Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 10 N°2 julio-diciembre de 2021 • ISSN-e 2357-5749 • Artículo Investigación • Páginas 28 a 50
DOI: https://doi.org/10.15446/rev.fac.cienc.v10n2.90438

# ENFOQUE BAYESIANO PARA OBTENER TASAS DE TRANSICIÓN EN UN MODELO DE MARKOV CON DOS ESTADOS RECURRENTES[a]

# BAYESIAN APPROACH TO OBTAINING TRANSITION RATES IN A MARKOV MODEL WITH TWO RECURRING STATES

ZULAY MARCELA GIRALDO BAUTISTA[b*] , JUAN CARLOS SALAZAR-URIBE[c] , RENÉ IRAL PALOMINO[d]

Recibido 9-09-2020, aceptado 20-04-2021, versión final 30-06-2021.
Artículo Investigación

**RESUMEN:** La aplicación de modelos de estados múltiples ha sido determinante a la hora de realizar estudios de datos longitudinales, tales como la observación de la progresión de una enfermedad en el tiempo, la recurrencia de una enfermedad, el seguimiento intermitente de la misma, entre otras; usualmente la forma cómo se mide el avance del fenómeno, es mediante el estado en el cual se pueda encontrar al sujeto en diferentes puntos en el tiempo. Las tasas de transición entre estados del fenómeno de estudio permiten evaluar si el individuo experimenta un cambio positivo o negativo del mismo, por tanto, se modela la manera como los individuos en cierta población transitan de un estado a otro a través del tiempo lo cual es importante para comprender su dinámica. Las tasas de transición en un modelo de Markov de dos estados recurrentes en función de covariables sé obtienen a través de un enfoque Bayesiano utilizando dos distribuciones apriori (No informativa e informativa); para esto se adoptó un esquema de análisis basado en el muestreador de Gibbs, mediante un estudio de simulación y aplicación a datos reales se ilustró el comportamiento de las tasas de transición bajo estas dos distribuciones y el efecto de una covariable.

**PALABRAS CLAVE:** Distribución Apriori; estados recurrentes; modelos de Markov; muestreador de Gibbs; probabilidad de transición; tasas de transición.

**ABSTRACT:** The application of multi-state models has been a decisive factor for studies of longitudinal data, such as observation of disease progression over time, recurrence of disease, intermittent monitoring, among others; usually the way to measure the progress of the phenomenon is to study the state in which the subject is found at different points in time. The transition rates between states of the phenomenon in study allows to assess whether the individual experiences a positive or negative change in its status,for this reason it is modeled how individuals in a certain population transit from one state to another through time, which is important to understand its dynamics. The transition rates in a Markov model of two recurrent states according to covariables are obtained by a Bayesian approach using two apriori distributions (Informative and Not Informative); to show this, an analysis scheme based on the Gibbs sampler

[a]Giraldo Bautista, Z. M., Salazar-Uribe, J. C. & Iral Palomino, R. (2021). Enfoque Bayesiano para obtener tasas de transición en un modelo de Markov con dos estados recurrentes. *Rev. Fac. Cienc.*, 10(**2**), 28–50. DOI: https://doi.org/10.15446/rev.fac.cienc.v10n2.90438

[b]Estadística, Magister en Ciencias - Estadística Facultad de Ciencias. Universidad Nacional de Colombia.
[*]Autor para la correspondencia: zmgirald@unal.edu.co
[c]Ph. D. en Estadística. Profesor Titular. Facultad de Ciencias. Universidad Nacional de Colombia
[d]M. Sc. en Estadística. Profesor Asociado. Facultad de Ciencias. Universidad Nacional de Colombia

was taken into account. Based on both a simulation study and an application to real data, it was possible to show the behavior of the transition rates under these two distributions and the effect of a covariable.

**KEYWORDS:** Apriori distribution; Gibbs sampler; Markov models; recurrent states, transition probability; transition rate.

# 1. INTRODUCCIÓN

Los fenómenos donde se presenta la recurrencia, - un evento recurrente es aquel que sucede en varias ocasiones para un mismo individuo (Cook & Lawless, 2007; Cardenas & Díaz, 2013) - son más frecuentes de lo que uno se imagina. Por ejemplo, los resfriados son recurrentes en la medida que una persona sana los contrae, se enferma y luego se alivia de nuevo y esto puede sucederle varias veces durante un determinado período de tiempo, o por ejemplo, una máquina empacadora de leche puede fallar y ser puesta en funcionamiento varias veces durante su vida útil (Baena & Salazar, 2006).

Una técnica que permite modelar las situaciones anteriores es la estimación de tasas de intensidad de transición, o simplemente tasas de transición que caracteriza un proceso de Markov con estados recurrentes las cuales proporcionan información directa del riesgo asociado de pasar de un estado a otro. Es conveniente asumir que estas tasas son funciones constantes del tiempo, (Kay, 1986). En este artículo se estudia una metodología general para modelar la posible interrelación entre el tiempo y la recurrencia de los estados del fenómeno de interes, donde las tasas de transición no dependen del tiempo, pero si dependen de algunas covariables involucradas en objeto de estudio.

Puede ser común encontrar trabajos de investigación con modelos de estados múltiples en diferentes áreas de conocimiento, pero particularmente en el área de la salud son bastante utilizados, estos son de ayuda para medir la progresión de enfermedades crónicas como el cáncer (Green & Byar, 2006), el VIH (Guihenneuc *et al.*, 2000) o la Artritis Reumatoide (Iral & Salazar, 2007)) en donde se estiman modelos de regresión exponenciales (Green & Byar, 2006), tasas de transición vía algoritmos estocásticos (Guihenneuc *et al.*, 2000), en trabajos como el de Iral & Salazar (2007) se muestra un modelo de Markov con tres estados donde estiman las tasas de transición por medio de un algoritmo de Newton - Raphson a través de ecuaciones de Kolmogorov, midiendo el efecto de covariables en la estimación. En Correa *et al.* (2010) se aborda el problema de estimación de las tasas de transición en modelo de Markov de tres estados por el método bayesiano MCMC basado en la discretización del soporte de la distribución, el cual es comparado con el reportado en Iral & Salazar (2007), en estudios más recientes Salazar et al. (2014) propone la estimación de tasas de transición en modelos de estados múltiples por medio del muestreador de Gibbs y comparado con el algoritmo de Newton - Raphson presentado en Iral & Salazar (2007) y el método bayesiano MCMC propuesto en Correa *et al.* (2010). De estos últimos estudios nombrados se tienen que los métodos bayesianos son efectivos y consistentes para abordar este tipo de problemas.

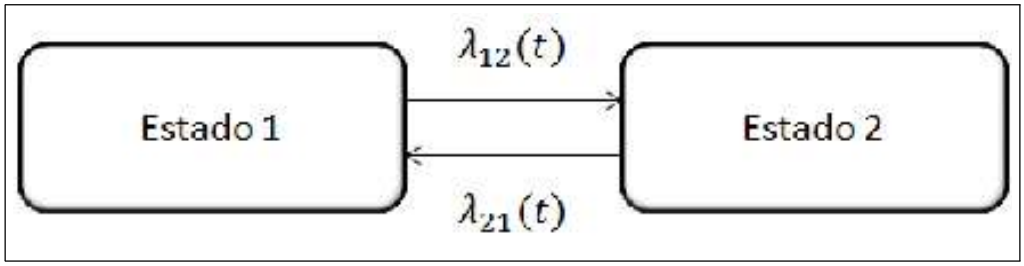


Figura 1: Proceso de Markov con dos estados recurrentes. Fuente: Elaboración Propia

A diferencia de los estudios mencionados anteriormente este artículo contempla el problema de recurrencia, el cual es tenido en cuenta en el modelo de estados múltiples con estados recurrentes, en trabajos como el de Jordan *et al.* (2008) se aborda la recurrencia a través de la aplicación de una cadena de Markov al problema de secuestro, aunque este tipo de problemas es principalmente abordado desde el análisis de supervivencia como por ejemplo Andersen y Gill (1982), Wei *et al.* (1989), Wang & Chang (1999), Peña *et al*. (2001), Martinez *et el.* (2009), Martinez *et al.* (2011) , Cardenas & Díaz (2013). El aporte más original de este artículo consiste en la predicción de las tasas de transición por medio de estadística Bayesiana. Especificamente se recurre al muestreador de Gibbs y se usan dos distribuciones Apriori una no informativa (Laplace) y otra informativa(Exponencial). Los detalles de este aporte se discuten en la sección 2. Por medio de un estudio de simulación se exploran las ventajas y las desventajas de las metodologías estudiadas (sección 3), además, también son estudiadas con datos longitudinales reales acerca de la recurrencia de ataques de virus informáticos a los computadores de una entidad bancaria en la sección 4 y finalmente, se discuten los méritos y las limitaciones de este enfoque en la sección 5.

## 2. EL MODELO

Los procesos de Markov han demostrado ser de mucha utilidad no solo en el estudio de algunas enfermedades tales como cirrosis, Alzheimer y esquizofrenia (Hendrie *et al.*, 2001; Harezlak *et al.*, 2003; Eichelsbacher & Ganesh, 2004), cáncer (Kay, 1986), entre otras, sino también para el análisis de fenómenos sociales como el secuestro (Jordan *et al.*, 2008) y tecnológicos como la recurrencia de ataques informáticos a equipos de cómputo (Valencia & Salazar, 2012).

Tomando el caso particular de ataques de virus informáticos a equipo informáticos se tiene un modelo de Markov con dos estados, en este caso: Sano e Infectado, los cuales son recurrentes debido a que tienen una probabilidad de regresar al estado anterior o simplemente quedarse en aquel que se encontraba inicialmente. La Figura 1 ilustra mejor el modelo de estados múltiples que puede utilizarse en este caso.

Sea $X(t)$ un proceso de Markov (con dos estados recurrentes), sea $S$ el espacio de estados, en el caso bajo estudio: $S=\{1,2\}$,donde se admiten las transisiones $1\to 1, 1\to 2, 2\to 1, 2\to 2$ y sea $P=[P_{ij}(t)]$ la matriz de probabilidades de transición del proceso de Markov $X(t)$, donde:

$$P_{ij}(t) = P[\text{Estado } j \text{ en t} \,|\, \text{Estado } i \text{ en t-1}]$$

Es posible relacionar las tasas de transición con las probabilidades de transición por medio de un sistema de ecuaciones diferenciales hacia adelante de Kolmogorov, (Bhat, 1984; Lawless, 2002).

Sea la matriz $P$ de probabilidades de transición asociada al proceso de Markov, la cual está relacionada con la matriz de tasas de transiciones $Q$ a través de un sistema de ecuaciones diferenciales de Kolmogorov hacia delante; estas matrices son de orden $k\times k$, donde $k$ representa el número de estados. Tomando el modelo que se ilustra en la Figura 1 las ecuaciones de Kolmogorov resultantes y sus soluciones exactas son:

$$\frac{\partial}{\partial t}P(t) = P(t)\,Q \quad ; \quad P(0) = I_k \,\text{con}\, Q=[\lambda_{ij}] \tag{1}$$

La solución a este sistema de ecuaciones está dada por:

$$P_{11}(t) = \frac{1}{\lambda_{12}+\lambda_{21}}\left[\lambda_{21}+\lambda_{12}e^{-(\lambda_{12}+\lambda_{21})t}\right] \tag{2}$$

$$P_{12}(t) = \frac{\lambda_{12}}{\lambda_{12}+\lambda_{21}}\left[1-e^{-(\lambda_{12}+\lambda_{21})t}\right] \tag{3}$$

$$P_{21}(t) = \frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}}\left[1-e^{-(\lambda_{12}+\lambda_{21})t}\right] \tag{4}$$

$$P_{22}(t) = \frac{1}{\lambda_{12}+\lambda_{21}}\left[\lambda_{12}+\lambda_{21}e^{-(\lambda_{12}+\lambda_{21})t}\right] \tag{5}$$

## 2.1. Tasas de transición como función de covariables

Para el proceso de Markov descrito en la Figura 1, se propone una parametrización para las tasas de transición de acuerdo al modelo de Andersen et al. (1993), tal como lo sugieren Kay (1986), Harezlak *et al.* (2003) y Salazar *et al.* (2003). Dicha parametrización es de la forma: $\lambda_{ij}=\lambda^*_{ij}e^{\beta_{ij}^T\mathbf{X}}$, donde el vector $\beta_{ij}$ mide los efectos del vector de covariables $X$ sobre las tasas de transición del estado $i$ al estado $j$. Para el caso de una sola covariable, usando el modelo de Andersen et al. (1993), las tasas de transición se expresan como:

$$\lambda_{12} = \lambda^*_{12}e^{\beta_{12}X} \tag{6}$$

$$\lambda_{21} = \lambda^*_{21}e^{\beta_{21}X} \tag{7}$$

## 2.2. Función de verosimilitud

Suponga un proceso de Markov de primer orden con dos estados recurrentes $\{1,2\}$. Para un sujeto $w$ considere las siguientes notaciones:

$M_w$ :Número de observaciones para el sujeto $w$.

$\{t_0^{(w)}, t_1^{(w)}, \ldots, t_{M_w}^{(w)}\}$ : los tiempos en los cuales el sujeto $w$ es monitoreado.

$S_i^{(w)}$ : Estado observado para el sujeto $w$ en el tiempo $t_i$.

$P_{S_{i-1},S_i}^{(w)}(t_i^{(w)} - t_{i-1}^{(w)})$ : Probabilidad de transición para el sujeto $w$ del estado $S_{i-1}$ al estado $S_i$ en el intervalo de tiempo $(t_{i-1}, t_i)$.

$\mathbf{T}$ : Es el vector que contiene todos los tiempos de monitoreo de todos los $n$ sujetos.

La contribución del $w$-ésimo individuo a la verosimilitud está dada por:

$$\prod_{i=1}^{M_w} P_{S_{i-1},S_i}^{(w)}(t_i^{(w)} - t_{i-1}^{(w)}) \ .$$

Por lo tanto, la verosimilitud para $n$ sujetos está dada por:

$$\prod_{w=1}^{n}\prod_{i=1}^{M_w} P_{S_{i-1},S_i}^{(w)}(t_i^{(w)} - t_{i-1}^{(w)}) \ .$$

Haciendo: $\tilde{t}_i^{(w)} = t_i^{(w)} - t_{i-1}^{(w)}$ y $\boldsymbol{\theta} = (\lambda_{12}^*, \lambda_{21}^*, \beta_{12}, \beta_{21})$, la verosimilitud se expresa como:

$$L(\boldsymbol{\theta}\,|\mathbf{X}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} P_{S_{i-1},S_i}^{(w)}\left(\tilde{t}_i^{(w)}\right) \ .$$

Para obtener una expresión que involucre todos los parámetros asociados al vector $\boldsymbol{\theta}$, se definen las siguientes variables indicadoras (Correa *et al.*, 2010):

$$\delta_{jk}^{(w,i)} = \begin{cases} 1 & \text{Si el sujeto } w \text{ pasa del estado } j \text{ al } k \text{ en la } i\text{-ésima observación} \\ 0 & \text{En otro caso} \end{cases}$$

Así, entonces la función de verosimilitud se expresa:

$$\begin{aligned} L(\boldsymbol{\theta}\,|X\,,\mathbf{T}) &= \prod_{w=1}^{n}\prod_{i=1}^{M_w} P_{11}^{\delta_{11}^{(w,i)}}(\tilde{t}_i^{(w)})\, P_{12}^{\delta_{12}^{(w,i)}}(\tilde{t}_i^{(w)})\, P_{21}^{\delta_{21}^{(w,i)}}(\tilde{t}_i^{(w)})\, P_{22}^{\delta_{22}^{(w,i)}}(\tilde{t}_i^{(w)}) \\ &= \prod_{w=1}^{n}\prod_{i=1}^{M_w} \left\{ \frac{1}{\lambda_{12}+\lambda_{21}} \left[\lambda_{21} + \lambda_{12}\, e^{-(\lambda_{12}+\lambda_{21})\tilde{t}_i^{(w)}}\right] \right\}^{\delta_{11}^{(w,i)}} \\ &\times \left\{ \frac{\lambda_{12}}{\lambda_{12}+\lambda_{21}} \left[1 - e^{-(\lambda_{12}+\lambda_{21})\tilde{t}_i^{(w)}}\right] \right\}^{\delta_{12}^{(w,i)}} \\ &\times \left\{ \frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}} \left[1 - e^{-(\lambda_{12}+\lambda_{21})\tilde{t}_i^{(w)}}\right] \right\}^{\delta_{21}^{(w,i)}} \\ &\times \left\{ \frac{1}{\lambda_{12}+\lambda_{21}} \left[\lambda_{12} + \lambda_{21}\, e^{-(\lambda_{12}+\lambda_{21})\tilde{t}_i^{(w)}}\right] \right\}^{\delta_{22}^{(w,i)}} \end{aligned}$$

Simplificando:

$$L(\boldsymbol{\theta}\,|\,X\,,\mathbf{T}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w}\left[\frac{1}{\lambda_{12}+\lambda_{21}}\right]\lambda_{12}^{\delta_{12}^{(w,i)}}\,\lambda_{21}^{\delta_{21}^{(w,i)}}\left[1-e^{-(\lambda_{12}+\lambda_{21})\tilde{t}_i^{(w)}}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}$$
$$\times\left[\lambda_{21}+\lambda_{12}\,e^{-(\lambda_{12}+\lambda_{21})\tilde{t}_i^{(w)}}\right]^{\delta_{11}^{(w,i)}}\left[\lambda_{12}+\lambda_{21}\,e^{-(\lambda_{12}+\lambda_{21})\tilde{t}_i^{(w)}}\right]^{\delta_{22}^{(w,i)}}$$

Usando la expresión del modelo de Andersen et al. (1993) se obtiene:

$$\begin{aligned}L(\boldsymbol{\theta}\,|\,X\,,\mathbf{T}) = &\prod_{w=1}^{n}\prod_{i=1}^{M_w}\left[\frac{1}{\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}}\right]\\
&\times[\lambda_{12}^{*}]^{\delta_{12}^{(w,i)}}\,e^{\beta_{12}X_w\,\delta_{12}^{(w,i)}}\,[\lambda_{21}^{*}]^{\delta_{12}^{(w,i)}}\,e^{\beta_{21}X_w\,\delta_{21}^{(w,i)}}\\
&\times\left\{1-e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}\\
&\times\left\{\lambda_{21}^{*}\,e^{\beta_{21}X_w}+\lambda_{12}\,e^{\beta_{12}X_w}\,e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{11}^{(w,i)}}\\
&\times\left\{\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}\,e^{\beta_{21}X_w}\,e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{22}^{(w,i)}}\end{aligned}$$

Donde cada una de las probabilidades de transición están dadas por la solución del sistema de ecuaciones diferenciales hacia adelante de Kolmogorov (ecuaciones 2, 3, 4, 5) y las covariables se involucran con la parametrización de tipo Andersen et al. (1993) (ecuaciones 6 y 7). Las densidades completas se pueden ver en los anexos de la sección 6.

## 2.3. Estimación bayesiana de las tasas de transición

La metodología bayesiana es muy útil para obtener aproximaciones de parámetros de interés (Gordon, 2001; Hans & Dunson, 2005). El no depender de supuestos asintóticos en las soluciones que se obtienen es una de las ventajas de la metodología bayesiana y todo el trabajo inferencial se realiza usando la distribución aposteriori (o posterior). La vigencia y utilidad de estos métodos de predicción justifican su uso para obtener estimaciones de las tasas de transición en un modelo de dos estados recurrentes.

Para utilizar el muestreador de Gibbs de acuerdo a Tanner (1996) se encuentran las distribuciones aposterioris y por medio de éstas las densidades condicionales con la cuales se programa el método del muestreador de Gibbs.

## 2.4. Distribuciones posteriores

Por el teorema de Bayes se sabe que la distribución posterior es proporcional a:

$$\xi(\theta|X) \propto L(\theta|X)\,\xi(\theta)$$

donde:
$L(\theta|X)$: Es la verosimilitud de los datos, definida en la sección anterior.
$\xi(\theta)$: Es la distribución apriori para los $\lambda_{ij}$.

De acuerdo al principio de la razón insuficiente de Laplace (Gordon, 2001), si el espacio parametral es finito, se puede utilizar una distribución apriori uniforme para reflejar ignorancia total. Para observar el efecto en las predicciones de las tasas de transición se selecciona una distribución apriori no informativa de la forma $\xi(\theta) = 1$ y una distribución apriori informativa exponencial con parámetros $\tau$ para $\lambda_{12}$ y $\alpha$ para $\lambda_{21}$, sea $\lambda_{ij}$ independientes y $\tau$, $\alpha$ conocidos.

Por lo tanto las distribuciones posteriores están dadas por:

$$\xi(\theta|X) \propto L(\theta|X)\xi(\theta) = \left[\prod_{w=1}^{n}\prod_{i=1}^{M_w} p_{11}^{\delta_{11}^{(w,i)}}(\tilde{t}_i^{(w)})\, p_{12}^{\delta_{12}^{(w,i)}}(\tilde{t}_i^{(w)})\, p_{21}^{\delta_{21}^{(w,i)}}(\tilde{t}_i^{(w)})\, p_{22}^{\delta_{22}^{(w,i)}}(\tilde{t}_i^{(w)})\right] 1 \tag{8}$$

$$\xi(\theta|X) \propto L(\theta|X)\xi(\theta) = \left[\prod_{w=1}^{n}\prod_{i=1}^{M_w} p_{11}^{\delta_{11}^{(w,i)}}(\tilde{t}_i^{(w)})\, p_{12}^{\delta_{12}^{(w,i)}}(\tilde{t}_i^{(w)})\, p_{21}^{\delta_{21}^{(w,i)}}(\tilde{t}_i^{(w)})\, p_{22}^{\delta_{22}^{(w,i)}}(\tilde{t}_i^{(w)})\right] \tau e^{-\lambda_{12}^{*}\tau}\alpha e^{-\lambda_{21}^{*}\alpha}$$

Para realizar la estimación se deben escribir las probabilidades de transición en términos de las tasas de transición por medio de las ecuaciones de Kolmogorov y el modelo de Andersen et al. (1993) por medio del cual se mide el efecto de las covariables.

Así, la distribución posterior para el vector de parámetros $\theta$ con una apriori no informativa es de la forma:

$$\xi(\theta|X) \propto L(\theta|X) \times 1$$

Se tiene entonces:

$$\begin{aligned}
L(\boldsymbol{\theta}\,|\,X\,,\mathbf{T}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} & \left[\frac{1}{\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}}\right] \\
& \times [\lambda_{12}^{*}]^{\delta_{12}^{(w,i)}}\, e^{\beta_{12}X_w\,\delta_{12}^{(w,i)}}\, [\lambda_{21}^{*}]^{\delta_{12}^{(w,i)}}\, e^{\beta_{21}X_w\,\delta_{21}^{(w,i)}} \\
& \times \left\{1 - e^{-\left[\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{12}^{(w,i)} + \delta_{21}^{(w,i)}} \\
& \times \left\{\lambda_{21}^{*}\, e^{\beta_{21}X_w} + \lambda_{12}\, e^{\beta_{12}X_w}\, e^{-\left[\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{11}^{(w,i)}} \\
& \times \left\{\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}\, e^{\beta_{21}X_w}\, e^{-\left[\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{22}^{(w,i)}}
\end{aligned}$$

Haciendo:

$$\begin{aligned}
A_w &= \left[\frac{1}{\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}}\right] \\
B_w^{(i)} &= \lambda_{21}^{*}\, e^{\beta_{21}X_w} + \lambda_{12}\, e^{\beta_{12}X_w}\, e^{-\left[\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}} \\
C_w^{(i)} &= \lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}\, e^{\beta_{21}X_w}\, e^{-\left[\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}} \\
D_w^{(i)} &= \left\{1 - e^{-\left[\lambda_{12}^{*}\, e^{\beta_{12}X_w} + \lambda_{21}^{*}\, e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}
\end{aligned}$$

Se obtienen las condicionales completas:

$$L(\lambda_{12}^{*}\,|\,\lambda_{21}^{*},\beta_{12},\beta_{21}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,[\lambda_{12}^{*}]^{\delta_{12}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}$$

$$L(\lambda_{21}^{*}\,|\,\lambda_{12}^{*},\beta_{12},\beta_{21}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,[\lambda_{21}^{*}]^{\delta_{21}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}$$

$$L(\beta_{12}\,|\,\lambda_{12}^{*},\lambda_{21}^{*},\beta_{21}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,e^{\beta_{12}X_w\delta_{12}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}$$

$$L(\beta_{21}\,|\,\lambda_{12}^{*},\lambda_{21}^{*},\beta_{12}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,e^{\beta_{21}X_w\delta_{21}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}$$

Y la distribución posterior para el vector de parámetros $\theta$ con una apriori informativa es de la forma:

$$\xi(\theta|X) \propto L(\theta|X) \times \tau e^{-\lambda_{12}^{*}\tau}\alpha e^{-\lambda_{21}^{*}\alpha}$$

Con $\tau$ y $\alpha$ conocidos, se tiene entonces:

$$\begin{aligned}
L(\boldsymbol{\theta}\,|\,X\,,\mathbf{T}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} & \left[\frac{1}{\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}}\right] \\
& \times [\lambda_{12}^{*}]^{\delta_{12}^{(w,i)}}\,e^{\beta_{12}X_w\,\delta_{12}^{(w,i)}}\,[\lambda_{21}^{*}]^{\delta_{12}^{(w,i)}}\,e^{\beta_{21}X_w\,\delta_{21}^{(w,i)}} \\
& \times \left\{1-e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}} \\
& \times \left\{\lambda_{21}^{*}\,e^{\beta_{21}X_w}+\lambda_{12}\,e^{\beta_{12}X_w}\,e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{11}^{(w,i)}} \\
& \times \left\{\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}\,e^{\beta_{21}X_w}\,e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}^{\delta_{22}^{(w,i)}} \\
& \times \tau e^{-\lambda_{12}^{*}\tau}\alpha e^{-\lambda_{21}^{*}\alpha}
\end{aligned}$$

Haciendo:

$$\begin{aligned}
A_w &= \left[\frac{1}{\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}}\right] \\
B_w^{(i)} &= \lambda_{21}^{*}\,e^{\beta_{21}X_w}+\lambda_{12}\,e^{\beta_{12}X_w}\,e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}} \\
C_w^{(i)} &= \lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}\,e^{\beta_{21}X_w}\,e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}} \\
D_w^{(i)} &= \left\{1-e^{-\left[\lambda_{12}^{*}\,e^{\beta_{12}X_w}+\lambda_{21}^{*}\,e^{\beta_{21}X_w}\right]\tilde{t}_i^{(w)}}\right\}
\end{aligned}$$

Se obtienen las condicionales completas:

$$L(\lambda_{12}^{*}\,|\,\lambda_{21}^{*},\beta_{12},\beta_{21}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,[\lambda_{12}^{*}]^{\delta_{12}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}\,\tau e^{-\lambda_{12}^{*}\tau}$$

$$L(\lambda_{21}^{*}\,|\,\lambda_{12}^{*},\beta_{12},\beta_{21}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,[\lambda_{21}^{*}]^{\delta_{21}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}\,\alpha e^{-\lambda_{21}^{*}\alpha}$$

$$L(\beta_{12}\,|\,\lambda_{12}^{*},\lambda_{21}^{*},\beta_{21}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,e^{\beta_{12}X_w\delta_{12}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}$$

$$L(\beta_{21}\,|\,\lambda_{12}^{*},\lambda_{21}^{*},\beta_{12}) = \prod_{w=1}^{n}\prod_{i=1}^{M_w} A_w\,e^{\beta_{21}X_w\delta_{21}^{(w,i)}}\,\left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}}\,\left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}}\,\left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)}+\delta_{21}^{(w,i)}}$$

Así, con las densidades condicionales completas encontradas para cada distribución apriori se implementa el método de estimación por medio del muestreador de Gibbs.

# 3. ESTUDIO DE SIMULACIÓN

Basados en el esquema de muestreo de Correa *et al.* (2010) se asume un proceso estocástico de Markov de dos estados que se denotarán 1 y 2 donde los estados son recurrentes; por tanto las transiciones son 1 a 1, 1 a 2, 2 a 1 y 2 a 2. Para el muestreador de Gibbs se asume una distribución apriori no informativa (Laplace igual 1) y una distribución apriori informativa (Exponencial con parámetros $\tau = 40$ para $\lambda_{12}$ y $\alpha = 91$ para $\lambda_{21}$ con $\lambda_{ij}$ independientes). Todas las simulaciones se llevaron a cabo usando el sotfware SAS con el procedimiento IML del SAS (2004).

Para ejecutar las simulaciones se tomaron las condiciones descritas a continuación. Primero, se simularon 1000 muestras de tamaños $n =$ 50, 100, 200, 400 unidades que contenían historias aleatorias de transiciones en el modelo de dos estados para los $n$ sujetos simulados; luego para cada tamaño muestral se generaron un máximo de 3 y 4 medidas repetidas por unidad. La variable edad del equipo de cómputo se incorpora en el modelo con tres categorías: 1: $\leq$ 72 semanas, 2: 72 - 117 semanas y 3: $>$ 117 semanas. La parametrización utilizada para las tasas de transición está basada en el modelo multiplicativo de Andersen et al. (1993), de la forma:

$$\lambda_{ij} = \lambda^*_{ij} e^{\beta_{ij} X} \tag{9}$$

donde X = (*gedad*). Los valores de los $\lambda^*_{ij}$ para cada grupo de edad son tomados de los reportados en Correa *et al.* (2010), las parametrizaciones usadas para obtener los valores de referencia fueron:

$$\lambda_{12} = 0.025 * e^{0.1127 * gedad} \tag{10}$$

$$\lambda_{21} = 0.011 * e^{0.1074 * gedad}$$

## 3.1. Resultados numéricos de las tasas de transición simuladas

En cada simulación se evaluó la distribución posterior y se calcularon las respectivas tasas de transición descritas por el modelo de Andersen et al. (1993) para cada valor de la covariable de interés. Luego, se tomó una muestra aleatoria con reemplazo de tamaño 1000 que contenía, además de los valores de cada una de las tasas de transición, los valores de probabilidad especificados por la distribución posterior. Usando cada una de estas muestras se calculó el promedio aritmético, el cual se trató como el respectivo estimador de las tasas de intensidad asociadas al modelo de dos estados recurrentes.

En las Tablas 1 y 2 se puede observar que independiente del número de escaneos, cuando el grupo de edad del equipo de cómputo es menor de 72 semanas las tasas de intensidad $\lambda_{21}$ están sobre estimadas respecto a los valores de referencia siendo menos sobrestimadas las predicciones de la distribución exponencial. Ahora las predicciones para $\lambda_{12}$ en este grupo de edad se muestran muy cercanas a los valores de referencia

Tabla 1: Tasas de Intensidad. Número de observaciones repetidas = 3

| | Descripción | Valores Covariable | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | | 2 | | 3 | |
| | Distribución Apriori | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ |
| 50 | Referencia | 0.028 | 0.012 | 0.031 | 0.014 | 0.035 | 0.015 |
| | Laplace | 0.029 | 0.036 | 0.032 | 0.041 | 0.036 | 0.046 |
| | Exponencial | 0.028 | 0.021 | 0.030 | 0,022 | 0.032 | 0.024 |
| 100 | Referencia | 0.028 | 0.012 | 0.031 | 0.014 | 0.035 | 0.015 |
| | Laplace | 0.027 | 0.022 | 0.029 | 0.024 | 0.030 | 0.025 |
| | Exponencial | 0,027 | 0.033 | 0.035 | 0.035 | 0.037 | 0.037 |
| 200 | Referencia | 0.028 | 0.012 | 0.031 | 0.014 | 0.035 | 0.015 |
| | Laplace | 0.028 | 0.028 | 0.031 | 0.032 | 0.035 | 0.035 |
| | Exponencial | 0.027 | 0.024 | 0.028 | 0.025 | 0.030 | 0.027 |
| 400 | Referencia | 0.028 | 0.012 | 0.031 | 0.014 | 0.035 | 0.015 |
| | Laplace | 0.027 | 0.027 | 0.030 | 0.030 | 0.034 | 0.034 |
| | Exponencial | 0.028 | 0.026 | 0.029 | 0.028 | 0.031 | 0.029 |

Tabla 2: Tasas de Intensidad. Número de observaciones repetidas = 4

| | Descripción | Valores Covariable | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | | 2 | | 3 | |
| | Distribución Apriori | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ |
| 50 | Referencia | 0.028 | 0.012 | 0.031 | 0.014 | 0.035 | 0.015 |
| | Laplace | 0.030 | 0.161 | 0.033 | 0.181 | 0.038 | 0.204 |
| | Exponencial | 0.028 | 0.022 | 0.029 | 0.023 | 0.031 | 0.024 |
| 100 | Referencia | 0.028 | 0.012 | 0.031 | 0.014 | 0.035 | 0.015 |
| | Laplace | 0.028 | 0.030 | 0.032 | 0.033 | 0.036 | 0.037 |
| | Exponencial | 0.027 | 0.023 | 0.029 | 0.025 | 0.030 | 0.026 |
| 200 | Referencia | 0.028 | 0.012 | 0.031 | 0.013 | 0.035 | 0.015 |
| | Laplace | 0.028 | 0.028 | 0.031 | 0.031 | 0.034 | 0.035 |
| | Exponencial | 0.028 | 0.025 | 0.029 | 0.027 | 0.031 | 0.028 |
| 400 | Referencia | 0.028 | 0.012 | 0.031 | 0.014 | 0.035 | 0.015 |
| | Laplace | 0.027 | 0.027 | 0.030 | 0.030 | 0.034 | 0.034 |
| | Exponencial | 0.027 | 0,026 | 0.029 | 0.028 | 0.031 | 0.029 |

especialmente para el tamaño de muestra igual a 100.

Para el grupo de edad del equipo de cómputo entre 72 y 117 semanas y la distribución Laplace, se observa que los valores de referencia de $\lambda_{12}$ se sobre estiman para el tamaño de muestra de 50, mientras que para los tamaños de muestra 100, 200 y 400 el valor predicho está subestimado, con más cercanía al valor de referencia la predicción cuando el tamaño de muestra es igual a 100. Ahora para la distribución exponencial se tiene que las predicciones $\lambda_{12}$ están subestimadas respecto a los valores de referencia excepto para el tamaño de muestra igual a 100. Para $\lambda_{21}$ en este grupo de edad las predicciones, con ambas distribuciones, se encuentra sobre estimadas pero menos con la distribución exponencial.

En el grupo de edad de mayores a 117 semanas con la distribución Laplace, se tiene que la tasa de intensidad estimada para $\lambda_{12}$ con el tamaño de muestra igual a 50 está muy cercano al valor de referencia, cuando el tamaño de muestra es de 100, 200 y 400 las predicciones para $\lambda_{12}$ son subestimadas.

Para $\lambda_{21}$ se puede observar que independiente del grupo de edad y la distribución apriori utilizada se sobre estiman las predicciones respecto a los valores de referencia. Es de notar que las predicciones de $\lambda_{21}$ realizadas con una distribución exponencial de parámetro 91 son menos sobre estimadas que las encontradas con una distribución Laplace. Además, se puede observar que a medida que el grupo de edad aumenta se observa un cambio en la predicción de la tasas, por tanto se muestra un efecto de la covariable en la obtención de las mismas.

### 3.2. Resultados distribucionales de las tasas de transición simuladas

En la Figura 2 se tiene que la distribución de las estimaciones para $\lambda_{12}$ es similar a la distribución de las estimaciones para $\lambda_{21}$ con una distribución apriori no informativa (Laplace). Se observa, también, que existe una diferencia por grupo de edad lo cual indica un efecto de la covariable en la estimación a nivel distribucional, note que en el grupo de edad 1 las distribución de las predicciones para $\lambda_{12}$ con distribución apriori exponencial con parámetro 40 es similar a las obtenidas con la distribución apriori no informativa, pero a medida que cambia el grupo de ésta esa similitud va cambiando.

Para las predicciones de $\lambda_{21}$ con un distribución apriori exponencial con parámetro 91 se tiene un comportamiento distribucional con valores más altos que los otros valores, además es un poco sesgada a la derecha, mientras que para las predicciones de $\lambda_{12}$ con una distribución apriori exponencial de parámetro 40 se tiene un comportamiento distribucional similar al obtenido con la distribución no informativa. Es importante destacar que en este caso también se observa el efecto de covariable edad pues a medida que el grupo de edad del equipo de cómputo aumenta los valores de las predicciones van disminuyendo.

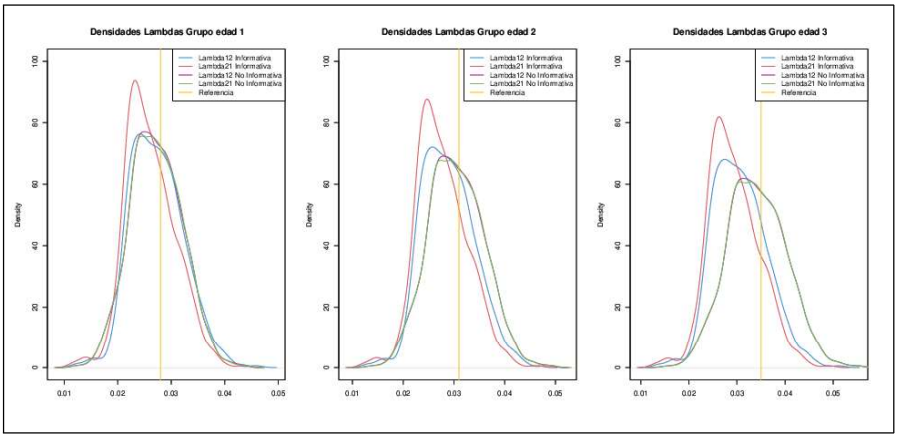


Figura 2: Densidades Tasas de transición. Fuente: Elaboración Propia

En las gráficas de la Figura 2 también se puede observar que las estimaciones de las tasas con una distribución no informativa tienen un comportamiento distribucional platicúrtico con colas no pesadas, en cuanto a la estimación de las tasas con una distribución informativa tienen comportamiento distribucional leptocúrtico. Gráficamente, se muestra una diferencia entre las predicciones con la distribución no informativa y la distribución informativa, por tanto se observa un efecto de la distribución apriori en las predicciones de las tasas.

En las Figuras 3 y 4 se puede observar que las cadenas simuladas se encuentran alrededor de una media y se ven estable a través del tiempo.

Para las predicciones de $\lambda_{12}$ con tres escaneos se nota que la distribución no informativa es la que tiene más cercanía con los valores de referencia mientras que la distribución exponencial con parámetro 40, en casi todos los grupos de edad, está por debajo de los valores de referencia y las predicciones de la distribución apriori no informativa (Figura 5). Para el caso de las predicciones de $\lambda_{21}$ se puede observar que para el grupo de edad 1 las predicciones con una apriori no informativa están muy por encima de los valores de referencia y los encontrados con distribución apriori exponencial con parámetro 91, este comportamiento es idéntico en los otros grupos de edades del equipo de cómputo pero es más evidente en el primer grupo (Figura 6). Este comportamiento no tiene un cambio muy significativo cuando se aumenta el número de escaneos a cuatro (Figuras 7 y 8).

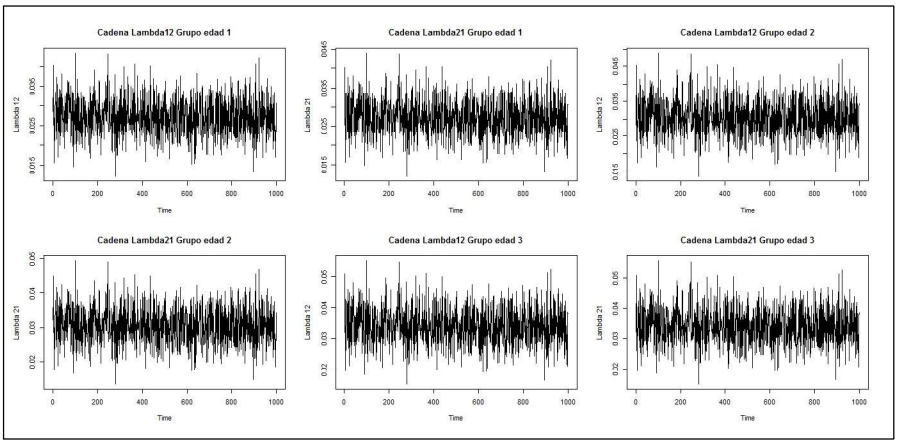


Figura 3: Cadenas de Markov apriori informativa Tamaño= 400 y medidas repetidas= 4. Fuente: Elaboración Propia

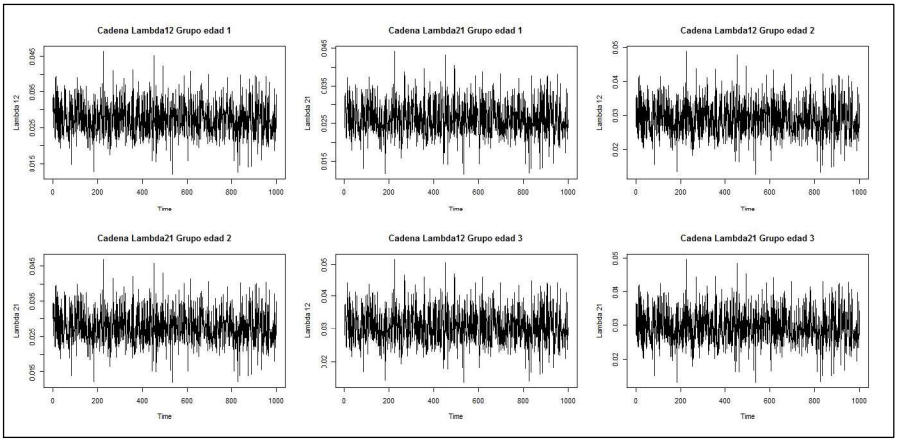


Figura 4: Cadenas de Markov apriori no informativa Tamaño= 400 y medidas repetidas = 4. Fuente: Elaboración Propia

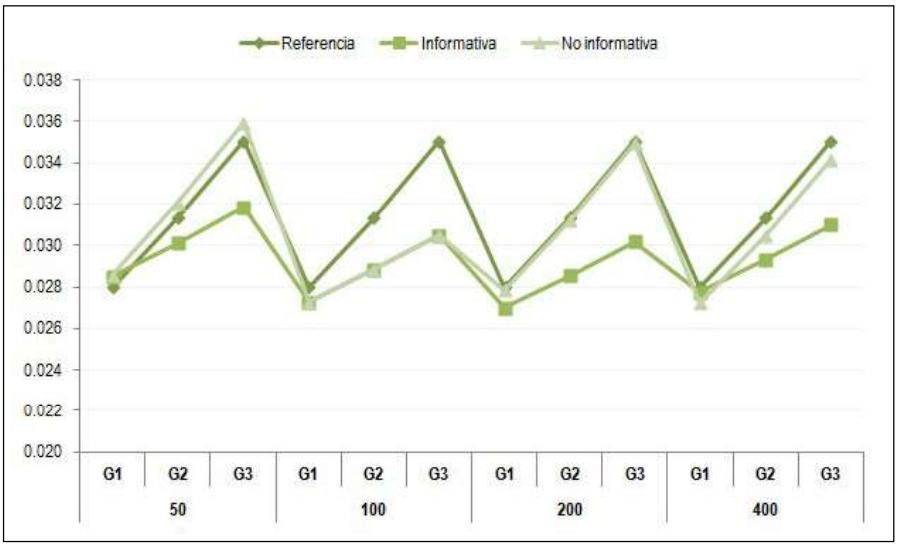


Figura 5: Estimación $\lambda_{12}$ con 3 escaneos. Fuente: Elaboración Propia

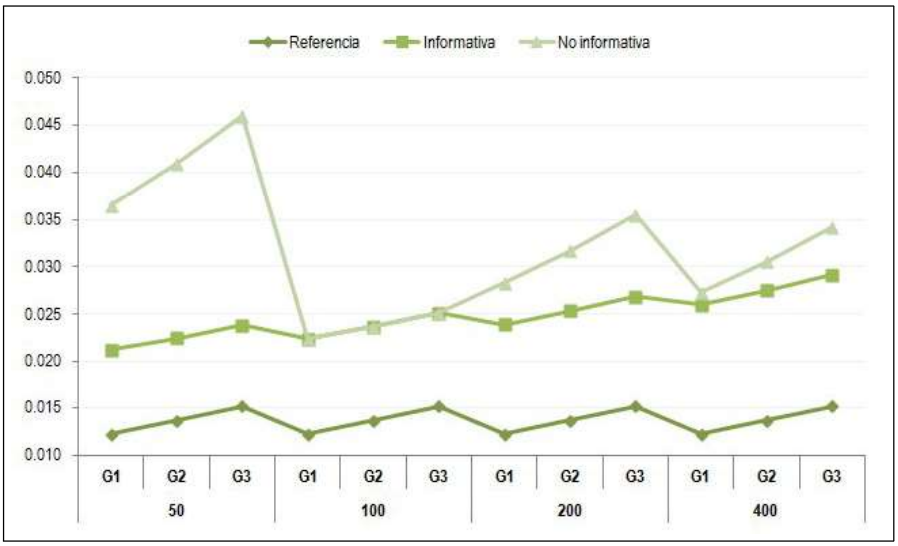


Figura 6: Estimación $\lambda_{21}$ con 3 escaneos. Fuente: Elaboración Propia

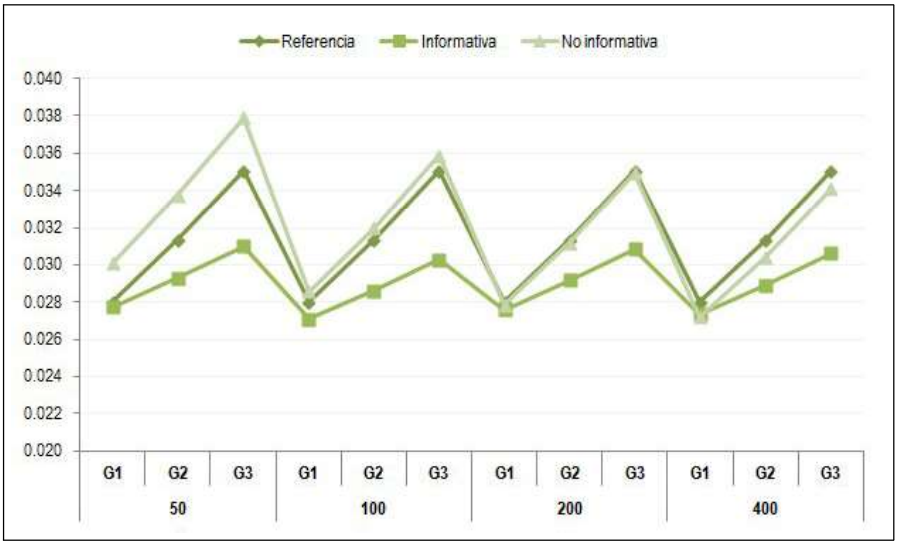


Figura 7: Estimación $\lambda_{12}$ con 4 escaneos. Fuente: Elaboración Propia

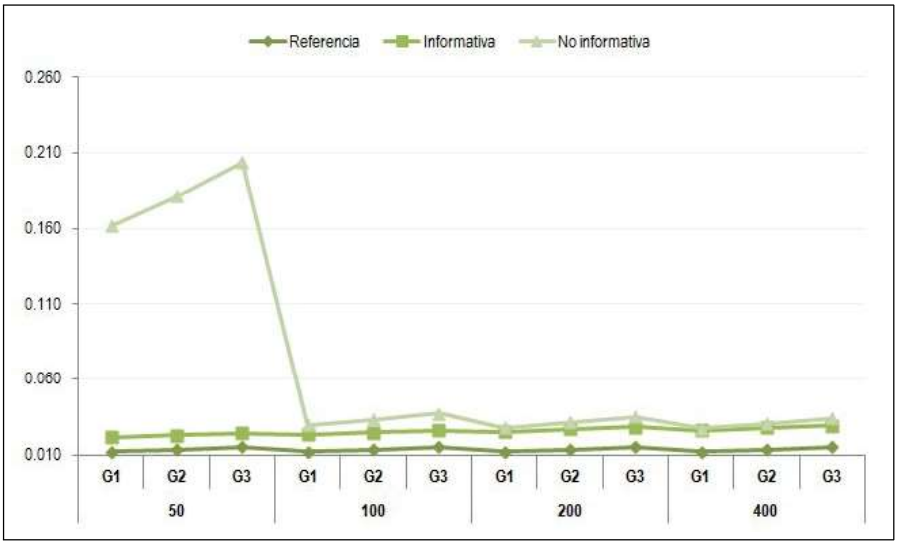


Figura 8: Estimación $\lambda_{21}$ con 4 escaneos. Fuente: Elaboración Propia

# 4. APLICACIÓN A DATOS REALES

Para ilustrar el modelo aquí expuesto se tomaron los datos recolectados en Valencia & Salazar (2012) de una entidad bancaria, estos provienen de una muestra aleatoria de 274 computadores, se tomó información en un periodo de 11 semanas, cada vez que se le registró información a un computador específico, se verificó si éste estaba en uno de estos dos posibles estados recurrentes: 1-sano ó 2-infectado.

La información recolectada está relacionada con: Edad de la máquina en meses, marca, clase, si tiene habilitado el puerto USB, cantidad de páginas web visitadas durante el período de observación, número de procesadores, tiempo de navegación y tipo de reloj del procesador. Para el modelo a evaluar sólo se tomó en consideración la variable clase, la cual es cualitativa con los siguientes valores: Portátil, CPU, Servidor. En la estructura de los datos utilizados, se tiene: la columna MAQUINA que identifica el código del equipo de cómputo observado, SEMANA que corresponde a la semana de observación del equipo de cómputo, la columna EP que indica el estado previo del equipo de cómputo, la columna EA que indica el estado actual del equipo, la columna DURACIÓN hace referencia al tiempo en semanas que transcurre entre cada observación y finalmente la columna CLASE que hace referencia al tipo de equipo de cómputo.
La matriz de transición cruda obtenida a partir de los datos, está dada por Frecuencia (Probabilidad de transición)), ver Tabla 3. La frecuencia, se refiere al número de cambios de estado observados en todos los computadores del SI (1: sano, 2: infectado. Por ejemplo, dado que un computador está infectado, la probabilidad de que pase a estar sano es 0.127.

Tabla 3: Matriz de transición cruda

| | 1 | 2 |
|---|---|---|
| 1 | 2.219 (0.873) | 322 (0.127) |
| 2 | 284 (0.600) | 189 (0.400) |

Se considera un modelo de dos estados recurrentes, donde las tasas de transición serán funciones de la covariable "Clase". Si $X$ denota la variable Clase, el modelo multiplicativo de Andersen et al. (1993) es de la forma:

$$\lambda_{ij} = \lambda^*_{ij}\, e^{\beta_{ij},X} \quad ; \quad i,j = 1,2\ .$$

Para poder incorporar en el modelo esta covariable, es necesario expresarla en una escala cuantitativa, por ejemplo 1 : CPU, 2: Portátil y 3: Servidor. El problema es que esta transformación no es adecuada ya que las nuevas categorías numéricas implican una jerarquía, mientras la variable original no es ordinal, y por lo tanto no es adecuado usar una escala numérica en reemplazo de una variable que no es ordinal. Como el interés está en evaluar el efecto de la covariable sobre la obtención de las tasas de intensidad de transición, se propone un modelo donde se tengan en cuenta las categorías de la variable clase.

Para ello defina las siguientes variables indicadoras:

$$X_1 = \begin{cases} 1; & \text{si tipo de computador es CPU} \\ 0; & \text{en otro caso} \end{cases}$$

$$X_2 = \begin{cases} 1; & \text{si tipo de computador es Portátil} \\ 0; & \text{en otro caso} \end{cases}$$

De esta manera, las tasas de intensidad de transición se expresan como:

$$\lambda_{ij} = \lambda^*_{ij}\, e^{\alpha_{ij}+\beta_{ij},X_1+\Gamma_{ij},X_2}; \qquad i,\ j = 1,\ 2. \tag{11}$$

El efecto asociado al computador tipo Servidor, se obtiene cuando $X_1 = 0$ y $X_2 = 0$.

Las soluciones del sistema de ecuaciones hacia adelante de Kolmogorov para el modelo de dos estados y la parametrización dada en la ecuación $(10)$, son usadas para construir la función de verosimilitud para el vector de parámetros $\boldsymbol{\theta}$. Para este caso particular, dicho vector está dado por:

$$\boldsymbol{\theta} = (\alpha_{12}, \lambda^*_{12}, \beta_{12}, \Gamma_{12}, \alpha_{21}, \lambda^*_{21}, \beta_{21}, \Gamma_{21})\ .$$

La estimación de estos parámetros se hará usando el muestreador de Gibbs. Se consideran dos situaciones: aprioris uniformes y aprioris exponenciales para $\lambda^*_{12}$ y para $\lambda^*_{21}$.

Para obtener los valores iniciales para $\lambda_{12}^{*}$ y para $\lambda_{21}^{*}$, se usan las ecuaciones descritas en sección 2.

$$\lambda_{ij}^{*(0)} = \frac{m_{ij}}{T_{ij}}, \ \beta_{ij}^{*(0)} \tag{12}$$

donde:
$m_{ij}$: representa el total de transiciones del estado i al estado j ,
$T_i$: representa el tiempo total en el estado i para todos los individuos,
$\beta_{ij}^{(0)}$: valores iniciales dados por el investigador.

Encontradas las ecuaciones para el vector de parámetros $\boldsymbol{\theta}$ se obtienen las predicciones para estos 8 parámetros con las cuales se calculan las tasas de intensidad de transición para $\lambda_{12}$ y $\lambda_{21}$ y las respectivas probabilidades de transición. Dado que la información fue recolectada en semanas, se muestra las probabilidades calculadas en un tiempo de una semana (Tabla 6) y dos semanas (Tabla 7) para cada una de las distribuciones aprioris.

De las Tablas 4 y 5 se puede observar que la probabilidades más altas en los tres tipos de computadores es la de estar en un estado sano, lo cual tiene sentido debido a los mantenimientos preventidos de los equipos de cómputo, los cuales están basados en mantener actualizado y tener corridas continuas de los sistemas antivirus lo cual influye en la probabilidad de mantenerse en este estado. Puede notarse también que a medida que el tiempo aumenta la probabilidad de quedarse en un estado de Infección disminuye lo cual también es explicado por el procedimiento realizado cuando se detecta un virus.

Cabe notar que el tipo de computadores más vulnerable a los ataques de virus son los servidores debido a que presentan la probabilidad más alta de transitar de un estado sano a un estado infectado, teniendo en cuenta que un servidor contiene información importante y de carácter confidencial de las compañías tiene sentido que sean los más atacados.

Tambien se puede ver que las probabilidades de pasar de un estado infectado a un estado sano son relevantes y mayores respecto a la probabilidad de pasar de un estado sano a infectado en los tres tipos de computadores esto se debe a que una de las formas de corregir este estado, una vez detectado el virus, es aplicar el antivirus influenciando así que los equipos tengan una mayor probabilidad de pasar de un estado de infección a un estado sano.

A nivel distribucional (Figura 9) y numérico no se observan diferencias significativas en las predicciones de las tasas de intensidad con las diferentes distribuciones aprioris escogidas para este estudio, cabe notar que en las densidades se percibe el impacto de aplicar un antivirus en la estimación de las tasas.

Tabla 4: Tasas de Intensidad

| Descripción | Valores Covariable | | | | | |
|---|---|---|---|---|---|---|
| | Portátil | | CPU | | Servidor | |
| Distribución Apriori | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ |
| Laplace | 0.196 | 0.209 | 0.210 | 0.789 | 1.012 | 1.031 |
| Exponencial | 0.196 | 0.209 | 0.210 | 0.788 | 1.011 | 1.031 |

Tabla 5: Intervalos de Credibilidad

| Descripción | Valores Covariable | | | | | |
|---|---|---|---|---|---|---|
| | Portátil | | CPU | | Servidor | |
| Distribución Apriori | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ | $\lambda_{12}$ | $\lambda_{21}$ |
| Laplace | (0.180, 0.202) | (0.191, 0.223) | (0.224, 0.210) | (0.720, 0.847) | (0.900, 1.128) | (0.928, 1.141) |
| Exponencial | (0.180, 0.202) | (0.189, 0.223) | (0.190, 0.224) | (0.720, 0.847) | (0.892, 1.128) | (0.919, 1.152) |

Tabla 6: Probabilidades de Transición con t = 1 Semana

| Descripción | Valores Covariable | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Portátil | | | | CPU | | | | Servidor | | | |
| Distribución Apriori | $P_{11}$ | $P_{12}$ | $P_{21}$ | $P_{22}$ | $P_{11}$ | $P_{12}$ | $P_{21}$ | $P_{22}$ | $P_{11}$ | $P_{12}$ | $P_{21}$ | $P_{22}$ |
| Laplace | 0.838 | 0.161 | 0.172 | 0.828 | 0.867 | 0.133 | 0.499 | 0.501 | 0.569 | 0.431 | 0.439 | 0.561 |
| Exponencial | 0.839 | 0.161 | 0.172 | 0.828 | 0.867 | 0.133 | 0.499 | 0.501 | 0.569 | 0.431 | 0.439 | 0.561 |

Tabla 7: Probabilidades de Transición con t = 2 Semanas

| Descripción | Valores Covariable | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Portátil | | | | CPU | | | | Servidor | | | |
| Distribución Apriori | $P_{11}$ | $P_{12}$ | $P_{21}$ | $P_{22}$ | $P_{11}$ | $P_{12}$ | $P_{21}$ | $P_{22}$ | $P_{11}$ | $P_{12}$ | $P_{21}$ | $P_{22}$ |
| Laplace | 0.730 | 0.269 | 0.287 | 0.713 | 0.818 | 0.182 | 0.682 | 0.318 | 0.513 | 0.487 | 0.496 | 0.504 |
| Exponencial | 0.731 | 0.269 | 0.287 | 0.713 | 0.818 | 0.182 | 0.682 | 0.318 | 0.513 | 0.487 | 0.496 | 0.504 |

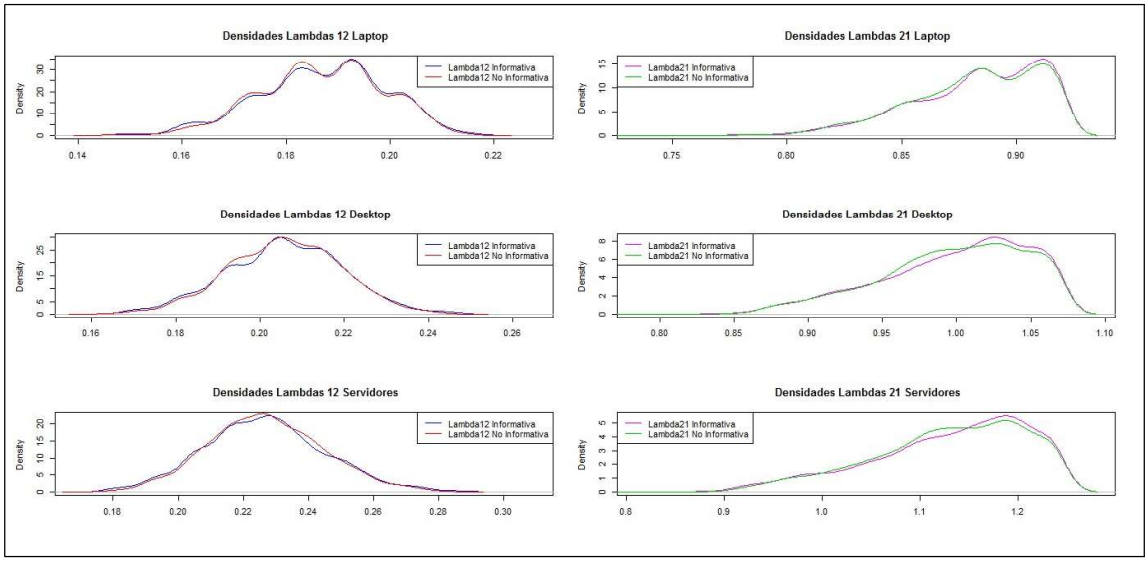


Figura 9: Densidades Tasas de transición distribuciones aprioris. Fuente: Elaboración Propia

# 5. CONCLUSIONES

## 5.1. Respecto al estudio de simulación

Por medio de un estudio de simulación se encontraron las predicciones para cada distribución apriori con diferentes números de individuos y diferentes números de escaneos, el cual se compara con los valores de referencia encontrados, de esto podemos observar que el comportamiento de las predicciones de $\lambda_{21}$ para los individuos más jóvenes pertenecientes al grupo 1 de edad son sobreestimadas respecto a los valores de referencia pero a medida que incrementa el tamaño de muestra el diferencial con valores de referencia disminuye independiente del número de medidas repetidas, para las predicciones de $\lambda_{12}$ en este grupo de edad se observan que para un tamaño de muesta igual a 50 se sobreestiman pero a medida que el tamaño de muestra aumenta la diferencia con los valores de referencia es mínima.

Es de notar que la covariable muestra un impacto en la estimación puesto que a medida que el grupo de edad aumenta los valores de referencia son mayores y las estimaciones también.

Respecto al comportamiento de las predicciones con las distribuciones de Laplace (Distribución no Informativa) y Exponencial con parámetros igual a 40 y 91 para $\lambda_{12}$ y $\lambda_{21}$ respectivamente (Distribución Informativa), las cuales fueron las escogidas como aprioris, se observan comportamientos diferentes en los tres grupos de edad y también a nivel distribucional principalmente en los picos figura 2 en los 3 grupos de edad, adicional se puede notar que el tamaño de muestra igual a 100 es el que mejor comportamiento presenta respecto a los valores de referencia, note también que los promedios estimados de las tasas de intensidad con una distribución apriori no informativa para $\lambda_{12}$ son más cercanos a los valores de referencia, mientras que para $\lambda_{21}$ las dos distribuciones aprioris son sobreestimadas pero la distribución a priori informativa es la que más cerca se encuentra a los valores de referencia (Figura 5 a 8). Por tanto se percibe un efecto en la estimación de las tasas de intensidad con las diferentes distribuciones aprioris estudiadas, por tanto podemos notar que para $\lambda_{12}$ funciona mejor una distribución Laplace mientras que para $\lambda_{21}$ ninguna de las dos distribuciones muestra una buen resultado.

## 5.2. Respecto a la aplicación a datos reales

En cuanto a la aplicación a datos reales tenemos que los equipo de computo más vulnerables son servidores por tanto se puede recomendar intensificar los escaneos semanales con el fin de reducir la probabilidad de transición de un estado sano a infectado, protegiendo así uno de los activos más importantes para la entidad, la información confidencial.

Por otro lado notamos que las distribuciones aprioris contempladas en este trabajo no tienen un impacto significativo en la estimación de las tasas de transición y por ende en la probabilidad de transición en estados recurrentes, es de notar que si se percibe un impacto en la estimación de $\lambda_{21}$ debido al tratamiento

aplicado cuando se detecta un virus en el sistema, lo que implica una probabilidad mayor de transitar de un estado infectado a sano en cualquier tipo computador.

Pese a que al introducir más covariables al modelo se espera una mejor estimación de las tasas de intensidad esto a su vez complejiza el modelo lo cual dificulta la estimación de las tasas e implica un proceso de simulación y computacional arduo que impacta los tiempos haciéndolos mas extenso en estos procesos. Sin embargo, para variables categóricas recolectadas longitudinalmente, este tipo de modelamiento muestra ser efectivo siempre y cuando se tengan suficientes datos y número de medidas repetidas apropiados.

# 6. ANEXOS

## 6.1. Densidades Condicionales

Partiendo de la verosimilitud encontrada y haciendo:

$$
\begin{aligned}
A_w &= \left[ \frac{1}{\lambda_{12}^* e^{\beta_{12} X_w} + \lambda_{21}^* e^{\beta_{21} X_w}} \right] \\
B_w^{(i)} &= \lambda_{21}^* e^{\beta_{21} X_w} + \lambda_{12} e^{\beta_{12} X_w} e^{-\left[\lambda_{12}^* e^{\beta_{12} X_w} + \lambda_{21}^* e^{\beta_{21} X_w}\right] \tilde{t}_i^{(w)}} \\
C_w^{(i)} &= \lambda_{12}^* e^{\beta_{12} X_w} + \lambda_{21} e^{\beta_{21} X_w} e^{-\left[\lambda_{12}^* e^{\beta_{12} X_w} + \lambda_{21}^* e^{\beta_{21} X_w}\right] \tilde{t}_i^{(w)}} \\
D_w^{(i)} &= \left\{ 1 - e^{-\left[\lambda_{12}^* e^{\beta_{12} X_w} + \lambda_{21}^* e^{\beta_{21} X_w}\right] \tilde{t}_i^{(w)}} \right\}
\end{aligned}
$$

Se obtienen las condicionales completas:

$$
\begin{aligned}
L(\lambda_{12}^* \,|\, \lambda_{21}^*, \beta_{12}, \beta_{21}) &= \prod_{w=1}^{n} \prod_{i=1}^{M_w} A_w \, [\lambda_{12}^*]^{\delta_{12}^{(w,i)}} \left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}} \left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}} \left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)} + \delta_{21}^{(w,i)}} \\
L(\lambda_{21}^* \,|\, \lambda_{12}^*, \beta_{12}, \beta_{21}) &= \prod_{w=1}^{n} \prod_{i=1}^{M_w} A_w \, [\lambda_{21}^*]^{\delta_{21}^{(w,i)}} \left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}} \left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}} \left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)} + \delta_{21}^{(w,i)}} \\
L(\beta_{12} \,|\, \lambda_{12}^*, \lambda_{21}^*, \beta_{21}) &= \prod_{w=1}^{n} \prod_{i=1}^{M_w} A_w e^{\beta_{12} X_w \delta_{12}^{(w,i)}} \left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}} \left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}} \left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)} + \delta_{21}^{(w,i)}} \\
L(\beta_{21} \,|\, \lambda_{12}^*, \lambda_{21}^*, \beta_{12}) &= \prod_{w=1}^{n} \prod_{i=1}^{M_w} A_w e^{\beta_{21} X_w \delta_{21}^{(w,i)}} \left[B_w^{(i)}\right]^{\delta_{11}^{(w,i)}} \left[C_w^{(i)}\right]^{\delta_{22}^{(w,i)}} \left[D_w^{(i)}\right]^{\delta_{12}^{(w,i)} + \delta_{21}^{(w,i)}}
\end{aligned}
$$

# Referencias

Andersen, P. & Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, **10**, 1100-1120.

Andersen, Pk., Borgan, Gill, R. D. & Keiding, N. (1993). Statistical Models Based on Counting Processes. Springer-Verlag. New York, U.S.A.

Baena, J. & Salazar-Uribe, J.C. (2006). Análisis de recurrencia de falla aplicado a la determinación del tiempo esperado de falla de una empacadora de líquidos en la Cooperativa Lechera Colanta. *Memorias XVI Simposio de Estadística*, Bucaramanga, Colombia

Bhat, U. (1984). Elements of applied Stochastic Processes. Wiley.

Cardenas, M. & Díaz, L. G. (2013). Un modelo de sobrevida multivariado para eventos recurrentes por sujeto con evento terminal: deserción de clientes en la industria de las Telecomunicaciones. Departamento de Estadística, Universidad Nacional de Colombia, Sede Bogotá. Bogotá, Colombia.

Cook,R J. & Lawless, J. F. (2007). The Statistical Analysis of Recurrent Events. Springer. New York, U.S.A.

Correa, J. C., Salazar-Uribe, J. C. & Iral, R. (2010). Aproximación bayesiana al problema de la estimación de las tasas de transición en un modelo de estados múltiples. *Memorias XX Simposio de Estadística*, Santa Marta, Colombia.

Eichelsbacher, P. & Ganesh, A. (2004). A shared random efect parameter approach for longitudinal dementia data with non-ignorable missing data. *Biometrics*, 23, 211-219.

Gordon, P. (2001), Bayesian statistical John Wiley Sons. Chichester.

Green, S. B. & Byar, D. P. (2006). The choice of treatment for cancer patients based on covariate information: Application to prostate Cancer. *Bulletin Cancer*, 67, 477-488.

Guihenneuc-Jouyaux, C., Richardson, S. & Longini, Jr, IM. (2000), Modeling Markers of Disease Progression by a Hidden Markov Process: Application to Characterizing CD4 Cell Decline. *Biometrics*, 56(**3**), 733-741.

Hans, C. & Dunson, D. (2005). Bayesian inference on umbrella orderings. *Biometrics*, 61, 1018-1026.

Harezlak, J., Gao, S. & Hui, S. L. (2003). An illnes-death stochastic model in the analysis of longitudinal dementia data. *Statistics in Medicine*, 22, 1465-1475.

Hendrie, H. C., Ogunniyi, A. & Hall, K. S., Baiyewu, O., Unverzagt, F. W., Gureje, O. & *et al.* (2001). Incidence of dementia and Alzheimer disease in 2 communities: Yoruba residing in Ibadan, Nigeria, and African Americans residing in Indianapolis, Indiana. *JAMA*, 285(**6**), 739-747.

Iral, R. & Salazar-Uribe, J. C.(2007). Estimación de funciones de intensidad en un modelo de Markov de tres estados bajo el efecto de covariables con datos longitudinales. Tesis de Maestría: Universidad Nacional de Colombia, Sede Medellín.

Joly, P. & Commenges, D.(1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS *Biometrics*, 55, 887-890.

Jordan, Y. , Lerma, L.F. and Toro, E.(2008). Aplicación de cadenas de Markov continuas a las estadísticas del secuestro en Colombia. *Scientia et Technica*, XIV(**38**), 235-240.

Kay, R.(1986). Treatment Effects in Competing-Risks Analysis of Prostate Cancer Data. *Biometrics*, 42(**1**), 203-211.

Lawless, J. F.(2002). Statistical Models and Methods for Lifetime Data. Wiley Series in Probability and Statistics.

Martínez, C., Ramírez, G. & Vásquez, M.(2009). Pruebas no paramétricas para comparar curvas de supervivencia de dos grupos que experimentan eventos recurrentes.*Revista Ingeniería U.C.*, 5, 45-55.

Martínez, C., Ramírez, G. & Vásquez, M.(2011). Pruebas estadísticas para comparar curvas de supervivencia de k grupos con eventos recurrentes.*Ingeniería Industrial. Actualidad y Nuevas Tendencias*, 6, 7-18.

Peña, E., Strawderman, R.& Hollander, M.(2001). Nonparametric estimation with recurrent event data.*The computer Journal*, 99, 1299-1315.

Salazar-Uribe, J. C., Tyas, S.L., Snowdon, D. A., Desrosiers, M. F., Riley, K. P., Mendiondo, M. S. & Kryscio, R. J. (2003), Estimating intensity functions on multi-state Markov models with application to the Nun Study. *JSM*, San Francisco, EEUU.

Salazar-Uribe, J.C., Iral, R., Calvo, E., Rojas, A., Hincapíe, M. E., Anaya, J. M. & Díaz, F. J. (2001), Three state Markov model: comparing three parameterizations of the transition intensity rate. Application to rheumatoid arthritis data.*Revista Colombiana de Estadística*, 30(**2**), 213-229.

Salazar-Uribe, J. C., Iral, R., Correa, J. C., Rojas, A. & Anaya, J. M.(2014). Enfoque bayesiano para obtener las tasas de transición en un modelo de estados múltiples. Aplicación a datos sobre artritis reumatoide.*Comunicaciones en Estadística*, 7(**2**), 201-220.

SAS Institute Inc. (2004). SAS/IML 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Tanner, Martin A. (1996). Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.Third Edition. Springer Series in Statistics. Springer-Verlag New York.

Valencia, G. A. & Salazar, J. C. (2012). A statistical approach to reduce malware inside an information system in banking sector.*Proceeding of the 2012 World Congress in Computer Science, Computer Engineering, and Applied Computing*, Las Vegas Nevada, EEUU.

Wang, M. & Chang, S. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, 94, 146-153.

Wei, L., Lin, D. & Weissfeld, L.(1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.