

Figura 51: Mapa de calor de p valores para pruebas de bondad de ajuste de distribuciones candidatas para la cuantía de los siniestros bajo test de: Anderson-Darling, Cramer Von Misses y Kolmogorov Smirnov.

Los gráficos anteriores muestran que estadísticamente, ninguna de las distribuciones candidatas tanto para colas ligeras como colas pesadas se ajustan al conjunto de datos de la cuantía de los siniestros dado que el p valor en todos los casos es cero y como este valor es menor que el nivel de significancia del 5 %, se rechaza que el conjunto de datos se ajusta a alguna de las distribuciones propuestas. Por otra parte, cuando se realizan las pruebas de hipótesis sobre las distribuciones lognormales por clase latente, se observa que los datos de las clases 3, 4 y 5 se ajustan estadísticamente para las pruebas de Anderson Darling, Cramer Von Misses y Kolmogorov-Smirnov. Sin embargo, este ajuste de distribución logrnomal estadísticamente hablando no se tiene para los conjuntos de datos de las clases 1 y 2. Entre todas las distribuciones candidatas, el 'mejor' ajuste para la cuantía de los siniestros se tiene cuando se propone una función de densidad de mixturas con componentes lognormal. A continuación, se presentan los resultados de la implementación metodológica de k-fold con $k = 5$ cuando se tiene un modelo clásico de distribución lognormal para las métricas de RMSE y MAPE:

Tabla 39: *La tabla presenta en su primera columna el nombre del indicador métricas promedio para RMSE o MAPE para train o test y en la segunda columna su respectivo valor.*

Indicador	Media
RMSE test score	$736,0652$US
MAPE test score	$7,040264\,\%$
RMSE train score	$337,3806$US
MAPE train score	$2,089818\,\%$

Los resultados demuestran que mediante la metodología de validación $k-$ fold bajo un enfoque de estimación de parámetros con modelos clásicos, la raíz del error cuadrático medio y el MAPE toman valores mayores a los reportados en el estudio de sensibilidad para los modelos Bayesianos de número de componentes fijo y aleatorio. Desde este punto de vista, los modelos Bayesianos propuestos presentan mejores estimaciones para datos faltantes cuando se realiza el proceso de validación cruzada. Sin embargo, resulta vital comparar los valores esperados de pérdida agregada para los siniestros. En términos de enfoque clásico la estimación del valor esperado de pérdida agregada se obtiene mediante Simulación Monte Carlo; y en el caso Bayesiano se obtiene mediante la distribución predictiva utilizando las iteraciones del algoritmo MCMC para muestrear valores de la distribución posterior de los parámetros en cada uno de los modelos. A continuación, se presentan los escenarios de pérdida agregada total por cuantía de los siniestros a los cuales debe hacer frente la compañía aseguradora bajo el enfoque de modelos clásicos y Bayesianos:

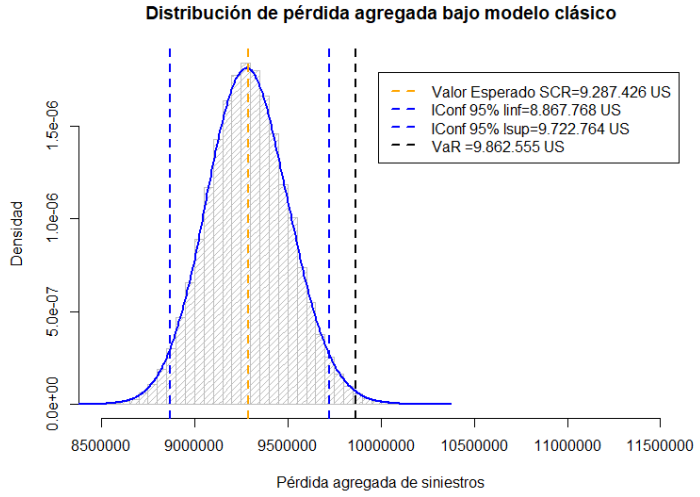

Figura 52: Distribución de pérdida agregada de siniestros bajo enfoque clásico mediante Simulación Monte Carlo de 100.000 iteraciones.

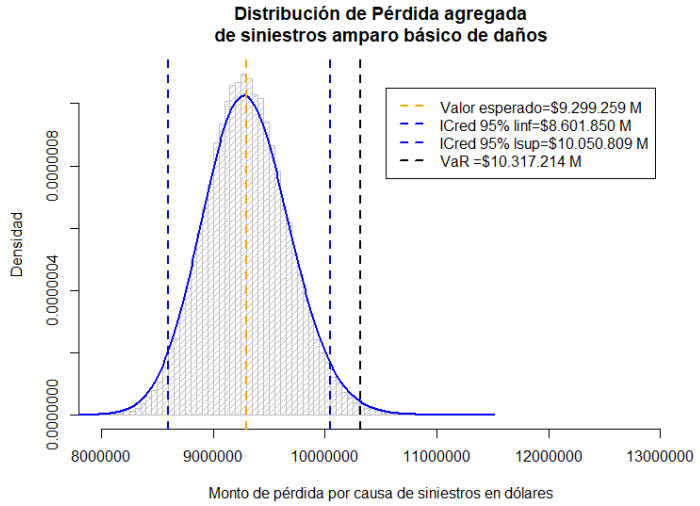

Figura 53: Distribución predictiva de riesgo para pérdida agregada de siniestros bajo enfoque Bayesiano.

Los resultados muestran que el valor esperado de pérdida agregada bajo el enfoque clásico es de $9'287.426$US; mientras que el valor esperado bajo el enfoque Bayesiano es de $9'299.259$US. En este caso el modelo Bayesiano presenta una 'mejor' estimación de pérdida agregada real de $9'314.604$US porque tiene la menor desviación porcentual con respecto al valor observado.

Al establecer un paralelo comparativo entre los modelos clásicos no solamente es importante la mejor estimación del monto de siniestros por medio del valor esperado, la cual será importante dentro de los requerimientos normativos vigentes, la mejor estimación que permita calcular las reservas de siniestros ocurridos avisados y no avisados. De esta manera, la bondad de ajuste cumple un papel importante desde el punto de vista estadístico porque si bien se obtienen valores esperados de pérdida agregada similares, en términos de las cuantías individuales de los siniestros, los modelos Bayesianos permitieron la obtención de valores menores para las métricas de MAPE y RMSE al momento de realizar la validación cruzada de $k-$ fold. El resultado en mención se puede explicar bajo el hecho de que los modelos Bayesianos con número de componentes fijo o aleatorio bajo sus tres propuestas de distribuciones previas no informativas mostraron un 'buen' ajuste al momento de calcular los valores p predictivos para media, desviación estándar, varianza y coeficiente de variación; mientras que bajo el enfoque clásico al hacer la segmentación de riesgos mediante clases con el algoritmo EM, estas, en particular las clases 1 y 2, no se ajustaron a una distribución lognormal al observar los resultados reportados en las pruebas de hipótesis de Anderson-Darling, Crammer Von Misses y kolmogorov-Smirnov.

6. Riesgo de suscripción en seguros de no vida.

El riesgo de suscripción en Colombia desde el Marco Integral de Supervisión de la Superintendencia Financiera se entiende como la probabilidad de que las primas 'cobradas' no sean suficientes para responder ante las obligaciones causadas por siniestros y gastos asociados. De esta manera, un modelo de riesgo a tiempo continuo que permita estimar la probabilidad de ruina en un horizonte de doce meses permite medir el riesgo de que se presente un escenario de insuficiencia de primas en la compañía aseguradora. Por otra parte, la normativa Europea de Solvencia II también se constituye en una forma posible de medir el riesgo de suscripción. Actualmente en el año 2021, esta Directiva se encuentra en estudio de aprobación para Colombia en acuerdo con Fasecolda y las diferentes compañías de seguros que se encuentran participando en los ejercicios de impactos cuantitativos que permitan hacer calibración de una fórmula estándar para el riesgo de suscripción y los requerimientos de capital regulatorios y obligatorios que deben constituir las entidades para su funcionamiento como respuesta a eventos catastróficos para que no se generen eventos de insolvencia y cierre de las compañías.

Es importante resaltar que el conjunto de datos de pólizas y siniestros de Automóviles de Estados Unidos en 2004 no se corresponde con un escenario de industria aseguradora como el de Colombia o a las disposiciones y considerandos expuestos por Marco Integral de la Superintendencia Financiera. En consecuencia, el uso de estos datos solamente es una aplicación ilustrativa de la metodología que permite medir el riesgo de suscripción en seguros de no vida bajo la implementación de un modelo Bayesiano de mezcla para la severidad de los siniestros.

Las etapas para la medición del riesgo de suscripción son:

- Tarifación del seguro de cobertura de daños por medio de los modelos Bayesianos de frecuencia y severidad.
- Estimación de reservas técnicas de primas y siniestros. En el caso de primas se considera la reserva de riesgos en curso o de prima no devengada y en el caso de siniestros se tienen en cuenta las reservas de siniestros ocurridos avisados y de siniestros ocurridos no avisados.
- Cálculo de estados financieros y resultados técnicos considerando como ingresos las primas 'ganadas' o devengadas y como egresos los pagos de siniestros y gastos administrativos, seguros y de personal.
- Estimación de probabilidad de ruina en un horizonte finito de 12 meses bajo distintos escenarios de capital.
- Cálculo de requerimiento de capital para riesgo de suscripción de volumen de primas y reservas para las compañías aseguradoras bajo documento de especificaciones técnicas de Solvencia II para Colombia en el estudio actual que adelanta Fasecolda con las diferentes entidades del sector.

6.1. Tarifación de seguros según segmentos de Modelo Bayesiano con número de componentes aleatorio.

Teniendo en cuenta el procedimiento expuesto en la sección 2.3, la técnica de tarifación se realiza de forma global para todo el portafolio de asegurados y por segmentos o clases latentes de los riesgos construidas mediante el algoritmo EM previo a la implementación del modelo Bayesiano de mezcla con número de componentes fijo o aleatorio. Una vez obtenidas las tasas turas de riesgo, se evalúa si resulta más adecuado realizar la tarifación de manera global o por segmentos teniendo en cuenta el aspecto de competitividad en el mercado de la industria. La metodología para calcular las tasas puras de riesgo por cada segmento h con $h \in \{1, \ldots, H = 5\}$ considera los mejores modelos Bayesianos de frecuencia y severidad que minimizan la raíz cuadrada del error cuadrático medio. El modelo de frecuencia óptimo para las estimaciones del número de siniestros ocurridos de manera semanal es el modelo Poisson de enlace lineal con parámetros no negativos con distribución previa de hiperparámetro $s^2 = 100$. Por otra

parte, el modelo de severidad óptimo para las estimaciones de las cuantías es el modelo Bayesiano de mezcla con número de componentes fijo bajo la propuesta no informativa 2. El proceso de cálculo de las tasas puras de riesgo para la tarifación de seguros de automóviles en la cobertura de daños de mayor o menor cuantía por cada clase latente h se presenta a continuación:

1. Para la iteración s del algoritmo MCMC del modelo de severidad guardar $\mu_{\mathbb{X},h}^{(s)}$ tal que:

$$\mu_{\mathbb{X},h}^{(s)} \leftarrow \exp\left\{\theta_h^{(s)} + \frac{(\sigma_h^2)^{(s)}}{2}\right\}.$$

2. Para la iteración s del algoritmo MCMC del modelo de frecuencia guardar $\mu_{\mathbb{N},h}^{(s)}$ tal que:

$$\mu_{\mathbb{N},h}^{(s)} \leftarrow w_h^{(s)} \cdot \frac{\sum_{i=1}^{52} M_i}{N},$$

   donde M_i es tal que:

$$M_i \overset{\mathrm{d}}{=} \text{Poisson}\left(N_i \cdot \left(\boldsymbol{\beta}^{(s)}\right)^T \cdot \mathbb{X}_i\right).$$

   Por otra parte, N_i es el número de riesgos asegurados en la semana i, $\boldsymbol{\beta}^{(s)}$ es el vector que contiene los parámetros del algoritmo MCMC para el enlace lineal de coeficientes no negativos del modelo Poisson, $\mathbb{X}$ es la matriz de diseño que contiene el intercepto y el índice de tiempo para la semana i y N es el total de siniestros ocurridos observados dentro de las 52 semanas.

3. Para la iteración s guardar $\mathbb{V}\text{ar}_h(S \mid \theta)^{(s)}$ tal que:

$$\mathbb{V}\text{ar}_h(S \mid \theta)^{(s)} \leftarrow \mu_{\mathbb{N},h}^{(s)} \cdot \left[\exp\left\{2 \cdot \theta_h^{(s)} + 2 \cdot (\sigma_h^2)^{(s)}\right\} - \left(\mu_{\mathbb{X},h}^{(s)}\right)^2\right] + \mu_{\mathbb{N},h}^{(s)} \cdot \left(\mu_{\mathbb{X},h}^{(s)}\right)^2.$$

4. Para la iteración s guardar $\mathbb{E}_h(S \mid \theta)^{(s)}$ tal que:

$$\mathbb{E}_h(S \mid \theta)^{(s)} \leftarrow \mu_{\mathbb{N},h}^{(s)} \cdot \mu_{\mathbb{X},h}^{(s)}.$$

Una vez hecho lo anterior en las S iteraciones del algoritmo MCMC, se estima el factor de credibilidad Z del histórico de datos y $1 - Z$ de la distribución posterior de la siguiente manera para cada clase latente h:

- Guardar $\mu_{\text{PV},h}$ y $\sigma_{\text{HM},h}^2$ tales que:

$$\begin{aligned}
\mu_{\text{PV},h} &\leftarrow \frac{1}{S}\sum_{s=1}^{S} \mathbb{V}\text{ar}_h(S \mid \theta)^{(s)} \\
\sigma_{\text{HM},h}^2 &\leftarrow \frac{1}{S-1}\sum_{s=1}^{S}\left[\mathbb{E}_h(S \mid \theta)^{(s)} - \tfrac{1}{S}\sum_{j=1}^{S}\mathbb{E}_h(S \mid \theta)^{(j)}\right]^2.
\end{aligned}$$

- Calcular el factor de credibilidad Z para los datos históricos:

$$Z_h \leftarrow \frac{N_h}{N_h + k_h},$$

  donde N_h es el total de riesgos asegurados de la clase latente h y $k_h = \frac{\mu_{\text{PV},h}}{\sigma_{\text{HM},h}^2}$.

- Guardar los valores de las medias histórica $\bar{S}_h$ y teórica $\mathbb{E}_h(S)$ para las cuantías de los siniestros:

$$\bar{S}_h \leftarrow \frac{\sum_{i=1}^{N_h} X_{i,h}}{N_h} \; ; \; \mathbb{E}_h(S) \leftarrow \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}_h(S \mid \theta)^{(s)}.$$

- Calcular la medida de riesgo M_h de exposición en tiempo y valor asegurado para la clase latente h:

$$M_h \leftarrow \frac{1}{365} \sum_{i=1}^{N_h} \left(\mathrm{SA}_{i,h} \cdot t_{i,h} \right),$$

  donde $\mathrm{SA}_{i,h}$ es el valor asegurado del riesgo i de la clase latente h y $t_{i,h}$ el tiempo de cobertura de su póliza o contrato de seguros.

- La tasa pura de riesgo P_h para la clase latente h es:

$$P_h \leftarrow 100 \cdot \frac{N_h}{M_h} \left[\cdot Z \cdot \bar{S}_h + (1 - Z) \cdot \mathbb{E}_h(S) \right].$$

- La prima comercial C_h sin impuestos de IVA por cada clase h se calcula teniendo en cuenta, por ejemplo, un valor del 10 % para gastos administrativos (GA), un 8 % para gastos comerciales o de intermediarios (GC) y 15 % para rentabilidad y utilidad (RU):

$$C_h \leftarrow \frac{P_h}{1 - GA - GC - RU}.$$

Dado que la tasa pura de riesgo depende de $\mathbb{E}_h(S)$ y a su vez esta depende de $\mathbb{E}_h(S \mid \theta)^{(s)}$, entonces:

$$P_h^{(s)} = 100 \cdot \frac{N_h}{M_h} \left[\cdot Z \cdot \bar{S}_h + (1 - Z) \cdot \mathbb{E}_h(S \mid \theta)^{(s)} \right].$$

En este caso, cada tasa pura por clase latente tiene una distribución posterior:

En las visualizaciones observadas anteriormente se aprecian las tasas puras de riesgo que se deben cobrar en cada segmento para que las primas sean suficientes exclusivamente para el pago de los siniestros. Adicionalmente, se observan los intervalos de credibilidad del 95 % para cada tasa y estos indican los valores mínimo y máximo que se pueden cobrar para que las primas recaudadas permitan hacer frente a los pagos de los siniestros sin lugar a posibles pérdidas. Para complementar los resultados anteriores, a continuación, se presentan los cálculos de cada factor de credibilidad, las medias muestrales y posteriores así como las medidas de riesgos y el valor de las tasas puras por cada clase latente:

Tabla 40: *La tabla presenta en cada una de sus columnas de izquierda a derecha las siguientes salidas: Clase latente, factor de credibilidad de la media muestral, media muestral, factor de credibilidad de la media posterior, media posterior, medida de riesgo y tasa pura de riesgo.*

Clase	Factor Z	Media muestral	Factor $1 - Z$	Media posterior	Medida de riesgo	Tasa pura
1	0,557	164,000	0,443	179,131	7039537	2,061 %
2	0,669	366,571	0,331	421,542	7514784	2,954 %
3	0,985	142,867	0,015	114,185	46914140	0,972 %
4	0,849	122,268	0,151	269,767	226470979	1,697 %
5	0,989	143,439	0,011	51,043	277618268	1,879 %

Realizando el cobro de una tasa pura como la anterior en cada segmento se aprecian los siguientes valores esperados y observados en el monto total de las cuantías de los siniestros:

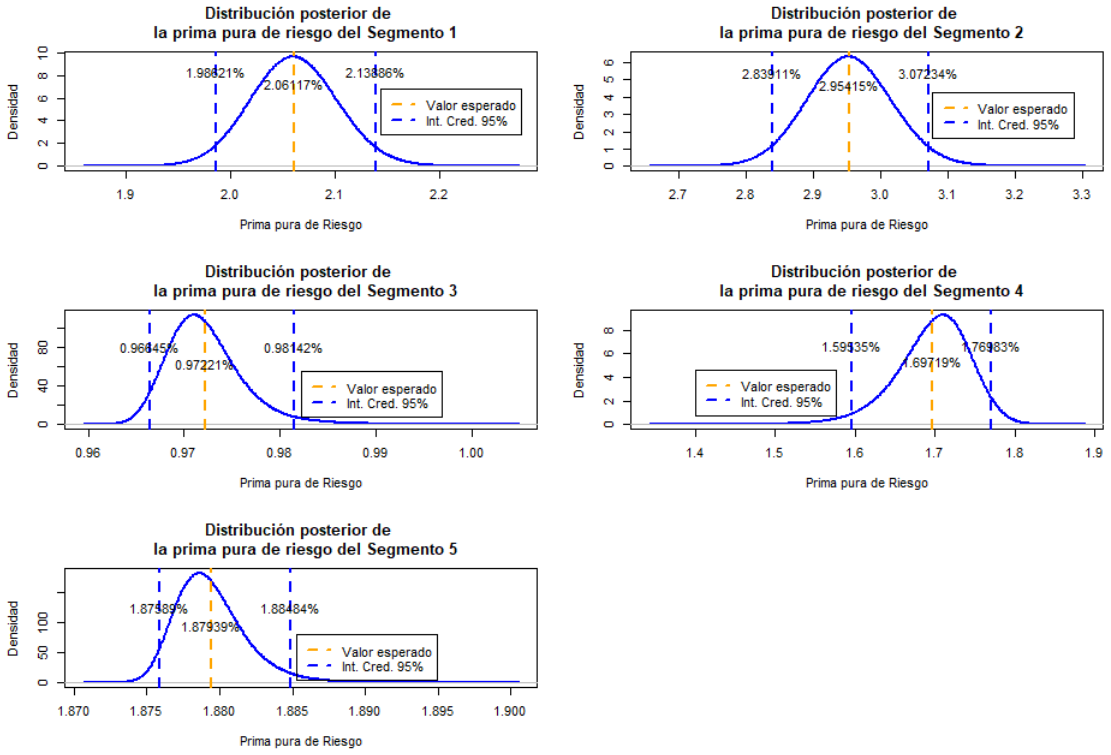

Figura 54: Distribuciones posteriores de las tasas puras de riesgo por clase latente.

Tabla 41: *La tabla presenta en sus columnas de izquierda a derecha las siguientes salidas: Clase latente, Monto esperado de siniestros al tarifar el seguro, monto real de los siniestros y desviación porcentual de monto esperado con respecto al monto observado.*

Segmento	Monto esperado de siniestros	Monto real de siniestros	Desviación
1	145097, 10 US	139400, 00 US	4, 087 %
2	221998, 00 US	211511, 20 US	4, 958 %
3	456104, 30 US	457461, 20 US	−0, 297 %
4	3843632, 60 US	3251338, 00 US	18, 217 %
5	5217536, 10 US	5254894, 10 US	−0, 711 %

Los resultados anteriores demuestran que la diferencia más importante que refleja una desviación entre el monto esperado con la tarifación y el monto observado ocurre en el segmento 4. La ventaja de la estadística Bayesiana reportando un intervalo de credibilidad del 95 % para la tasa pura de riesgo es que se puede ajustar esta desviación observada en la clase latente, considerando el límite inferior como el valor mínimo que se puede cobrar para responder ante las obligaciones causadas por los siniestros. Si es escoge como tasa pura el valor de 1, 59535 %, el monto esperado de los siniestros es de 3613004, 8US y en ese caso la desviación con respecto al monto total observado es de 11, 12 %.

Por otra parte, los valores de las tasas comerciales para cada segmento utilizando un porcentaje de gastos administrativos y de personal del 10 %, un valor del 8 % para gastos comerciales, de publicidad e intermediación y un 15 % de rentabilidad o utilidad, se aprecian a continuación:

Tabla 42: *La tabla presenta a continuación en sus columnas de izquierda a derecha: La clase latente, tasa comercial para la venta del contrato de seguros de autos para cobertura de daños de mayor y/o menor cuantía y tasa pura de riesgo.*

Segmento	Tasa comercial	Tasa pura
1	$3,076378\,\%$	$2,0611734\,\%$
2	$4,409179\,\%$	$2,9541497\,\%$
3	$1,451061\,\%$	$0,9722107\,\%$
4	$2,533113\,\%$	$1,6971855\,\%$
5	$2,805063\,\%$	$1,8793922\,\%$

Los resultados mostrados anteriormente deben ser contrastados con las tasas puras de riesgo y comercial globales para todo el portafolio dado que es necesario verificar si la mejor decisión es lanzar al mercado un producto para automóviles realizando una tasación por segmentos o para todo el portafolio dado que es necesario garantizar la competitivas del producto con respecto a otras compañías. La metodología para la tasación global presenta algunos ajustes sobre $\mu_{\mathbb{X}}$ y $\mu_{\mathbb{N}}$ para cada iteración s del algoritmo MCMC:

$$\mu_{\mathbb{X}}^{(s)} = \sum_{h=1}^{H} \left(w_h \cdot \exp\left\{ \theta_h^{(s)} + \frac{(\sigma_h^2)^{(s)}}{2} \right\} \right) \; ; \; \mu_{\mathbb{N}}^{(s)} = \frac{\sum_{i=1}^{52} M_i}{N},$$

donde M_i es tal que:

$$M_i \overset{\mathrm{d}}{=} \text{Poisson}\left(N_i \cdot \left(\boldsymbol{\beta}^{(s)} \right)^T \cdot \mathbb{X}_i \right).$$

Observe que N_i es el número de riesgos asegurados en la semana i y N es el total de asegurados del portafolio. Por otra parte los valores de $\mathbb{V}\text{ar}(S \mid \theta)^{(s)}$ y $\mathbb{E}(S \mid \theta)^{(s)}$ son:

$$\begin{aligned} \mathbb{V}\text{ar}(S \mid \theta)^{(s)} &= \left[\mathbb{E}\left((\mathbb{X}^2)^{(s)} \right) - \left(\mu_{\mathbb{X}}^{(s)} \right)^2 \right] \cdot \mu_{\mathbb{N}}^{(s)} + \mu_{\mathbb{N}}^{(s)} \cdot \left(\mu_{\mathbb{X}}^{(s)} \right)^2 \\ \mathbb{E}(S \mid \theta)^{(s)} &= \mu_{\mathbb{X}}^{(s)} \cdot \mu_{\mathbb{N}}^{(s)}. \end{aligned}$$

El valor del factor de credibilidad Z de la media muestral es Z tal que:

$$Z = \frac{N}{N + k},$$

donde N es el total de riesgos asegurados en todo el portafolio, $k = \frac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}}$, $\mu_{\text{PV}} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{V}\text{ar}(S \mid \theta)^{(s)}$ y $\sigma^2_{\text{HM}} = \frac{1}{S-1} \sum_{s=1}^{S} \left[\mathbb{E}(S \mid \theta)^{(s)} - \frac{1}{S} \sum_{j=1}^{S} \mathbb{E}(S \mid \theta)^{(j)} \right]^2$. Adicionalmente los valores para la medida de riesgo y exposición, la media muestral y la media teórica se obtienen de forma análoga al método de tarifación de la tasa pura de riesgo en cada clase latente h.

Tal y como se apreció en cada clase latente, de forma global para todo el portafolio de asegurados la tasa pura de riesgo tiene una distribución posterior tal y como se evidencia en la siguiente visualización. De este modo, el valor esperado de tasa a cobrar para cada asegurado es del $1,62207\,\%$ con un intervalo de credibilidad del $95\,\%$ cuyos límites indican que el valor mínimo a tarifar es del $1,60948\,\%$ y el valor máximo es de $1,63484\,\%$.

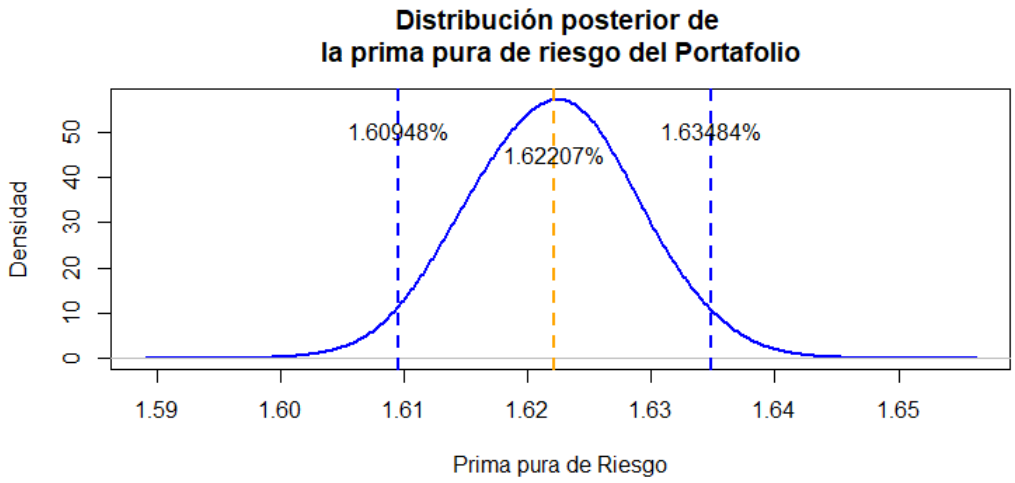

Figura 55: Distribución posterior de la tasa pura de riesgo para todo el portafolio de asegurados.

Complementando los resultados presentados anteriormente, se relacionan a continuación el factor de credibilidad, las medias muestrales y teórica, la medida de riesgo y las tasas puras y comercial utilizando un porcentaje de gastos administrativos y de personal del 10 %, gastos comerciales y de intermediarios del 8 % y rentabilidad del 15 %:

Tabla 43: *La tabla en su primera columna reporta los insumos para el cálculo de la tasa pura de riesgo mientras que la segunda columna indica el valor asociado a cada insumo.*

Indicador	Valor
Factor de credibilidad	0,96702
Media muestral	137,27020
Media posterior	74,32941
Medida de riesgo	565557708
Tasa pura	$1,6221\,\%$
Tasa comercial	$2,4210\,\%$

La interpretación proporcionada por la tabla anterior permite concluir que para que las primas permitan cubrir únicamente los montos de los siniestros es necesario realizar el cobro de una tasa del 1.62 % por valor asegurado del vehículo. Bajo este escenario el valor esperado de pago de los siniestros es de 9′299.259 US; mientras que el valor observado de pago es de 9′314.604 US. De este modo, la desviación entre los resultados es del $-0,1647\,\%$, lo cual indica que se tiene una 'buena' tasa pura de riesgo para tarifar el seguro. Por otra parte, la tasa comercial para todo el portafolio toma un valor del $2,4210\,\%$ y resulta ser más económica para todos los segmentos excepto el 3, donde el valor de la tasación fue de $1,451061\,\%$. En ese sentido, la tasa comercial puede ser bastante competitiva para todos los segmentos excepto en el 3, donde se puede proponer un nuevo producto con su respectivo clausulado definiendo el tipo de vehículos que se aseguran y generando una tarifación comercial con tasa del 1.451061 %. Para fines prácticos, se realiza el estudio del producto con la tasa global comercial del $2,4210\,\%$ para todo el portafolio.

6.2. Simulación base de datos de emisión y pagos de siniestros en seguros de Automóviles.

Como resultado posterior a la escogencia de las tasas pura y comercial para la tarifación del portafolio de asegurados, se hace una simulación de un conjunto de datos que integra los datos fijos de vehículos asegurados en Estados Unidos en 2004 y los datos aleatorios de fecha de emisión del contrato o póliza, así como las fechas de cobertura del vehículo asegurado ante daños de menor o mayor cuantía que puedan generar la siniestralidad. Los datos fijos son: valor asegurado, edad del vehículo, años de uso del vehículo, género del conductor del vehículo asegurado, área de tránsito del vehículo, tipo de vehículo, segmento o clase latente del riesgo, valor de la prima comercial y gastos administrativos o comerciales asociados a la expedición de la póliza. La simulación se realiza teniendo en cuenta que un vehículo puede estar asegurado entre uno y doce meses bajo la siguiente función de masa de probabilidad $p(x)$:

$$p(X = x) = \begin{cases} 0.85 & , \text{si } x = 12 \\ 0.013\bar{6} & , \text{si } x \in \{1, \ldots, 11\} \end{cases}$$

En consecuencia, la probabilidad es máxima cuando el vehículo será asegurado por un año con un valor del 85 %. El valor de la prima comercial PC_i sin impuesto IVA en dólares para el riesgo i de valor asegurado SA_i se calcula de la siguiente manera:

$$\text{PC} = \frac{\text{SA}_i \cdot C \cdot (\text{Fec Vig Hasta} - \text{Fec Vig Desde})}{365},$$

donde C es la tasa comercial tal que $C = 2.4210\,\%$, Fec Vig Hasta es la fecha final de vigencia de la póliza y Fec Vig Desde es la fecha inicial de vigencia de la póliza. Por otra parte, los gastos administrativos corresponden al 10 % y los gastos comerciales al 8 % de la prima del contrato de seguros. De esta manera, la base simulada tiene una estructura similar a la que presenta en seguida:

Tabla 44: *Las columnas de la tabla presentan la siguiente información al hacer lectura de la misma en sentido de izquierda a derecha: Llave que identifica la póliza, endoso o anexo, fecha de emisión, fecha de inicio de vigencia de la póliza, fecha de fin de vigencia de la póliza, valor de la prima comercial, valor de gastos administrativos y valor de gastos comerciales.*

Llave	Endoso	Fec Emi	Fec Vig Desde	Fec Vig Hasta	Prima	GA	GC
01-03-1	Nuevo cont	4/01/2021	4/01/2021	4/12/2021	646,890	64,689	51,751
01-03-2	Nuevo cont	4/01/2021	4/01/2021	4/01/2022	1267,634	126,763	101,411
01-03-3	Nuevo cont	4/01/2021	4/01/2021	4/01/2022	227,574	22,757	18,206
01-03-4	Nuevo cont	4/01/2021	4/01/2021	4/10/2021	626,527	62,653	50,122
01-03-5	Nuevo cont	4/01/2021	4/01/2021	4/09/2021	183,744	18,374	14,699
01-03-6	Nuevo cont	4/01/2021	4/01/2021	4/01/2022	796,508	79,651	63,721
01-03-7	Nuevo cont	4/01/2021	4/01/2021	4/01/2022	130,734	13,073	10,459

Típicamente la llave de la póliza se utiliza como una variable identificadora para el contrato de seguros, mientras que el endoso o anexo en este caso hace alusión a nuevos contratos o emisiones originales de vez primera de las pólizas de automóviles con amparo básico de daños de mayor y/o menos cuantía. En la práctica, el endoso suele contener anexos sobre la emisión original donde se contratan o se cancelan nuevas coberturas como por ejemplo la sustracción del vehículo o protección ante desastres naturales. De esta manera los endosos suelen afectar los valores asegurados y los valores de la prima. Así mismo, los endosos indican si una póliza se renueva o no una vez termina su vigencia o si el contrato es revocado por

el asegurado o el asegurador. Para la simulación, note que las fechas de inicio de vigencia coinciden con las fechas de emisión o expedición de la póliza. Sin embargo, en el negocio de la industria aseguradora esto no necesariamente ocurre porque se puede hacer una emisión de un contrato de seguros en forma anticipada al inicio de vigencia de este.

Adicionalmente a lo anterior, es necesario realizar una simulación de un conjunto de datos que contenga la gestión de pagos de la compañía aseguradora. Dicha gestión es de vital importancia porque es necesario tener en cuenta los siguientes considerandos dentro el negocio asegurador:

- Una vez ocurrido y avisado el siniestro, no necesariamente el valor de su cuantía es determinado instantáneamente porque es necesario tomar la reclamación del asegurado para hacer una investigación exhaustiva que permita decidir en primera instancia si se realiza o no el pago. Por otra parte, en caso de pago es necesario considerar los hechos que causaron el siniestro para así calcular la cuantía y este proceso puede tomar un tiempo.
- Los pagos se realizan en diferentes periodos de manera que con el último pago se liquida el valor total de la cuantía del siniestro.
- Los asegurados suelen hacer la reclamación de su siniestro una vez ocurre. Sin embargo, otra proporción de asegurados pueden realizar la reclamación un tiempo después. Por lo anterior, la compañía aseguradora debe constituir reservas de siniestros ocurridos avisados y no avisados puesto que los pagos muestran solamente los resultados de la gestión, pero no necesariamente contienen toda la información del pago de los siniestros.

De acuerdo con los aspectos mencionados anteriormente, se realiza una simulación de la gestión de pagos considerando que estos suceden bajo una progresión geométrica de razón $\frac{1}{2}$ en el tiempo. Adicionalmente la probabilidad de que el asegurado de aviso a su siniestro dentro del mismo año en curso es del 80 %. En consecuencia, un conjunto de datos que simule la gestión de pagos tiene la siguiente estructura en la práctica:

Tabla 45: *La información contenida en las columnas de la tabla cuando se realiza su lectura de izquierda a derecha reporta: llave de identificación del siniestro ocurrido, llave de la póliza, valor del siniestro en dólares, fecha de ocurrencia del siniestro, fecha de aviso del siniestro, función indicadora de aviso del siniestro dentro del año en curso, fecha de pago y valor del pago.*

Llave stro	Llave	Valor stro	Fec Siniestro	Fec Aviso	Ind. Aviso	Fec Pago	Pago
01-03-1397-17	01-03-1397	3660,09	5/01/2021	5/01/2021	1	5/02/2021	1830,05
01-03-1398-18	01-03-1398	25594,22	5/01/2021	5/01/2021	1	5/02/2021	12797,11
01-03-407-4	01-03-407	200,00	6/01/2021	6/01/2021	1	6/02/2021	100,00
01-03-409-6	01-03-409	924,94	6/01/2021	6/01/2021	1	6/02/2021	462,47
01-03-621-7	01-03-621	5832,04	6/01/2021	6/01/2021	1	6/02/2021	2916,02
01-03-1023-11	01-03-1023	9053,00	6/01/2021	6/01/2021	1	6/02/2021	4526,50

Observe que la llave del siniestro identifica el siniestro y en consecuencia, una póliza puede tener diferentes llaves de siniestro dado que puede ocurrir al menos un siniestro a un vehículo asegurado. Por otra parte, columnas como el valor del siniestro o la función que indica si el siniestro ha sido avisado o no dentro del año en curso no son observables de manera instantánea en las bases de datos de las compañías aseguradoras y se conocen tiempo después de comenzar a realizar la gestión de pagos. Como en esta aplicación práctica se conocen los valores de los siniestros, es posible validar que el valor acumulado de los siniestros liquidados o pagos dentro del primer año no se corresponden con las cuantías totales de los siniestros tal y como se aprecia en la siguiente tabla:

Tabla 46: *Las columnas de la tabla en orden de izquierda a derecha contienen la siguiente información: corte de fin de mes para pagos acumulados de los siniestros del año en curso, valor acumulado de los pagos en dólares hasta la fecha de corte, número de siniestros ocurridos acumulados a la fecha de corte y valor real de los siniestros acumulado a la fecha de corte.*

Fecha de corte	Monto de pagos	número de siniestros	Valor real siniestros
31/01/2021	0	107	514664,5
28/02/2021	189834,3	259	948029,6
31/03/2021	490071,1	476	1527217,4
30/04/2021	847631,1	761	2290179,4
31/05/2021	1363717,6	1124	3092505,2
30/06/2021	1917207,3	1496	3822371,2
31/07/2021	2462948,6	1930	4720747,6
31/08/2021	3104262,2	2428	5682758,3
30/09/2021	3799498,1	2917	6543378,6
31/10/2021	4501543,3	3418	7412393,4
30/11/2021	5164991,2	3901	8150717,6
31/12/2021	5851297	4624	9314604,4

Note que el valor pagado por obligaciones causadas por siniestros de mayor y/o menor cuantía en automóviles durante el primer año del producto corresponde a 5′851.297US; mientras que el valor real de los pagos corresponde a $9'314.604,4$US. En consecuencia, el valor de la reserva de siniestros R que debe constituir la compañía corresponde a 31 de diciembre de 2021 es:

$$R = 9'314.604,4 \text{ US} - 5'851.297 \text{ US} = 3'463.307,4 \text{ US}.$$

El valor de esta reserva de siniestros contempla tanto el valor de reserva de siniestros avisados como no avisados dentro del año en curso. Sin embargo, las normativas vigentes en Colombia contemplan este cálculo desagregado tanto a siniestros ocurridos avisados antes de la fecha de cálculo de la reserva como de siniestros ocurridos no avisados a la fecha de cálculo de la reserva.

El documento de Régimen de Seguros expedido por Fasecolda define en su capítulo VI, conceptos como reserva de siniestros ocurridos avisados y no avisados a la fecha de cálculo bajo el Decreto Único Financiero 2555 de 2010 modificado por el artículo 1 del Decreto 2973 de 2013. La reserva de siniestros ocurridos avisados y de ahora en adelante RSAP es aquella que 'corresponde al monto de recursos que debe destinar la entidad aseguradora para atender los pagos de los siniestros ocurridos una vez estos hayan sido avisados, así como los gastos asociados a éstos, a la fecha de cálculo de esta reserva'. Por otra parte, la reserva de siniestros ocurridos y no avisados a la fecha de corte se denomina IBNR y 'representa una estimación del monto de recursos que debe destinar la entidad aseguradora para atender los futuros pagos de siniestros que ya han ocurrido, a la fecha de cálculo de esta reserva, pero que todavía no han sido avisados a la entidad aseguradora'.

Dado el control que se tiene sobre la simulación, es posible conocer tal y como se dijo anteriormente características no observables en los conjuntos de datos que permiten estudiar la gestión de pagos de la compañía aseguradora en su primer año. Estas características se relacionan con el aviso de los siniestros dentro del año en curso, así como el monto total a pagar por el siniestro. De esta manera los valores de reserva de siniestros a constituir en cada corte de forma acumulada en el primer año se presentan a continuación:

Tabla 47: *La información de las columnas de izquierda a derecha reporta: fecha de corte, valor total de la reserva, número de siniestros avisados en el año en curso, valor de la reserva de siniestros avisados, número de siniestros no avisados y valor de la reserva de siniestros no avisados.*

Fecha de corte	Reserva total	siniestros avisados	RSAP	siniestros no avisados	IBNR
31/01/2021	514664,5	82	398671,2	25	115993,3
28/02/2021	758195,3	198	590972,3	61	167223
31/03/2021	1037146,2	372	723920,5	104	313225,7
30/04/2021	1442548,3	612	1027772,6	149	414775,7
31/05/2021	1728787,6	899	1109179,6	225	619608
30/06/2021	1905163,9	1192	1090382,6	304	814781,3
31/07/2021	2257799	1544	1282627,2	386	975171,8
31/08/2021	2578496,1	1949	1427491,4	479	1151004,7
30/09/2021	2743880,5	2331	1385580,6	586	1358299,9
31/10/2021	2910850,1	2723	1379603,4	695	1531246,7
30/11/2021	2985726,5	3113	1346257,9	788	1639468,5
31/12/2021	3463307,4	3689	1584175	935	1879132,4

Observe que el valor real de la reserva que debe constituir la compañía aseguradora para siniestros ocurridos avisados a 31 de diciembre de 2021 es de $1'584.175$ US; mientras que el valor de la reserva IBNR que debe constituir a la misma fecha es de $1'879.132$US. Respondiendo a las normativas vigentes, las compañías aseguradoras deben realizar la mejor estimación de estas reservas dado que existe incertidumbre con respecto a: el monto total que tienen los siniestros ocurridos avisados con respecto a los pagos realizados y el monto total que tienen los siniestros ocurridos no avisados. Aunque en la práctica se utilizan metodologías de triangulación determinísticas y estocásticas para el cálculo de las reservas, el mejor modelo Bayesiano de frecuencia y severidad que minimiza el error cuadrático medio para predecir el número de siniestros ocurridos y el monto de sus cuantías hasta el tiempo t puede realizar la estimación de los valores de las reservas de siniestros mediante la siguiente metodología. La mejor estimación de la reserva de siniestros avisados a una fecha de cálculo se obtiene de la siguiente manera:

- Sea N_a el número de siniestros ocurridos avisados dentro de un tiempo t. Para cada iteración s del algoritmo MCMC se estima la cuantía de los siniestros $S_a^{(s)}$ de la siguiente manera:

$$S_a^{(s)} \leftarrow \sum_{i=1}^{N_a} X_i,$$

donde X_i es tal que:

$$\begin{aligned} X_i \mid Z_{i,h} &\overset{\mathrm{d}}{=} \text{lognormal}\left(\theta_h^{(s)}, (\sigma_h^2)^{(s)}\right) \\ Z_{i,h} &\overset{\mathrm{d}}{=} \text{Multinomial}\left(w_1^{(s)}, \ldots, w_H^{(s)}\right). \end{aligned}$$

- Guardar el valor de la reserva de siniestros ocurridos avisados $\text{RSAP}^{(s)}$ en la iteración s tal que:

$$\text{RSAP}^{(s)} \leftarrow S_a^{(s)} - P_t,$$

donde P_t es el monto total de pagos de siniestros ocurridos avisados hasta el tiempo t.

- Calcular la mejor estimación de las reservas RSAP como la media posterior de $\text{RSAP}^{(s)}$:

$$\text{RSAP} \leftarrow \frac{1}{S}\sum_{s=1}^{S}\text{RSAP}^{(s)}.$$

La metodología de estimación para el cálculo de la reserva IBNR se realiza de la siguiente manera:

- Para cada iteración del algoritmo MCMC se estima el total de siniestros $N_t^{(s)}$ que se espera ocurran hasta el tiempo t en semanas de manera que:

$$N_t^{(s)} \leftarrow \sum_{i=1}^{t} M_i,$$

  donde M_i es tal que:

$$M_i \overset{\text{d}}{=} \text{Poisson}\left(N_i \cdot \left(\boldsymbol{\beta}^{(s)}\right)^T \cdot \mathbb{X}_i\right).$$

  Observe que la matriz de diseño $\mathbb{X}$ es de tamaño $t \times 2$, y N_i es el número de riesgos asegurados en la semana i.

- Guardar el valor esperado del monto total de los siniestros $S^{(s)}$ tal que:

$$S^{(s)} \leftarrow \sum_{i=1}^{N_t^{(s)}} X_i,$$

  donde X_i es tal que:

$$\begin{aligned} X_i \mid Z_{i,h} &\overset{\text{d}}{=} \text{lognormal}\left(\theta_h^{(s)}, (\sigma_h^2)^{(s)}\right) \\ Z_{i,h} &\overset{\text{d}}{=} \text{Multinomial}\left(w_1^{(s)}, \ldots, w_H^{(s)}\right). \end{aligned}$$

- Obtener el valor de la reserva IBNR en la iteración s tal que:

$$\text{IBNR}^{(s)} \leftarrow S^{(s)} - \text{RSAP}^{(s)}.$$

- La mejor estimación de la reserva IBNR a corte del periodo t es:

$$\text{IBNR} \leftarrow \frac{1}{S}\sum_{s=1}^{S}\text{IBNR}^{(s)}.$$

Observe que para esta reserva no se descuentan pagos realizados por la compañía aseguradora dado que aún no se han avisado estos siniestros a la fecha de cálculo de la reserva. Por otra parte, la mejor estimación del número de siniestros ocurridos no avisados $N_{\text{IBNR}}^{(s)}$ viene dada en cada iteración del algoritmo MCMC por:

$$N_{\text{IBNR}}^{(s)} \leftarrow N_t^{(s)} - N_a.$$

De este modo en términomo computacionales, la mejor estimación del número de siniestros ocurridos no avisados N_{IBNR} a la fecha de cálculo de la reserva está dada por:

$$N_{\text{IBNR}} = \mathbb{E}(N_t - N_a) = \frac{1}{S}\sum_{s=1}^{S}\left(N_t^{(s)} - N_a\right).$$

En consecuencia, de lo anterior, se obtienen las mejores estimaciones de las reservas IBNR, RSAP y de siniestros mediante las metodologías anteriores a corte de 31 de diciembre de 2021:

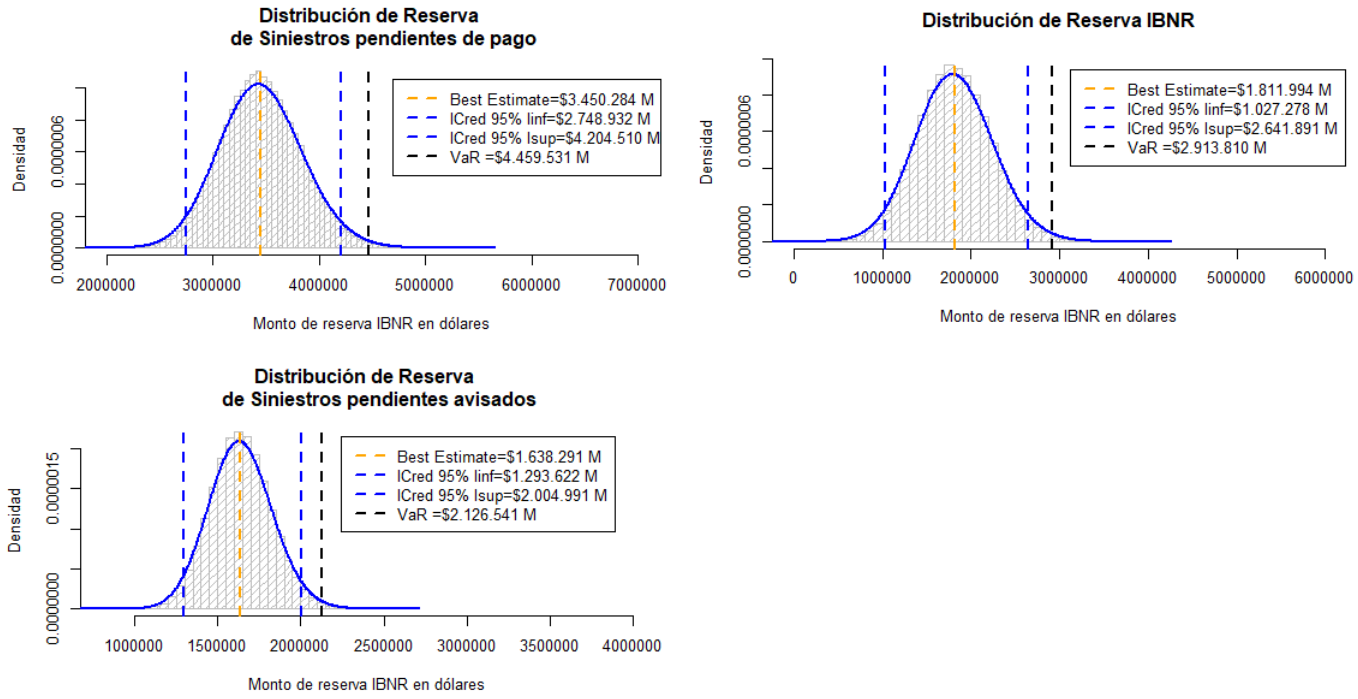

Figura 56: Distribuciones predictivas para reserva de siniestros pendientes de pago, reserva de siniestros IBNR y reserva de siniestros ocurridos avisados.

Los resultados obtenidos anteriormente para la mejor estimación de las reservas de siniestros se pueden resumir de la siguiente manera en comparación con los valores reales observados en la tabla 45:

Tabla 48: *La tabla presenta la siguiente información en el orden de sus columnas de izquierda a derecha: Tipo de reserva de siniestros, su mejor estimación mediante el modelo Bayesiano de frecuencia y severidad, el valor real y la desviación porcentual de la mejor estimación respecto al valor real que debería constituir la compañía aseguradora en reservas de siniestros.*

Reserva	Mejor estimación	Valor real	Desviación
Siniestros pendientes de pago	3450284,0	3463307,4	-0,376 %
RSAP	1638291,0	1584175,0	3,416 %
IBNR	1811994,0	1879132,4	-3,573 %

La interpretación sugiere que las mejores estimaciones de las reservas técnicas de siniestros con respecto a su valor real se desvían en menos de un 4 %, lo cual puede ser un 'buen' indicador al momento de constituir las reservas que se encuentran en las normativas vigentes para el sector asegurador colombiano.

Otra de las reservas importantes a considerar para el análisis de riesgo de suscripción es la reserva de riesgos en curso o reserva de prima no devengada. La interpretación de esta reserva indica que es el monto de prima emitida que aún no ha ganado la aseguradora porque los riesgos asegurados aún no han finalizado su vigencia. De esta manera, cuando una compañía aseguradora hace emisión de una póliza, el monto de prima comercial sin IVA no es 'ganado' por la compañía hasta que el riesgo termine de estar asegurado en términos de su fecha de fin de vigencia. En consecuencia, la compañía aseguradora a una fecha de corte o cálculo tiene unas primas ganadas o devengadas de los días de cobertura que ya han

transcurrido para los riesgos asegurados, mientras que tiene una prima no 'ganada' y se constituye en su reserva de riesgos en curso de los días que aún a esta fecha hacen falta para que los riesgos finalicen su contrato de seguros. Al respecto, Fasecolda mediante el documento de régimen de seguros que compila las normativas vigentes para el sector asegurador colombiano, en su capítulo VI citando el Decreto Único Financiero 2555 de 2010 modificado por el artículo 1 del Decreto 2973 de 2013 considera que 'es aquella que se constituye para el cumplimiento de las obligaciones futuras derivadas de los compromisos asumidos en las pólizas vigentes a la fecha de cálculo. La reserva de riesgos en curso está compuesta por la reserva de prima no devengada y la reserva por insuficiencia de primas'. Para la implementación de la presente metodología no será considerada la reserva de insuficiencia de primas.

La reserva de prima no devengada RPND se calcula de la siguiente forma:

$$\text{RPND} = \text{PC} \cdot \text{factor}_{\text{ND}},$$

donde PC es la prima comercial sin impuesto de IVA que se cobra al asegurado para celebrar su contrato de seguros y $\text{factor}_{\text{ND}}$ es el factor de prima que aún no ha ganado la aseguradora a causa de que el riesgo asegurado continua en vigor a la fecha de cálculo o valoración. El factor de prima no devengada para las fechas de: fin de vigencia Fec Vig Hasta, inicio de vigencia Fec Vig Desde y valoración o cálculo de la reserva de riesgos en curso Fec val para un contrato de seguros se define tal que:

$$\text{factor}_{\text{ND}} = \begin{cases} 1 & \text{, si Fec val} < \text{Fec Vig Desde} \\ 0 & \text{, si Fec val} > \text{Fec Vig Hasta} \\ \frac{\text{Fec Vig Hasta} - \text{Fec val}}{\text{Fec Vig Hasta} - \text{Fec Vig Desde}} & \text{, en otro caso.} \end{cases}$$

Como ilustración de esta metodología de cálculo, se presentan a continuación los resultados de primas emitidas, primas devengadas y reserva de primas no devengadas junto con gastos administrativos y comerciales causados por las pólizas de seguros de automóviles suscritas en 2021 dentro del conjunto de datos simulado:

Tabla 49: *La tabla presenta información de sus columnas en orden de izquierda a derecha de la siguiente manera: Fecha de corte, primas emitidas acumuladas a la fecha de corte, primas devengadas acumuladas a corte, reserva de riesgos en curso a fecha de corte, gastos adminsitrativos y de personal a fecha de corte y gastos comerciales o de intermediación a fecha de corte.*

Fecha corte	Primas Emitidas	Primas Devengadas	RPND	GA	GC
31/01/2021	2460991	123100	2337891	246099,1	196879,2
28/02/2021	4772427	442377,8	4330049	477242,7	381794,2
31/03/2021	7629272	1011245	6618027	762927,2	610341,7
30/04/2021	10051370	1784385,5	8266985	1005137	804109,6
31/05/2021	12895540	2799921,6	10095618	1289554	1031643,2
30/06/2021	14872291	3979405,1	10892886	1487229,1	1189783,3
31/07/2021	16876027	5382567,9	11493459	1687602,7	1350082,1
31/08/2021	19402537	6968726,2	12433811	1940253,7	1552203
30/09/2021	21211395	8670635,5	12540759	2121139,5	1696911,6
31/10/2021	22921990	10572519,7	12349471	2292199	1833759,2
30/11/2021	24952260	12553235,1	12399025	2495226	1996180,8
31/12/2021	27033143	14733671	12299472	2703314,3	2162651,5

La interpretación sugiere que aunque en el año 2021 se emitieron pólizas por valor de 27′033.143US solamente se ganaron 14′733.671US. Dado que los contratos de seguros de autos para la cobertura de amparo básico de daños de mayor y/o menor cuantía se celebran como máximo a un año, la prima que se espera ganar para el próximo año es de 12′299.472US. Es importante aclarar que las primas emitidas a diciembre de 2021 solamente se 'ganarán' hasta 31 de diciembre de 2022.

En términos del factor de prima no devengada, observe que en caso de que la fecha de valoración sea menor que la fecha de inicio, el factor de prima no devengada es 1 porque no se ha 'ganado' aún la prima emitida dado que a la fecha de cálculo esta se ha facturado de manera anticipada a su inicio de vigencia. Sin embargo, para el caso puntual de la presente metodología no se cuenta con eventos de esta naturaleza dentro del proceso de simulación del conjunto de datos. Por otra parte, el factor de prima no devengada toma el valor de cero cuando toda la prima ya ha sido 'ganada' por la compañía aseguradora dado que, a la fecha de cálculo de la reserva, el riesgo ya no se encuentra vigente.

6.3. Modelo de Riesgo a tiempo continuo y probabilidad de ruina.

Teniendo en cuenta tanto gastos como siniestos en una compañía aseguradora, el modelo de riesgo a tiempo continuo del proceso estocástico $\{C_t\}_{t\geq 0}$ se puede adaptar de la siguiente manera:

$$C_t = \begin{cases} u & \text{si } t = 0 \\ u + P(t) - S(t) - G(t) & \text{si } t > 0, \end{cases}$$

donde u es el capital inicial de la compañía aseguradora, $S(t)$ es el monto total de los siniestros a cargo de la aseguradora hasta el tiempo t y $G(t)$ es el monto de gastos de la compañía. La cuantificación del riesgo de suscripción se hace obteniendo la probabilidad de ruina con horizonte finito $\psi(u, t_0)$. A partir de los anterior, generar una partición con n puntos de $[0, t_0]$ igualmente espaciados $t_{0,1}, t_{0,2}, \ldots, t_{0,n-1}, t_0$ tales que:

$$t_{0,1} < t_{0,2} < ... < t_{0,n-1} < t_0.$$

Realizando S_{MCMC} iteraciones para obtener muestras posteriores de los vectores de parámetros $\boldsymbol{\lambda}^{(s)}$ para el modelo Bayesiano de frecuencia y $\boldsymbol{\theta}^{(s)}$ para el modelo Bayesino de severidad; el procedimiento para la obtención de $\psi(u, n)$ se presenta a continuación:

1. Definir k valores $\{u_1, \ldots, u_K\}$ para el capital inicial de la aseguradora. Si $0 \in \{u_1, \ldots, u_K\}$, el modelo del proceso estocástico $\{C_t\}$ permite realizar la proyección del estado de financieros de la compañía aseguradora hasta el tiempo t_0.
2. Por cada u_k obtener S_{MCMC} muestras de $S(t)^*$, donde $S(t)^*$ es el monto de siniestros liquidados (este monto se desagrega como el pago total a la fecha de cálculo y la reserva de siniestros pendientes de pago constituida por la compañía aseguradora) y $t \in \{t_{0,1}, t_{0,2}, \ldots, t_{0,n-1}, t_0\}$. El proceso se realiza de la siguiente manera:
   - Muestrear un valor para $N(t)^{(s)}$ por cada s con $s \in \{1, \ldots, S_{\text{MCMC}}\}$ de manera que:
   $$N(t)^{(s)} \overset{\text{d}}{=} p\left(N(t) \mid \boldsymbol{\lambda}^{(s)}\right).$$
   - Por cada s con $s \in \{1, \ldots, S_{\text{MCMC}}\}$, obtener $N(t)^{(s)}$ muestras de la cuantía de los siniestros $X_i^{(s)}$ con $i \in \{1, \ldots, N(t)^{(s)}\}$ de manera que:
   $$X_i^{(s)} \overset{\text{iid}}{=} p\left(X_i \mid \boldsymbol{\theta}^{(s)}\right).$$
   - Guardar $S(t)^{*,(s)}$ como:
   $$S(t)^{*,(s)} \leftarrow \textstyle\sum_{i=1}^{N(t)^{(s)}} X_i^{(s)}.$$

3. Obtener las primas devengadas $P(t)$ y los gastos $G(t)$ de la compañía para $t \in \{t_{0,1}, t_{0,2}, \ldots, t_{0,n-1}, t_0\}$, considerando los porcentajes de gastos administrativos y de personal GA aplicados en el proceso de cálculo de la reserva de riesgos en curso y el procedimiento de tarifación de seguros para el cálculo de la prima comercial, respectivamente.

4. Estimar la probabilidad de ruina de la compañía aseguradora $\psi(u_k, t_0)$ en un horizonte finito $t_0 > 0$ para todo $k \in \{1, \ldots, K\}$ de la siguiente manera:

   - Guardar $\{C_t\}^{(s)}$ como:

   $$\{C_t\}^{(s)} \leftarrow u_k + P(t)^{(s)} - S(t)^{(s)} - G(t)^{(s)},$$

   para todo $t \in \{t_{0,1}, t_{0,2}, \ldots, t_{0,n-1}, t_0\}$.
   - Guardar en $\tau^{(s)}$ un tiempo de paro definido como una variable indicadora tal que:

   $$\tau^{(s)} \leftarrow \begin{cases} 1 & \text{si } \{C_t\}^{(s)} \leq 0 \text{ para algún } t \in \{t_{0,1}, t_{0,2}, \ldots, t_{0,n-1}, t_0\} \\ 0 & \text{en otro caso .} \end{cases}$$

   - Guardar $\psi(u_k, t_0)$ como:

   $$\psi(u_k, t_0) \leftarrow \frac{1}{S_{\text{MCMC}}} \sum_{s=1}^{S_{\text{MCMC}}} \tau^{(s)}.$$

5. Realizar una visualización de un gráfico de dispersión para u_k y $\psi(u_k, t_0)$ con $k \in \{1, \ldots, K\}$.

Note que para estimar la probabilidad de ruina de una compañía aseguradora en un horizonte infinito, se deben realizar los anteriores pasos con t_0 lo suficientemente grande para garantizar la convergencia de $\psi(u_k, t_0)$ a $\psi(u_k)$. Al implementar la simulación se obtuvieron los siguientes valores de probabilidad de ruina para diferentes niveles de capital u en dólares:

Tabla 50: *La tabla presenta la siguiente información en sus columnas en orden de izquierda a derecha: Capital inicial u de la compañía aseguradora, probabilidad de ruina en un horizonte finito de 12 meses bajo el nivel de capital constituido para producto de automóviles en amparo básico de daños de mayor y/o menor cuantía, límites inferior y superior del intervalo de credibilidad del* 95 % *para probabilidad de ruina.*

Capital u	Probabilidad de ruina $\psi(u)$	límite inferior 95 %	límite superior 95 %
0 US	0,24	0,00	0,50
172413, 793 US	0,08	0,00	0,25
344827, 586 US	0,03	0,00	0,15
517241, 379 US	0,00	0,00	0,05
689655, 172 US	0,00	0,00	0,00
862068, 966 US	0,00	0,00	0,00

De esta manera, los resultados permiten comprobar que en un escenario de no constitución de capital para la compañía aseguradora al hacer el negocio de suscripción de pólizas de automóviles por un año con amparo básico de daños por mayor y/o menor cuantía, la probabilidad esperada de que, las primas devengadas no sean suficientes para hacer frente a obligaciones de gastos y siniestros es del 24 %. Por otra parte, esta probabilidad está entre un escenario optimista del 0 % y un escenario pesimista del 50 % al determinar su intervalo de credibilidad del 95 %. Sin embargo, las normativas vigentes exigen que para el funcionamiento de un Ramo y una Compañía de Seguros Generales o No Vida se requiere un capital

obligatorio por lo cual, el escenario de comenzar el negocio sin capital es interpretable pero no es un evento posible en la práctica.

Note que, en la tabla presentada anteriormente, a medida que se constituye un mayor volumen de capital para la compañía aseguradora, menor es la probabilidad de entrar en insolvencia y a su vez, menores son los límites del intervalo de credibilidad del 95 % para dicha probabilidad. De este modo, la probabilidad de entrar en insolvencia es nula cuando la aseguradora constituye un capital de 862.068, 966US, con lo cual garantizaría el cumplimiento de sus obligaciones de pagos de siniestros y gastos. Así pues, la aseguradora en un momento hipotético alcanzaría a consumir todo este capital en mención para cubrir sus obligaciones.

Complementando los resultados de la tabla anterior, a continuación, se presenta la visualización de probabilidad de ruina con su respectivo intervalo de credibilidad del 95 % según la variación de capital inicial constituido:

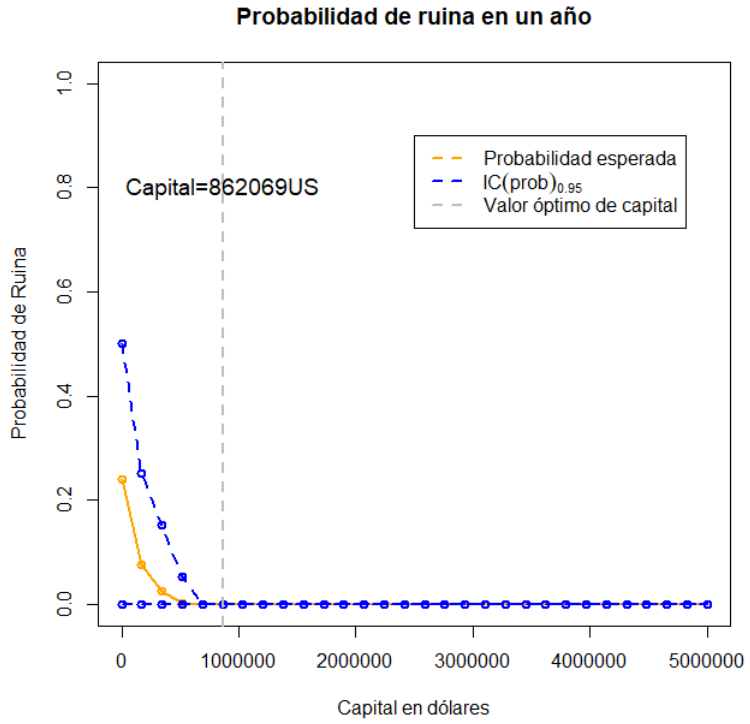

Figura 57: Gráfico de probabilidad de ruina en un horizonte temporal de 12 meses para negocio asegurador de automóviles con amparo básico de cobertura de daños por mayor y/o menor cuantía según diferentes niveles de volumen de capital constituido.

En términos de la distribución de pérdida agregada causada por siniestros, es posible observar mediante las distribuciones predictivas en cada clase latente o segmento, el valor esperado de siniestros a pagar. La pérdida agregada para cada segmento h se obtiene mediante iteraciones s del algoritmo MCMC de la siguiente manera:

- El número de sinestros ocurridos $N_h^{(s)}$ en la iteración s de la clase h hasta el tiempo t es tal que:

$$N_h^{(s)} \leftarrow w_h^{(s)} \cdot \sum_{i=1}^{t} M_i, \tag{22}$$

  donde M_i es tal que:

$$M_i \overset{\mathrm{d}}{=} \text{Poisson}\left(N_i \cdot \left(\boldsymbol{\beta}^{(s)}\right)^T \cdot \mathbb{X}_i\right).$$

Note que $\boldsymbol{\beta}^{(s)}$ es el vector en la iteración s de las muestras posteriores de parámetros del mejor modelo predictivo de frecuencia para el número de siniestros ocurridos, N_i es el número de riesgos asegurados en la semana i y $\boldsymbol{X}$ es la matriz de diseño con la primera columna como intercepto y la segunda los índices de tiempo hasta el tiempo t.

- El monto de pérdida agregada S_h del segmento h en la iteración s es:

$$S_h^{(s)} \leftarrow \sum_{i=1}^{N_h^{(s)}} X_{i,h},$$

donde $X_{i,h}$ es tal que:

$$X_{i,h} \overset{\mathrm{d}}{=} \text{lognormal}\left(\theta_h^{(s)}, \left(\sigma_h^2\right)^{(s)}\right).$$

Una vez obtenidas las distribuciones de pérdida agregada en cada uno de los segmentos h con $h \in \{1, \ldots, 5\}$, la pérdida agregada total esperada para todo el portafolio es de $9'299.259$US; mientras que el valor observado del monto total de los siniestros es de $9'314.604,4$US. Los resultados de pérdida agregada por segmento con sus respectivos intervalos de credibilidad del 95 % se presentan en seguida:

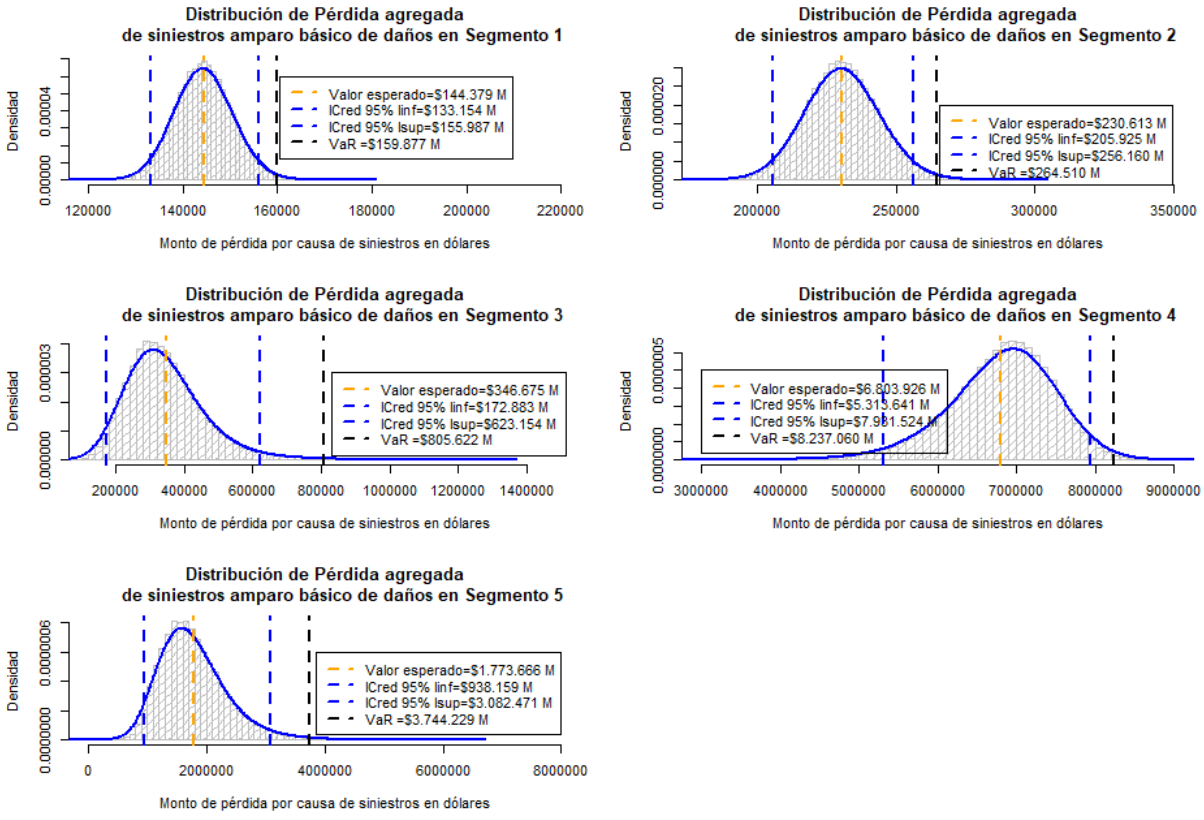

Figura 58: Distribuciones predictivas de pérdida agregada de siniestros para cada clase latente.

Las pérdidas presentadas anteriormente aún no están amparadas bajo un contrato de reaseguro proporcional o no proporcional. Estos contratos tienden a eliminar una tasa sobre el monto de pérdida agregada,

pero implican el aumento de la prima comercial que se cobra al asegurado al incluir contratos de reaseguro que típicamente afectan cuantías muy altas asociadas posiblemente a 'grandes' valores asegurados. Sin embargo, el efecto de un contrato de reaseguro sobre la mitigación de la pérdida causada por siniestros será analizado en la siguiente sección.

Las simulaciones realizadas anteriormente para obtener el valor esperado de probabilidad de ruina permiten obtener un indicador denominado por la Superintendencia Financiera de Colombia como índice combinado. Este indicador muestra un resultado técnico que reporta el porcentaje sobre ingresos o primas devengadas que 'gana' o 'pierde' la compañía aseguradora cuando realiza el pago de sus obligaciones de siniestros y gastos.

índice combinado y resultado técnico.

El índice combinado I según la Superintendencia Financiera de Colombia es un indicador que se calcula mediante la siguiente ecuación:

$$I = 100 \cdot \frac{\text{Monto de Gastos} + \text{Monto de Siniestros}}{\text{Monto de Primas Devengadas}}.$$

Tomando las simulaciones obtenidas para $P(t)$, $G(t)$ y $S(t)$ correspondientes a los montos de primas devengadas, gastos y siniestros, respectivamente en el proceso de riesgo a tiempo continuo para estimar el valor esperado de probabilidad de ruina en un horizonte temporal de 12 meses, es posible calcular la distribución predictiva del índice combinado I por medio de S iteraciones del algoritmo MCMC. De este modo, se define $I^{(s)}$ en cada iteración $s \in \{1, \ldots, S\}$ de la siguiente forma:

$$I^{(s)} = 100 \cdot \frac{G(t)^{(s)} + S(t)^{(s)}}{P(t)^{(s)}}.$$

Si el tiempo t es horizonte temporal de doce meses, se tiene la distribución predictiva del índice combinado:

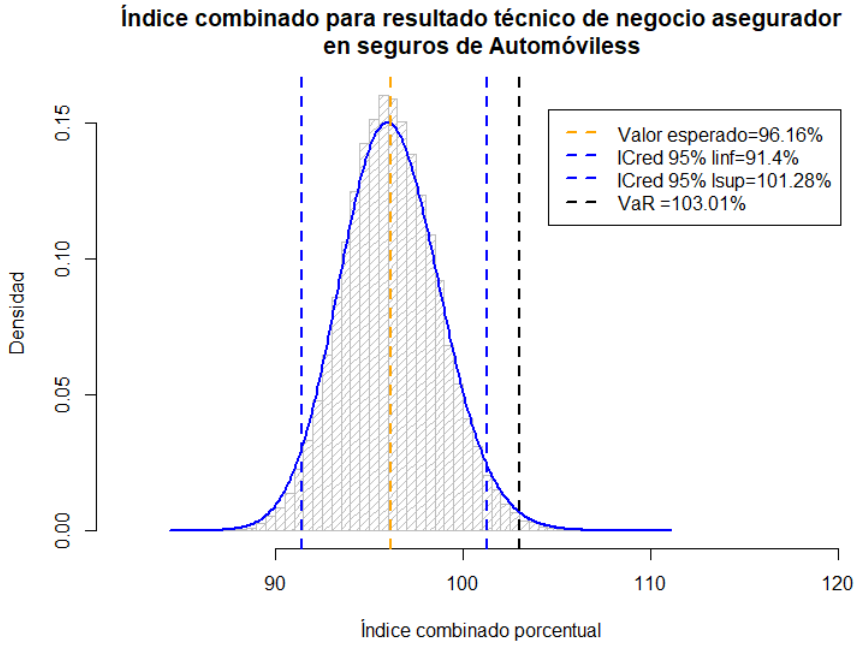

Figura 59: Distribución predictiva del índice combinado.

Los resultados permiten evidenciar que al hacer el lanzamiento de un producto de automóviles con amparo básico de daños por mayor y/o menor cuantía, se espera que a 31 diciembre de 2021 el índice combinado tome un valor del $96,16\,\%$. Adicionalmente se espera que el valor del índice combinado se encuentre entre $91,4\,\%$ y $101,28\,\%$ con una probabilidad del $95\,\%$. En términos de rentabilidad frente al negocio de suscripción de pólizas de automóviles bajo esta cobertura de daños por mayor y/o menor cuantía se tienen los siguientes escenarios para la compañía aseguradora:

Tabla 51: *La tabla presenta la siguiente información en sus columnas al realizar lectura de la misma de izquierda a derecha: Escenario para la compañía aseguradora, valor del índice combinado, resultado técnico porcentual y monto de rentabilidad o utilidad en dólares.*

Escenario	Índice combinado	Resultado técnico	Monto de rentabilidad
Optimista	91,400 %	8,600 %	$1'267.095,706$ US
Central	96,160 %	3,840 %	$565.772,9664$ **US**
Pesimista	101,280 %	-1,280 %	$-188.590,9888$ **US**

En términos generales es posible concluir que mediante el proceso de tarifación para cobertura de daños por mayor y/o menor cuantía en seguros de Automóviles, se espera que en el escenario más optimista la compañía obtenga al final del año 2021 una utilidad de $1'267.095,706$ US. En un escenario central, la utilidad esperada es de $565.772,9664$ **US**; mientras que en un escenario pesimista se esperan pérdidas por $188.590,9888$ **US**. Sin embargo, al establecer una articulación de estos resultados con la probabilidad de ruina esperada en un horizonte temporal de 12 meses, se puede afirmar que la expectativa para la compañía aseguradora es de ganancia dado que la probabilidad esperada de que las primas devengadas no sean suficientes para hacer frente a obligaciones de gastos y siniestros a corte de diciembre 31 de 2021 es del $24\,\%$. En contraste, la probabilidad de que las primas sean suficientes para responder a las obligaciones inclusive generando utilidades es del $76\,\%$.

6.4. Simulaciones para contratos de reaseguro proporcional y no proporcional.

Una forma de mitigar las pérdidas causadas por las cuantías de los siniestros es realizar un contrato de reaseguro proporcional o no proporcional que le permita a la compañía aseguradora ceder algunas de sus obligaciones de pagos a entidades reaseguradoras que tienen determinada participación dentro del contrato. Típicamente, las compañías realizan la contratación de un reaseguro proporcional y no proporcional, pero deben incluir este costo en la tasa comercial que se cobra al asegurado al momento de establecer el contrato de seguros. Debido a lo anterior, la compañía aseguradora debe ser 'cuidadosa' con las tarifas y el cobro de estas dado que requiere ser competitiva frente a otras compañías del sector que se encuentren ofreciendo el mismo producto. En esta aplicación práctica que ilustra la metodología de un modelo Bayesiano de mezcla que permite medir el riesgo de suscripción en seguros de no vida o Seguros Generales, se definirán las siguientes condiciones de contratación para un reaseguro proporcional o no proporcional:

- Condición reaseguro proporcional cuota parte: En el año de suscripción 2021, si el siniestro de daños por mayor y/o menor cuantía de un vehículo asegurado cuesta al menos 15.000 US, la compañía aseguradora asume el $60\,\%$ de las obligaciones mientras que el restante $40\,\%$ es cedido a las compañías reaseguradoras pactadas dentro del contrato considerando sus porcentajes de participación.

- Condición reaseguro no proporcional XL: Si al hacer la cesión del $40\,\%$ en la condición de cuota parte para las obligaciones de los siniestros con cuantías de al menos 15.000 US, y la compañía retiene más de 2.000 US, tendrá una retención prioritaria y hará cesiones de reaseguro no proporcional en cuatro capas de la siguiente manera:

1. El valor de la retención prioritaria es de 2.000 US.
2. El máximo valor cedido en la primera capa es de 5.000 US
3. El máximo valor cedido en la segunda capa es de 8.000 US
4. El máximo valor cedido en la tercera capa es de 10.000 US
5. El máximo valor cedido en la cuarta capa es de 15.000 US

- Condición de límites en capas de reaseguro no proporcional XL: La primera capa de reaseguro tiene una cobertura de hasta 90.000 US, la segunda capa un límite de hasta 144.000 US, la tercera capa un límite de hasta 180.000US y la cuarta capa un límite de hasta 270.000US. En caso de que los límites de una capa se encuentren al tope, la compañía aseguradora está facultada para realizar su respectivo reinstalamento ampliando de este modo los límites de cobertura. En caso de no tener reinstalamentos para aquellas capas que se encuentran llenas, la compañía asume como retención propia el pago de las obligaciones que no puedan ser asumidas por las capas del contrato.

Desde el punto de vista de la toma de decisiones, el modelo Bayesiano de frecuencia y severidad definido para la medición del riesgo de suscripción permite estudiar de forma anticipada a la ocurrencia de los siniestros, si el contrato de reaseguro bajo cada una de sus condiciones es un negocio 'inteligente' para la compañía aseguradora. Mediante un estudio de simulación es posible estimar: el porcentaje del monto límite que será cubierto por las capas de reaseguro no proporcional al hacer cesión de las obligaciones, así como las tasas de eliminación de pérdida que genera el contrato de reaseguros y cada una de las condiciones de reaseguro cuota parte y no proporcional. El estudio de simulación se realiza mediante la siguiente metodología de S iteraciones:

1. Para cada iteración s tal que $s \in \{1, \ldots, S\}$, guardar el número esperado de siniestros $N(t = 52)^{(s)}$ que ocurrirán dentro de un año (se asume que en el año se tienen 52 semanas y de allí que $t = 52$):

$$N(t = 52)^{(s)} \leftarrow \sum_{i=1}^{t} M_i,$$

donde M_i es tal que:

$$M_i \overset{\text{d}}{=} \text{Poisson}\left(N_i \cdot \left(\boldsymbol{\beta}_{\text{post}}\right)^T \cdot \mathbb{X}_i\right),$$

de manera que N_i es el total de riesgos asegurados en la semana i, $\boldsymbol{\beta}_{\text{post}}$ es la media posterior del vector $\boldsymbol{\beta}$ de parámetros del modelo de frecuencia Poisson con enlace lineal de coeficientes positivos, y $\mathbb{X}_i$ es el vector de la $i-$ésima fila de la matriz de diseño $\mathbb{X}$ con columnas de: intercepto e índices entre 1 y t.

2. En la iteración s estimar las cuantías de los $N(t = 52)^{(s)}$ siniestros ocurridos de la siguiente manera:

$$\begin{aligned} X_{i,h}^{(s)} \mid Z_{i,h} &\overset{\text{d}}{=} \text{lognormal}\left(\theta_{h,\text{post}}, \sigma^2_{h,\text{post}}\right) \\ Z_{i,h} &\overset{\text{d}}{=} \text{Multinomial}(w_{1,\text{post}}, \ldots, w_{5,\text{post}}), \end{aligned}$$

para todo $i \in \{1, \ldots, N(t = 52)^{(s)}\}$, donde $\theta_{h,\text{post}}, \sigma^2_{h,\text{post}}, w_{h,\text{post}}$ sson las medias posteriores de cada parámetro del modelo Bayesiano de mezcla con número de componentes fijo para cada clase h con $h \in \{1, \ldots, 5\}$.

3. Bajo las condiciones de contratos de reaseguro proporcional y no proporcional definidas, establecer en cada iteración s las obligaciones de siniestros retenidas por la compañías aseguradora y cedidas a las contrapartes reaseguradoras para cada siniestro $X_i^{(s)}$ con $i \in \{1, \ldots, N(t = 52)^{(s)}\}$, considerando que los siniestros ocurren de forma aleatorio dentro del año en curso. Dado el caso que las capas se encuentren llenas, es necesario asignar los valores que estaban destinados a estas como retenciones de la compañía.

El estudio de simulación se realizó con un total de iteraciones $S = 1.000$, encontrando que si la compañía acepta esta propuesta de contrato se enfrentaría a tener las dos primeras capas llenas en todos los escenarios y de este modo, asumiendo estas obligaciones como fondos propios en caso de no hacer reinstalmentos de estas capas. Para observar el porcentaje esperado de consumo de las capas con respecto a sus límites pactados en el contrato de reaseguro además de sus respectivos intervalos de credibilidad del 95 % se presentan a continuación las distribuciones de estas cesiones:

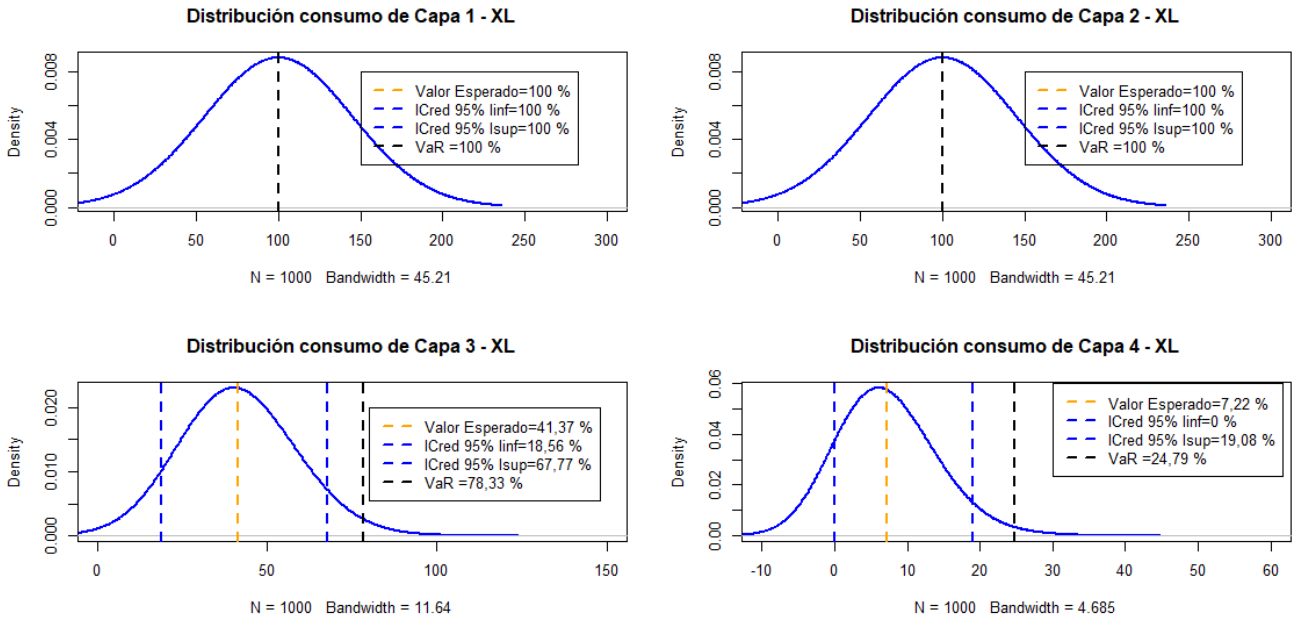

Figura 60: Distribuciones del consumo de las capas de reaseguro no proporcional XL según sus límites pactados en el contrato.

Los resultados presentados anteriormente permiten tomar importantes decisiones con respecto a la aceptación de este contrato de reaseguro previo a la ocurrencia de los siniestros porque en el 100 % de los escenarios simulados, las capas 1 y 2 alcanzaron su límite del 100 %, razón por la cual se requieren reinstalamentos de estas dos capas para absorber posibles pérdidas a las que deba hacer frente la compañía. Por otra parte, se observa que la capa 3 alcanza un consumo esperado del $41,37$ % y con una probabilidad del 95 %, se espera que el consumo de la capa se encuentre entre el $18,56$ % y el $67,77$ %. Con relación a la capa 4 se observa que el consumo esperado es del $7,22$ % de manera que con una probabilidad del 95 % el consumo estará entre el 0 % y el $7,22$ %. En caso de que en la práctica, la capa 4 tenga un consumo nulo, se tiene un indicador de que el contrato no representa un 'buen' negocio para la compañía aseguradora porque ha pactado una condición que no utiliza para respaldar sus obligaciones frente a los siniestros más 'severos' de manera que ha debido retener montos de pago en sus fondos propios a causa del llenado de las capas 1 y 2. La medida de VaR corresponde al percentil $99,5$ % de la distribución, de manera que en un escenario que sucede 1 entre 200 veces, las capas 3 y 4 presentan un consumo del $78,3$ % y $24,79$ %, respectivamente.

Por otra parte, la tasa de eliminación de pérdida t_E para la compañía aseguradora por el contrato de reaseguros se puede estimar de la siguiente manera:

$$t_E = 100 \cdot \frac{\mathbb{E}(S) - \mathbb{E}(S \mid L)}{\mathbb{E}(S)},$$

donde S es el monto de pérdida agregada del proceso estocástico $S = \sum_{i=1}^{N(t)} X_i$ y $S \mid L$ es el monto retenido por la compañía una vez cedidas sus obligaciones a los reaseguradores bajo condiciones y límites L pactados en el contrato. Las tasas de eliminación de pérdida para la compañía aseguradora por causa del contrato de reaseguros y de sus condiciones cuota parte y XL se presentan a continuación con las distribuciones que presentan los respectivos intervalos de credibilidad del 95 % asociados a las tasas de eliminación esperadas:

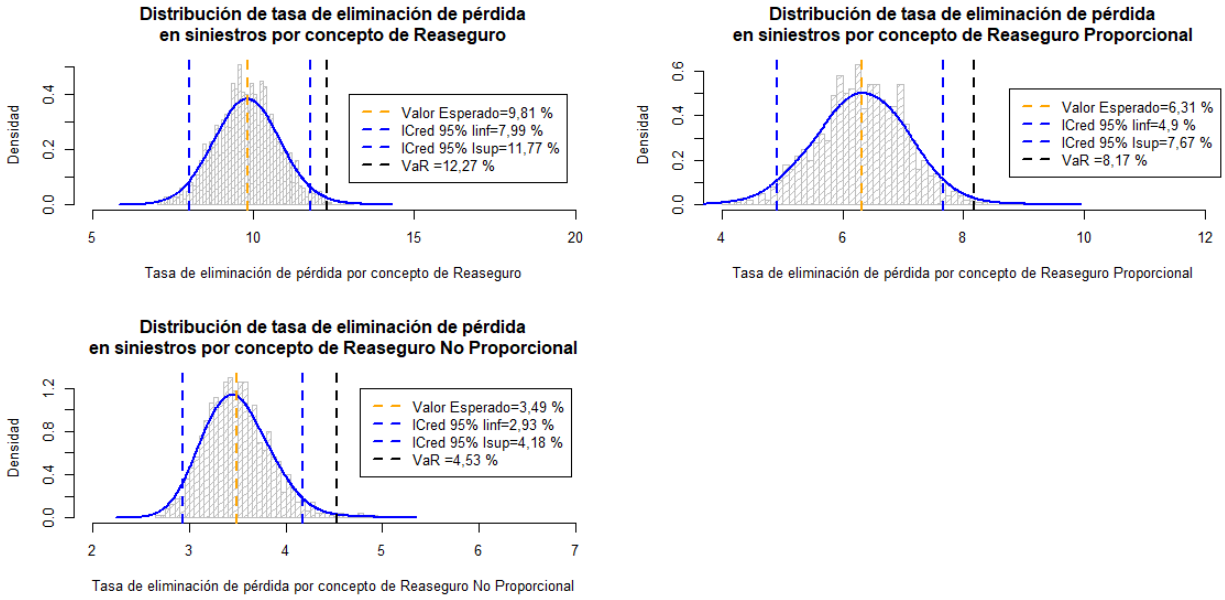

Figura 61: Distirbuciones de las tasas de eliminación de pérdida para la compañía aseguradora por efectos de contrato de reaseguro.

Los resultados permiten observar que la tasa de eliminación de pérdida esperada por la aceptación del contrato de reaseguros es del $9,81\,\%$ y con una probabilidad del $95\,\%$ esta se encuentra entre $7,99\,\%$ y $11,77\,\%$. Haciendo una desagregación de las condiciones de contrato de reaseguro proporcional y no proporcional se tiene que la tasa de eliminación de pérdida esperada para la condición cuota parte es del $6,31\,\%$ de manera que con una probabilidad del $95\,\%$ esta tasa se encuentra entre $4,9\,\%$ y $7,67\,\%$. La condición no proporcional XL permite evidenciar una tasa esperada de eliminación de pérdida del $3,49\,\%$ con límites entre $2,93\,\%$ y $4,18\,\%$ con una probabilidad del $95\,\%$. Finalmente, en un escenario que ocurre 1 entre 200, la tasa de eliminación de pérdida por causa de todo el contrato de reaseguro es del $12,27\,\%$; mientras que esta tasa para el reaseguro proporcional es de $6,31\,\%$ y para el reaseguro no proporcional es de $4,53\,\%$.

En conclusión, para la toma de decisiones del contrato de reaseguro óptimo para la compañía es necesario mostrar mediante simulaciones y predicciones que las capas no alcancen su límite para que no deba asumirse con fondos propios el valor de proporciones de cuantías 'severas' causadas por los siniestros. Adicionalmente, el contrato óptimo debe tener la mayor tasa esperada de eliminación de pérdida para la compañía aseguradora.

6.5. Requerimiento de capital para riesgo de suscripción en seguros de no vida bajo Directiva de Solvencia II.

Dentro de los ejercicios de impactos cuantitativos que viene adelantando Fasecolda actualmente en conjunto con las compañías aseguradoras del país, se está estableciendo una fórmula estándar que permita

definir una normativa con un capital regulatorio que deben constituir las aseguradoras para que ante eventos inesperados o catastróficos puedan responder a estas obligaciones sin entrar en estados de insolvencia. Al respecto, fue expedido un documento conocido como estudio para la implementación de una regulación basada en riesgos en Colombia para octubre de 2020. En este documento se considera que el riesgo de suscripción de no vida en una compañía aseguradora depende de: el riesgo de primas y reservas, riesgo de caída y riesgo catastrófico. Este último se relaciona con eventos de siniestros causados por la naturaleza como por ejemplo terremotos, inundaciones, maremotos, etc. Para ilustrar la metodología y el aporte de los modelos Bayesianos de mezcla con número de componentes fijo o aleatorio para la severidad de los siniestros en Directivas como la de Solvencia II, se considera únicamente el volumen de primas y reservas bajo el conjunto de datos de pólizas y siniestros de automóviles para la cobertura de amparo básico de daños por mayor y/o menor cuantía. Escogiendo el 31 de diciembre como fecha de corte, el requerimiento de capital para riesgo de suscripción de volumen de primas y reservas es calculado a partir de los siguientes input:

- P_s: Es la mejor estimación de las primas que devengará la compañía aseguradora entre el 1 de enero de 2022 y el 31 de diciembre de 2022 en la línea de negocio s, para el caso automóviles.
- $P_{\text{last},s}$: Es el total de primas devengadas entre el 1 de enero de 2021 y el 31 de diciembre de 2021 para la línea de negocio de automóviles.
- $\text{FP}_{\text{ex},s}$: Es el valor presente del monto esperado de primas a devengar entre el 1 de enero de 2023 y el 31 de diciembre de 2023 para contratos de seguros de autos que no hayan finalizado su vigencia y que fueron suscritos en 2021.
- $\text{FP}_{\text{fut},s}$: Es el valor presente del monto esperado de primas a devengar entre el 1 de enero de 2023 y el 31 de diciembre de 2023 nuevos contratos de seguros de autos que hayan sido celebrados entre el 1 de enero de 2022 y el 31 de diciembre de 2022. Se definen como nuevos contratos aquellas pólizas de seguros que sean renovadas o que correspondan a pólizas de nuevos clientes.
- PCO_s: Es la mejor estimación de los siniestros pendientes de pago de la compañía aseguradora en la línea de negocio de autos a corte de 31 de diciembre de 2021.

Mediante una metodología de simulación que pretende considerar las iteraciones de cálculo para las reservas IBNR y RSAP, es posible realizar el cálculo de requerimiento de capital mediante los siguientes pasos y supuestos de crecimiento para la compañía aseguradora en 2022.

Supuestos e hipótesis.

1. Una vez vencida la póliza de autos par cobertura de daños por mayor y/o menor cuantía, la probabilidad de renovación por el mismo periodo es del 98 %.
2. La proporción de contratos por concepto de nuevos clientes que la compañía aseguradora pretende celebrar en 2022 corresponde al 5 % de las pólizas renovadas.

Pasos para cálculo de requerimiento de capital.

Realizar S iteraciones de manera que:

- En la iteración s generar la variable de renovación R_i para el contrato i de modo que:

$$R_i \overset{\text{d}}{=} \text{Bernouilli}(\theta = 0.98),$$

para todo $i \in \{1, \ldots, M\}$, donde M representa el número de asegurados en 2021. Posteriormente guardar el conjunto de datos de aquellas pólizas que renovarán su contrato con la compañía aseguradora en 2020.

- Calcular las fechas de inicio y fin de vigencia de los contratos renovados sumando un año a cada una.
- Calcular la prima comercial, gastos administrativos y gastos comerciales asociados a cada póliza renovada en 2022.
- Sea M_{2022} el número de contratos que renovaron su póliza con la compañía aseguradora. Dicho de otro modo,

$$M_{2022} = \sum_{i=1}^{M} R_i.$$

  En la iteración s, calcular el tamaño de muestra n de contratos para nuevos clientes tal que $n = 0.05 \cdot M_{2022}$. Obtener una muestra estratificada MAS de los contratos renovados con tamaño n cuyos estratos son los segmentos o clases latentes del modelo Bayesiano de mezcla. Paso seguido a este, construir un conjunto de datos que una las pólizas renovadas con la muestra obtenida. Note que este conjunto contiene todos los nuevos contratos de 2022.
- Calcular la reserva de prima no devengada $\text{RPND}^{(s)}$ de la iteración s calculando el valor de esta reserva por cada nuevo contrato i. Esta última reserva se denota por $\text{RPND}_i^{(s)}$ y la fecha de cálculo es 31 de diciembre de 2022. Guardar $\text{RPND}^{(s)}$ tal que:

$$\text{RPND}^{(s)} \leftarrow \sum_{i=1}^{M_{\text{nuevos}}} \text{RPND}_i^{(s)},$$

  donde M_{nuevos} representa el total de nuevos contratos en 2022.
- Guardar el monto de la mejor estimación de las primas $P_s^{(s)}$ que se devengarán a corte de diciembre 31 de 2022:

$$P_s^{(s)} \leftarrow \text{RPND}_{2021} + \sum_{i=1}^{M_{\text{nuevos}}} \text{PC}_i - \text{RPND}^{(s)},$$

  donde PC_i es la prima comercial sin impuesto de IVA del contrato i y RPND_{2021} es la reserva de prima no devengada a corte de 31 de diciembre de 2021.
- Guardar el monto de primas devengadas $P_{\text{last},s}^{(s)}$ a corte de 31 de diciembre de 2021:

$$P_{\text{last},s}^{(s)} \leftarrow \sum_{i=1}^{M} \text{PC}_{i,2021} - \text{RPND}_{2021},$$

  donde $\text{PC}_{i,2021}$ es la prima comercial de cada contrato de seguros suscrito en 2021.
- Guardar $\text{FP}_{\text{ex},s}^{(s)}$ tal que:

$$\text{FP}_{\text{ex},s}^{(s)} \leftarrow 0.$$

  Este resultado es consecuencia de que los contratos de seguros de autos se han definido con un límite máximo de vigencia de un año. De este modo no se tienen primas por devengar en 2023 de contratos suscritos en 2021.
- Guardar $\text{FP}_{\text{fut},s}^{(s)}$ tal que:

$$\text{FP}_{\text{fut},s}^{(s)} \leftarrow \text{RPND}^{(s)} \cdot (1+v)^{-1},$$

  donde v es una tasa de interés técnico del 4 % para actualizar flujos futuros a valor presente.
- Guardar el monto de siniestros pendientes de pago $\text{PCO}_s^{(s)}$ de la iteración s tal que:

$$\text{PCO}_s^{(s)} \leftarrow \text{IBNR}^{(s)} + \text{RSAP}^{(s)}.$$

- Construir el vector columna $\mathrm{V}_s^{(s)}$ de tamaño 22×1, de manera que la $i-$ésima componente de la $s-$ésima iteración es tal que:

$$\mathrm{V}_{i,s}^{(s)} \leftarrow \begin{cases} 0 & ,\text{si } i \neq 2 \\ \text{máx}\left\{P_s^{(s)}, P_{\text{last},s}^{(s)}\right\} + \mathrm{FP}_{\text{ex},s}^{(s)} + \mathrm{FP}_{\text{fut},s}^{(s)} + \mathrm{PCO}_s^{(s)} & ,\text{si } i = 2. \end{cases}$$

  La definición de un vector es consecuencia de que la estimación de volumen de primas y reservas para riesgo suscripción se realiza mediante productos matriciales. Por otra parte, los componentes del vector $\mathrm{V}_s^{(s)}$ diferentes al componente 2 toman el valor de 0 porque corresponden a líneas de negocio de seguros de no vida diferentes a automóviles. En contraste con lo anterior, para el componente 2 se relaciona un cálculo relacionado con los ítems de primas devengadas en Solvencia II para automóviles dado que el conjunto de datos de estudio se asocia con esta línea de negocio.

- Guadar el vector columna $\boldsymbol{\sigma}_s^{(s)}$ de tamaño 22×1 en la iteración s tal que su componente i es:

$$\boldsymbol{\sigma}_{i,s}^{(s)} \leftarrow \left[\frac{\sigma_{\text{prem,i}} \cdot \left(\mathrm{V}_{i,s}^{(s)} - \mathrm{PCO}_{i,s}^{(s)}\right)^2 + \sigma_{\text{prem,i}} \cdot \sigma_{\text{res,i}} \cdot \left(\mathrm{V}_{i,s}^{(s)} - \mathrm{PCO}_{i,s}^{(s)}\right) \cdot \mathrm{PCO}_{i,s}^{(s)} + \sigma_{\text{res,i}} \cdot \left(\mathrm{PCO}_{i,s}^{(s)}\right)^2}{\mathrm{V}_{i,s}^{(s)}}\right]^{\frac{1}{2}},$$

  donde los valores para $\sigma_{\text{prem,i}}$ y $\sigma_{\text{res,i}}$ con $i \in \{1, \ldots, 22\}$ se encuentran en el documento de Estudio para la implementación de una regulación basada en riesgos en Colombia (Fasecolda, 2020), y $\mathrm{PCO}_{i,s}^{(s)}$ está definido como sigue:

$$\mathrm{PCO}_{i,s}^{(s)} = \begin{cases} 0 & ,\text{si } i \neq 2 \\ \mathrm{PCO}_s^{(s)} & ,\text{si } i = 2. \end{cases}$$

  Esta última definición se asocia nuevamente con el hecho de que el conjunto de datos estudiado se relaciona con una línea de negocio de automóviles y esta corresponde al componente 2 del vector anterior. De este modo, en las demás líneas de negocio de seguros de no vida se asume un valor de 0.

- Guardar en la iteración s el vector columna $\boldsymbol{S}_v^{(s)}$ de tamaño 22×1 cuya componente i está dada por:

$$\boldsymbol{S}_{i,v}^{(s)} = \boldsymbol{\sigma}_{i,s}^{(s)} \cdot \mathrm{V}_{i,s}^{(s)}.$$

- Determinar el requerimiento de capital $\mathrm{SCR}^{(s)}$ en la iteración s de la siguiente manera:

$$\mathrm{SCR}^{(s)} \leftarrow 3 \cdot \left(\left(\boldsymbol{S}_v^{(s)}\right)^T \cdot \Sigma_c \cdot \boldsymbol{S}_v^{(s)}\right)^{\frac{1}{2}},$$

  donde la matriz Σ_c de tamaño 22×22 es la matriz de correlaciones entre riesgo de primas y reservas para cada línea de negocio de seguros de no vida. Los detalles de esta matriz se encuentran en Estudio para la implementación de una regulación basada en riesgos en Colombia (Fasecolda, 2020). Observe que en la ecuación anterior el valor de la raíz cuadrada del producto matricial se encuentra multiplicada por 3. Desde el punto de vista de la Directiva de Solvencia II, esto sucede porque tres volatilidades se corresponden con un nivel de estrés 'alto' para el volumen del requerimiento de capital. Sin embargo, en cada una de las iteraciones s con $s \in \{1, \ldots, S\}$, el cálculo se realizó con una o dos volatilidades para fines comparativos del volumen de capital bajo cada uno de estos escenarios.

Los resultados de las distribuciones predictivas se presentan a continuación. Se consideran distribuciones predictivas desde el punto de vista de que el requerimiento de capital $\text{SCR}^{(s)}$ depende de los valores de las reservas técnicas de siniestros ocurridos avisados y ocurridos no avisados, las cuales están en función de las iteraciones del algoritmo MCMC del modelo Bayesiano de frecuencia y de mezcla con número de componentes fijo o aleatorio para la severidad de los siniestros.

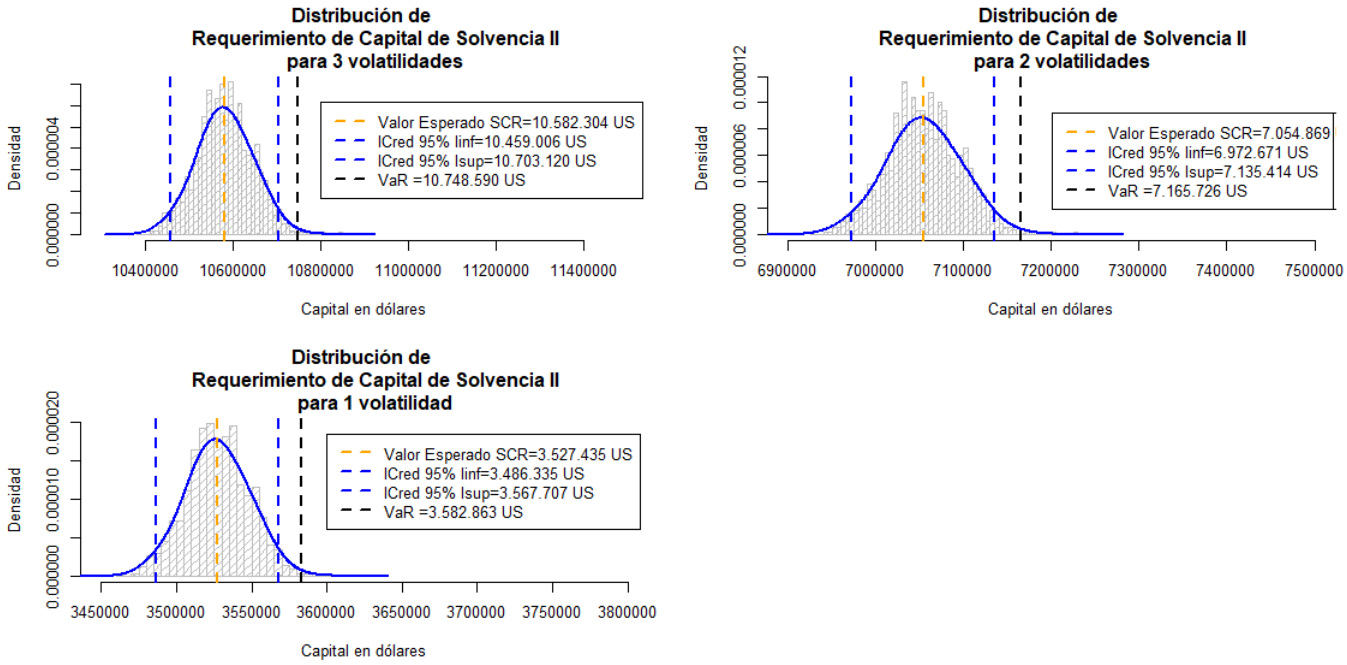

Figura 62: Distribuciones predictivas de los requerimientos de capital a una, dos y tres volatilidades.

Los resultados muestran que desde el punto de vista del volumen de primas y reservas para el riesgo de suscripción de una línea de negocio de automóviles con un conjunto de datos de interés con el cual se ilustró la metodología del presente documento para un amparo básico de daños por mayor y/o menor, el capital regulatorio a constituir es de 10′582.304 US. Sin embargo, si la calibración de la fórmula estándar se realiza posteriormente con una o dos volatilidades, el capital regulatorio corresponde a 3′582.863US o 7′054.869US.

7. Conclusiones

Los modelos Bayesianos para frecuencia y severidad de los siniestros considerando una mezcla con número de componentes aleatorio y distribución lognormal para este último, se constituyen en una herramienta fundamental para la mejor estimación del monto de pérdida agregada en siniestros y el cálculo de una tarifa de seguros de no vida, dado que superan posibles falacias que tienen los modelos clásicos cuando no resulta posible ajustar una distribución univariada a un conjunto de datos multimodal o cuando no es posible garantizar la bondad de ajuste desde el punto de vista estadístico para un modelo de mezcla clásico en todas en sus componentes.

Las garantías de tarifas 'justas' que permiten mitigar el riesgo de suscripción en productos y seguros de no vida se relacionan directamente con la robustez de modelos de frecuencia Bayesianos y modelos de mezcla con número de componentes aleatorio para la severidad de los siniestros dado que los estudios de simulación, sensibilidad, bondad de ajuste y estimación adecuada de datos faltantes en metodologías de validación cruzada como k-fold.

La toma de decisiones para compañías aseguradoras con respecto al riesgo de suscripción basada en modelos Bayesianos de frecuencia y severidad con número de componentes aleatorio permite definir por medio de simulaciones y distribuciones predictivas el negocio más inteligente desde el punto de vista del negocio más 'inteligente' en el largo plazo de un contrato de reaseguros que permita el mayor beneficio de la compañía con relación a la máxima tasa de eliminación de pérdidas por causa de siniestros, la menor retención y el mayor costo beneficio como consecuencia de no tener que realizar reinstalamentos en algunas de las capas de reaseguro no proporcional.

Los avances en materia de modelos Bayesianos de frecuencia y modelos de mezcla con número de componentes aleatorio para la cuantía de los siniestros representan una potencial aplicación para el negocio asegurador que va desde el monitoreo de cálculos estocásticos para los resultados técnicos y/o rentabilidad del negocio y la probabilidad de insolvencia hasta una fácil adaptabilidad con respecto a una Directiva de Solvencia y una regulación basada en riesgos como la que se encuentra proponiendo Fasecolda actualmente en el sector asegurador colombiano en compañías de las diferentes entidades del sector que se encuentran participando en los ejercicios de impactos cuantitativos que permitan el cálculo y la calibración de una fórmula estándar para el capital regulatorio del riesgo de suscripción.

Una de las preguntas fundamentales que permitió tan importante avance en metodologías de modelos Bayesianos para la medición del riesgo de suscripción en seguros de no vida surgió en las metodologías de los modelos clásicos para abordar distribuciones univariadas dado que estas parecen presentar dificultades en distribuciones multimodales para la severidad de los siniestros. Sin embargo, en caso de mezclas con distribuciones de colas pesadas donde no sea posible encontrar fácilmente un número óptimo y aleatorio de clases latentes sumado a costos computacionales importantes, la adaptación del modelo Bayesiano de mezcla con distribución lognormal y número de componentes aleatorio a una versión no paramétrica con número infinito de clases resultaría de gran aporte para la generación de tarifas en productos de seguros asociadas a conjuntos de datos que reúnan estas condiciones y se encuentren fuera del alcance de la metodología propuesta. De esta manera, el desarrollo de modelos no paramétricos se constituye en un desafío y una importante oportunidad para estudios futuros tal y como lo presentan Jain, S & Neal, R. (2012) al implementar el algoritmo de fusión y separación de clases latentes para un proceso Dirichlet (DP) vía cadenas de Markov de Monte Carlo.

8. Bibliografía

Referencias

[1] Bulhmann, H. Straub, E. (1972) Credibility for Loss Ratios *(Translated by C.E. Brooks) ARCH*

[2] Philbrick, S. (1981) An Examination of Credibility Concepts *Proceedings of the Casuality Actuarial Society*

[3] Tse, Y. (2009) Nonlife Actuarial Models Theory, Methods and Evaluation *Cambridge University Press, New York*

[4] EIOPA 11/163. (2011) Calibration of the Premium and Reserve Risk Factors in the Standard Formula of Solvency II. *Report of the Joint Working Group on Non - Life and Health NSLT Calibration*

[5] EIOPA - 14 - 322. (2014) The underlyng assumptions in the standard formula for the Solvency Capital Requeriment calcilation.

[6] Gelman, A. et al. (2014) Bayesian Data Analysis *Texts in Statistical Science, Third Edition, New York*

[7] khapeava, T. (2014) Credibility Modeling with Applications *Thesis presented in a partial fulfilment of the requirements or the degree of Master of Science (M.Sc.) in Computational Sciences, Laurentian University*

[8] Migon, H. et al. (2015) Statistical Inference An Integrated Approach *Texts in Statistical Science, Second Edition, New York*

[9] Jain, S & Neal, R. (2012) A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model *Journal of Computational and Graphical Statistics*

[10] SuperIntendencia Financiera de Colombia (2003). Disposiciones Especiales aplicables a la gestión de riesgos en las entidades aseguradoras *Parte II Título IV,Bogotá, Colombia*

[11] Blanco, L. (2005). Proabilidad *Universidad Nacional de Colombia, Bogotá*

[12] Rincón, L. (2012). Introducción a la teoría de Riesgo *Ciudad Universitaria UNAM, México D.F*

[13] Rincón, L. (2012). Introducción a los procesos estocásticos *Ciudad Universitaria UNAM, México D.F*

[14] Barañaño, A. De la Peña, E. Garayeta, A (2016). Medición del Riesgo de suscripción mediante modelos internos de solvencia II *Trabajo realizado en proyecto UFI 11/51 Dirección Empresarial y Gobernanza Territorial y Social de la UPV/EHU, Bilbao, España*

[15] Omary, C. Nyambura, S. Wairimu, J (2018). Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions. *Journal of Mathematical Finance*

[16] García, M (2014). Modelación de pérdida agregada aplicada a siniestros de cáncer de mama: Caso de una empresa aseguradora. *Universidad Autónoma del Estado de México, Toluca, 2014*

[17] Régimen de Seguros. *Libro digital de Fasecolda, http://publicaciones.fasecolda.com/regimen-de-seguros/*

[18] Estudio para la implementación de una regulación basada en riesgos en Colombia. *Fasecolda, 15 de octubre de 2020*

[19] Asan, U. & Ercan, S. (2012). An Introduction to Self-Organizing Maps. *ResearchGate*