Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 10 N°1 enero-junio de 2021 • ISSN-e 2357-5749 • Artículo Investigación • Páginas 6 a 19
DOI: https://doi.org/10.15446/revfaccienc.v10n1.84237

# A PROPOSAL FOR ESTIMATING INTENSITY FUNCTIONS IN A THREE-STATE MODEL IN THE PRESENCE OF ARBITRARY CENSORING[a]

# UNA PROPUESTA PARA ESTIMAR FUNCIONES DE INTENSIDAD EN UN MODELO DE TRES ESTADOS EN PRESENCIA DE CENSURA ARBITRARIA

LUISA FERNANDA ROSALES-CERQUERA[b*], RENÉ IRAL-PALOMINO[c], MAURICIO ALEJANDRO MAZO-LOPERA[d], JUAN CARLOS SALAZAR-URIBE[e]

Recibido 18-12-2019, aceptado 22-09-2020, versión final 16-10-2020.
Research paper

**ABSTRACT:** Multi-state are useful tools to model the dynamics of recurring processes over time or some changing phenomenon over time. This paper presents a methodology to estimate time-dependent intensity functions in the presence of interval censoring and right-hand censoring when considering a three-state model like the disease-death one. The likelihood function is deduced mathematically, which incorporates information that has been collected longitudinally, as well as the different modes of censoring. This likelihood should be optimized numerically with the help of a Gauss quadrature since in that expression there is an integral which is related to censored units. A piecewise function-based method is explored through a simulation study to obtain an estimate of the intensities.

**KEYWORDS:** Statistics; intensity function; Markov model; censored data; simulation study.

**RESUMEN:** Los modelos de múltiples estados conforman una importante familia de herramientas estadísticas que sirven para modelar la dinámica de procesos recurrentes a través del tiempo o algún fenómeno cambiante en el tiempo. En este trabajo se presenta una metodología para estimar funciones de intensidad dependientes del tiempo en presencia de censura de intervalo y censura a derecha cuando se considera un modelo de tres estados del estilo enfermedad - muerte. Se deduce matemáticamente la función de verosimilitud, la cual incorpora información que ha sido recolectada longitudinalmente, así como los diferentes modos de censura. Esta verosimilitud se debe optimizar numéricamente con la ayuda de una cuadratura de Gauss ya que en dicha expresión aparece una integral que se relaciona con unidades censuradas. Se explora, por medio de un estudio de simulación, un método basado en funciones por tramos para

[a]Rosales Cerquera, L. F. Iral Palomino, R.; Mazo, M. & Salazar - Uribe, J . C. (2021). A proposal for estimating intensity functions in a three-state model in the presence of arbitrary censoring. *Rev. Fac. Cienc.*, 10 (**1**), 6–19. DOI: https://doi.org/10.15446/revfaccienc.v10n1.84237

[b]M.Sc. in Statistics. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín
[*]Corresponding author: lufrocea@gmail.com
[c]M.Sc. in Statistics. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín
[d]Ph.D. in Statistics. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín
[e]Ph.D. in Statistics. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín

obtener una estimación de las intensidades.

**PALABRAS CLAVE:** Estadística; función de intensidad; modelo de Markov; datos censurados; estudio de simulación.

# 1. INTRODUCTION

Multi-state models have received attention from the scientific community in recent years, as can be seen in the works of Joly *et al.* (1998), Joly and Commengues (1999), Harezlak *et al.* (2003), Kay (1986), Kryscio *et al.* (2006), Salazar *et al.* (2005), Salgado *et al.* (2018), Tovar & Salazar (2009), Yeh *et al.* (2010), Wei *et al.* (2014). These types of models are useful for studying the role of covariates in the progression or regression of certain diseases such as cancer (Kay, 1986), dementia (Kryscio *et al.*, 2006), rheumatoid arthritis (Salazar *et al.*, 2007), cardiovascular disease (Hajihosseini *et al.*, 2016), and AIDS (Frydman, 1992). A simple example of a multi-state model is the classic survival analysis in which only two states are considered: alive and death and it is perhaps one of the models most studied in the statistical literature and one of its ojectives is to estimate the hazard function (Cox, 1972; Hedeker *et al.*, 2000).
In the case of multi-state models, these hazard functions are called transition functions which are formally defined later.

The estimation of the hazard function is complicated when working with longitudinal data due mainly to censoring and dependency between observations. Harezlak *et al.* (2003) estimate these transition functions in a three state model but does not take into account neither right nor interval censoring modes. To achieve an effective estimation of the transition functions in the presence of longitudinal data, censoring and models with more than two states, it is necessary to reformulate the classic treatment suggested by Cox (1972) and Andersen *et al.* (1993) and adapt it to these types of situations. Making this adaptation is one of the main objectives of this work.

If a disease can be classified in several states according to its severity, assessing the risk of a person transiting between the different states is clinically important (Kryscio *et al.*, 2006; Salazar *et al.*, 2007; Frydman, 1992). For example, if the estimated risk of transition from a moderate to a severe state of the disease turns out to be unusually large, a prevention plan can be implemented. It is this clinical importance that motivates this work without ignoring that the methodology developed here can be applied in other areas that are not necessarily clinical. In this work, a method of estimating the time-dependent transition functions is proposed for a three-state model (see figure 1) in the presence of censoring.

The first advantage of the method proposed here is that it directly yields estimates of the intensity function that quantify risk. Another advantage of the methodology is that it provides estimates for the variance of the estimates of the transition functions, which allows to build confidence intervals. When censoring modes are taken into account (especially interval censoring) to estimate intensity functions, the resulting

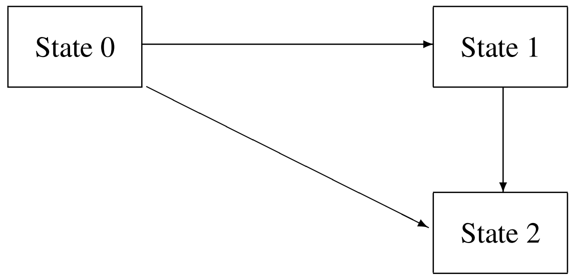



Figure 1: Three-state model with an absorbing state (State 2). Source: Elaborated by authors.

likelihood function acquires a complicated form from a mathematical point of view and this makes its evaluation and maximization more complex and thus it requires numerical methods for optimization. This work constitutes a generalization of the work proposed by Joly *et al.* (1998), Joly and Commengues (1999) although the estimation strategy adopted here is different from the one proposed by them; however, the likelihood function has a remarkable resemblance.

## 2. THE MODEL

A three-state model ($k=3$) is proposed to estimate the transition functions. Formally, the transition intensity function from state $j$ to state $l$, $j,l=1,\cdots,k$ is defined as:

$$\alpha_{jl}(t) = \lim_{\Delta_t \to 0} \frac{P\left[X(t+\Delta_t)=l \mid X(t)=j\right]}{\Delta_t}.$$

Specifically, an idea proposed by Faddy (1976) is used to estimate the time-dependent intensity functions by assuming that they are piecewise constant. For a multi-state model, a functional representation can be found for the different transition intensities when they depend on time if it is assumed that the $\boldsymbol{A}$ matrix, the intensity matrix, is piecewise constant, that is, $\boldsymbol{A}(t)=\boldsymbol{A}_h$, for $\tau_{h-1}<t\leq\tau_h$, $t_0=0$, where the $\tau$ define the classes in the piecewise function.

In this case, for the transitions from $i$ to $j$ a piecewise function can be formulated as,

$$\alpha_{ij}(t) = \begin{cases} \beta_1 & \text{if } \tau_0 < t \leq \tau_1, \\ \beta_2 & \text{if } \tau_1 < t \leq \tau_2, \\ \vdots & \\ \beta_h & \text{if } \tau_{h-1} < t \leq \tau_h, \end{cases}$$

where the $\beta$´s are estimated using the logarithm of the likelihood function. The methodology proposed in this work constitutes a generalization of the one proposed by Joly *et al.* (1998) and Joly and Commengues

(1999) in the sense that the model considered here allows more general transitions; for example, it allows transitions from State 0 to 2 without having to go through State 1. With this level of flexibility, the likelihood function can be deduced mathematically, which in principle is much more complex than that proposed by those authors. Unfortunately, this function does not have a closed analytical form which implies that its evaluation and subsequent maximization is laborious and must be treated numerically.

## 3. THE LIKELIHOOD FUNCTION

A number of independent, complete or incomplete observations (for instance, medical records) from a stochastic process is assumed $\{X(t), t \geq 0\}$ with $X(t) = 0$, 1, 2, which is observed for a certain period of time at irregular time intervals. At each visit, each patient is registered with information that is collected in a vector of the form: $(\text{Disease status, Time between visits, Covariates})'$. This generates longitudinal data and censoring.

The medical records are divided into four groups as follows:

(1) If the subject during the study was seen only once or if the subject during the study period remained in State 0 (Right censoring).

(2) The subject arrived at the study being in state one or passed from State 0 to 1.

(3) Subject changes from State 0 to 1 and from State 1 to 2 while the study is in progress.

(4) The individual entered the study in State 0 and he/she goes to State 2.

Observation $X_i$ is said to be interval censored if it belongs to $A_i = [L_i, R_i]$; right censored if it belongs to $A_i = [L_i, \infty)$, and left censored if it belongs to $A_i = [0, R_i)$. Let $X_i^{01}$ be the elapsed time in State 0 by the subject $i$ (it can be interval censored). Let $X_i^{12}$ be the time spent in State 1 by the $i$-th subject, $X_i^{01}$ and $X_i^{12}$ are assumed independent. $T_i$ is the time in which the subject $i$ was last seen; $T_i$ can be the right censored time for the first transition from State 0 to 1, the right censored time for the transition from State 1 to 2, or the time for the transition from State 0 to 1. Let $X_i^{02}$ be the elapsed time in State 0 before entering the absorbent state, and $T_d$ the time in which the subject $i$ was seen in the absorbent state (in a clinical study, for example, the absorbent state may be death, where the exact time of the event is usually known). That is, $X_i^{02} = T_d - L_i$, where $L_i$ is the beginning of the study for a particular subject.

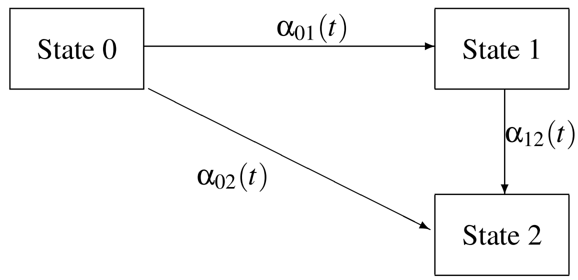



Figure 2: Time dependent intensity functions for the three-state model. Source: Elaborated by authors.

To build the likelihood function, we propose the following indicator variables:

$$\delta_{01} = \begin{cases} 0 & \text{If } i\text{-th subject is censored for the first transition} \\ 1 & \text{If } i\text{-th subject goes from State 0 to State 1} \end{cases}$$

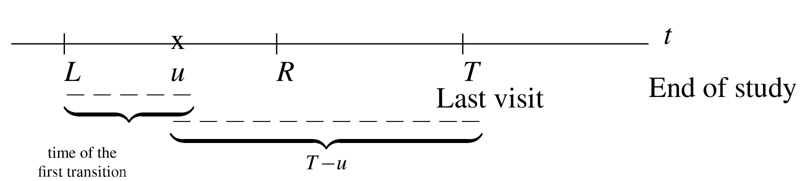
$$\delta_{12} = \begin{cases} 0 & \text{If } i\text{-th subject is censored for the second transition} \\ 1 & \text{If } i\text{-th subject goes from State 1 to State 2} \end{cases}$$

$$\delta_{02} = \begin{cases} 0 & \text{If } i\text{-th subject is right censored or visits State 1} \\ 1 & \text{If } i\text{-th subject goes from State 0 to State 2 without visiting State 1} \end{cases}$$

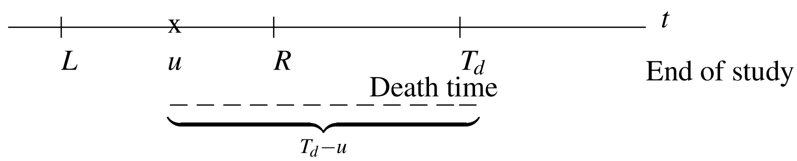
There exists four possible scenarios for a specific subject or unit:

1. $\delta^i_{01} = 0, \delta^i_{12} = 0, \delta^i_{02} = 0$. $i$-th subject has been seen once during the follow up. We do not have additional information about him/her.

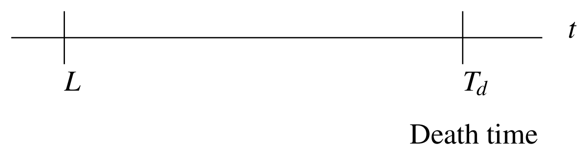

2. $\delta^i_{01} = 1, \delta^i_{12} = 0, \delta^i_{02} = 0$. $i$-th subject goes from 0 to 1 sometime in the interval [L,R]. We observe times $u$ (time for first transition) and $T - u$ (elapsed time from the first transition). We do not know if he/she is still alive.



3. $\delta_{01}^{i} = 1, \delta_{12}^{i} = 1, \delta_{02}^{i} = 0$. Similar to previous case, but we know the exact death time ($T_d$). We observe $u$: time of first transition, $T_d - u$: elapsed time from first transition.




4. $\delta_{01}^{i} = 0, \delta_{12}^{i} = 0, \delta_{02}^{i} = 1$. We know death time and that $i$-th subject has not acquired the disease during the study $T_d - L$.




## 3.1. Contributions to the likelihood in each case

Let $A_{lk}(t)$ be the cumulative intensity function associated with the time spent in state $l$, with $l$ = 0, 1 and $k = l + 1$ for the progressive model that we are considering.

- The contribution from subject $i$ to the likelihood is given by (Scenario 1):

$$\int_{L_i{}^0}^{\infty} e^{-A_{01}(u)} du$$

  Since $P(X > u) = S_x(u) = e^{-A_{01}(u)}$, for the first transition from State 0 to 1, when $X$ is right censored.

- A subject's contribution to likelihood is as follows (Scenario 2):

$$\int_{L_i{}^0}^{R_i^0} \alpha_{01}(u) e^{-A_{01}(u)} e^{-A_{12}(T_i - u)} du$$

  There was a transition from 0 to 1 which is: $f_x(u) = \alpha_x(u) e^{A_x(u)}$ and the second transition (from State 1 to 2) is censored, then $P(X > T - u) = e^{-A_x(T-u)}$.

- The contribution of a subject to the likelihood is given by (Scenario 3):

$$\int_{L_i{}^0}^{R_i^0} \alpha_{01}(u) e^{-A_{01}(u)} \alpha_{12}(T_i - u) e^{-A_{12}(T_i - u)} du$$

  Since $f_x(u) = \alpha_x(u) e^{-A_x(u)}$ for the first transition from 0 to 1, and transition from 1 to 2 is $f_x(T_d - u) = \alpha_x(T_d - u) e^{-A_x(T_d - u)}$

- The contribution of a subject to the likelihood is given by (Scenario 4):

$$\int_{L_i{}^0}^{R_i^0} \alpha_{02}(u) e^{-A_{02}(u)} du$$

  Since $f_x(u) = \alpha_x(u) e^{-A_x(u)}$ for a transition from State 0 to 2.

### 3.2. The log-likelihood

Considering the discussion of previous numeral, we have that the log-likelihood function is as follows:

$$\ell = \sum_{i=1}^{n} \ln\left\{ \int_{L_i^0}^{R_i^0} \left(e^{-A_{01}(u)}\right)^{1-\delta_{02}^i} \left(\alpha_{01}(u)e^{-A_{12}(T_i-u)}\right)^{\delta_{01}^i} \left(\alpha_{12}(T_i-u)\right)^{\delta_{12}^i} \left(\alpha_{02}(u)e^{-A_{02}(u)}\right)^{\delta_{02}^i} du \right\}$$

Note that this function has no closed analytical form and it requires numerical methods for optimization. To evaluate the integral, the Gauss quadrature method will be used with 16 points (Abramowitz & Stegun, 1972):

$$\int_a^b f(y)dy \approx \frac{b-a}{2}\sum_{i=1}^{n} w_i f(y_i)$$

$$y_i = \left(\frac{b-a}{2}\right)x_i + \left(\frac{b+a}{2}\right)$$

with

$$f(u) = \left(e^{-A_{01}(u)}\right)^{1-\delta_{02}^i} \left(\alpha_{01}(u)e^{-A_{12}(T_i-u)}\right)^{\delta_{01}^i} \left(\alpha_{12}(T_i-u)\right)^{\delta_{12}^i} \left(\alpha_{02}(u)e^{-A_{02}(u)}\right)^{\delta_{02}^i},$$

$a = {L_i}^0$, $b = R_i^0$, $w_i$ are the weights, and $x_i$ are the abscissas. Subsequently, the likelihood function is optimized using the Newton-Raphson algorithm which iteratively calculates an estimate for the vector $\boldsymbol{\theta} = (\alpha_{01}, \alpha_{02}, \alpha_{12})'$.

The longitudinal nature of the data is taken into account in the model using a one-step transition matrix whose probabilities are obtained as follows. The probability of transition from state $j$ to state $l$ in $(s,t)$ is defined as:

$P_{jl}(s,t) = P[X(s) = l \mid X(t) = j]$ with $s > t$ and $j \leq l$, where $j = 0,1$, $l = 0,1,2$. The transition matrix used to simulate the states is defined as:

$$\boldsymbol{P}(t) = \begin{pmatrix} P_{00}(s,t) & P_{01}(s,t) & P_{02}(s,t) \\ 0 & P_{11}(s,t) & P_{12}(s,t) \\ 0 & 0 & 1 \end{pmatrix}$$

where, $\sum_{j=0}^{2} P_{ij} = 1$, $i = 0,1,2$ y $P_{ij} \geq 0$. Under the assumption of constant intensities, the solution of the Kolmogorov forward equations (which relates transition probabilities to intensities (Bhat, 1984)) $\frac{d\boldsymbol{P}(t)}{dt} = \boldsymbol{P}(t)\boldsymbol{A}$, has a closed form in this setting and it is given by:

$$\begin{aligned}
P_{00}(t) &= \exp\{-(\alpha_{01}+\alpha_{02})t\} \\
P_{01}(t) &= \frac{\alpha_{01}}{\alpha_{12}-\alpha_{01}-\alpha_{02}}[\exp\{(\alpha_{01}+\alpha_{02})t\} - \exp\{(-\alpha_{12})t\}] \\
P_{02}(t) &= 1 - P_{00}(t) - P_{01}(t) \\
P_{11}(t) &= \exp\{-(\alpha_{12})t\} \\
P_{12}(t) &= 1 - P_{11}(t) \\
P_{22}(t) &= 1
\end{aligned}$$

We used these results to conduct the simulation study.

## 4. SIMULATION STUDY

In order to illustrate the methodology, a simulation study was carried out, where a three-state model with an absorbent state was considered. The purpose of this study is to evaluate the ability of the proposed method to estimate the time-dependent transition functions from healthy to sick ($\alpha_{01}(t)$), healthy to death ($\alpha_{02}(t)$) and from sick to death ($\alpha_{12}(t)$) where values reported in Harezlak *et al.* (2003) are taken as a reference. Additionally, we want to observe how good the method preserves the level of significance when two types of interval estimates are used: Wald type intervals (Van der Vaart, 2000) and Lee type intervals (Lee, 1992). Since we have to simulate different states during visits over time, the dependency due to longitudinal data is taken into account in the model using a one-step transition matrix:

$$\boldsymbol{P}(t) = \begin{pmatrix} P_{00}(s,t) & P_{01}(s,t) & P_{02}(s,t) \\ 0 & P_{11}(s,t) & P_{12}(s,t) \\ 0 & 0 & 1 \end{pmatrix}$$

Where, $\sum_{j=0}^{2} P_{ij} = 1$, $i = 0,1,2$ y $P_{ij} \geq 0$. We have also considered a homogeneous time model in which $\alpha_{jl}(t)$ is:

$$\alpha_{ij}(t) = \begin{cases} \alpha_{jl}^{(0)}, & \text{for } 70 \leq t, \\ \alpha_{jl}^{(1)}, & \text{for } 70 < t \leq 80, \\ \alpha_{jl}^{(2)}, & \text{for } t > 80. \end{cases}$$

Then, we use the solution of the Kolmogorov Forward equation discussed in the previous section. The simulation study consists of 1000 repetitions under the following conditions: Sample sizes: 800, 1500, 2500, and 3000 (typical sample sizes in epidemiological studies), maximum number of visits: 4 and 5, (to reflect the longitudinal nature of the data), first order Markov dependence (to incorporate the dependence on the observations). As Faddy (1976) suggests, age was used considering three groups: below 70, between 70 and 80, and over 80. To optimize the log-likelihood function we used the Newton-Raphson algorithm in conjunction with the Gauss quadrature. All simulations were performed with the SAS IML package V9.2 SAS (2011), licensed to the Universidad Nacional de Colombia. Figures 3 and 4 show the standard deviation of the estimators for the three intensities taking into account the number of visits. The ranges of the standard deviations are between 0.001 and 0.020, having the same pattern for the three age groups. As the sample size increases the standard deviation for the estimates decreases, as expected, being the group of those under 70 years where the smallest standard deviation was observed and the opposite case was presented in the group of those over 80 years, this behavior is repeated for the three intensities for both cases (four and five visits).

As an example, for $N = 1500$, a piecewise estimate of the time-dependent intensity functions is given by (see Figure 5).

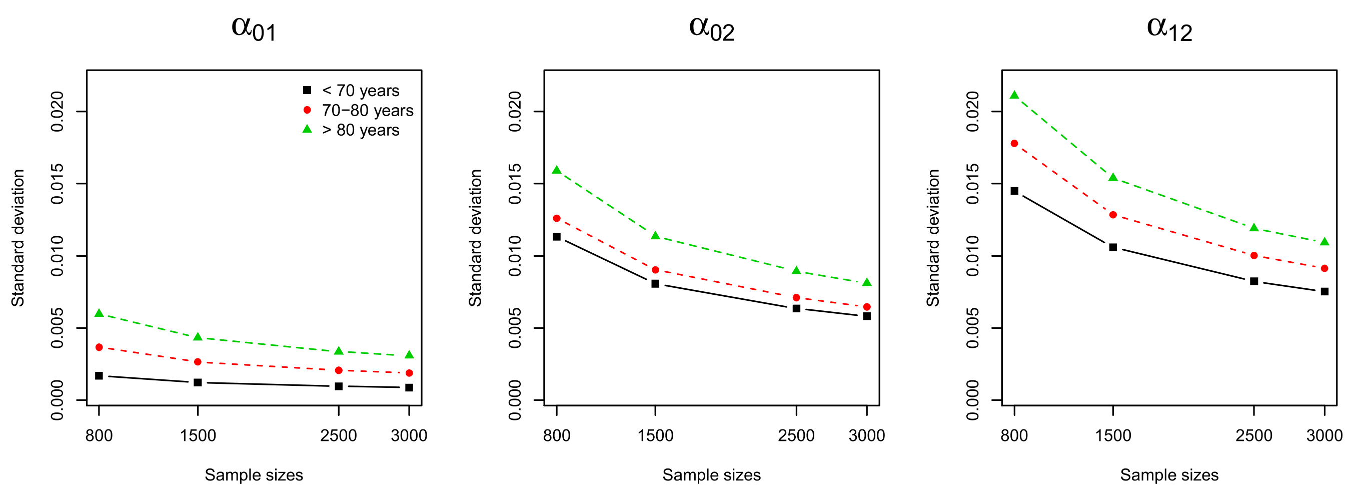



Figure 3: Mean standard deviation by sample size in the case of 4 visits. Source: Elaborated by authors.

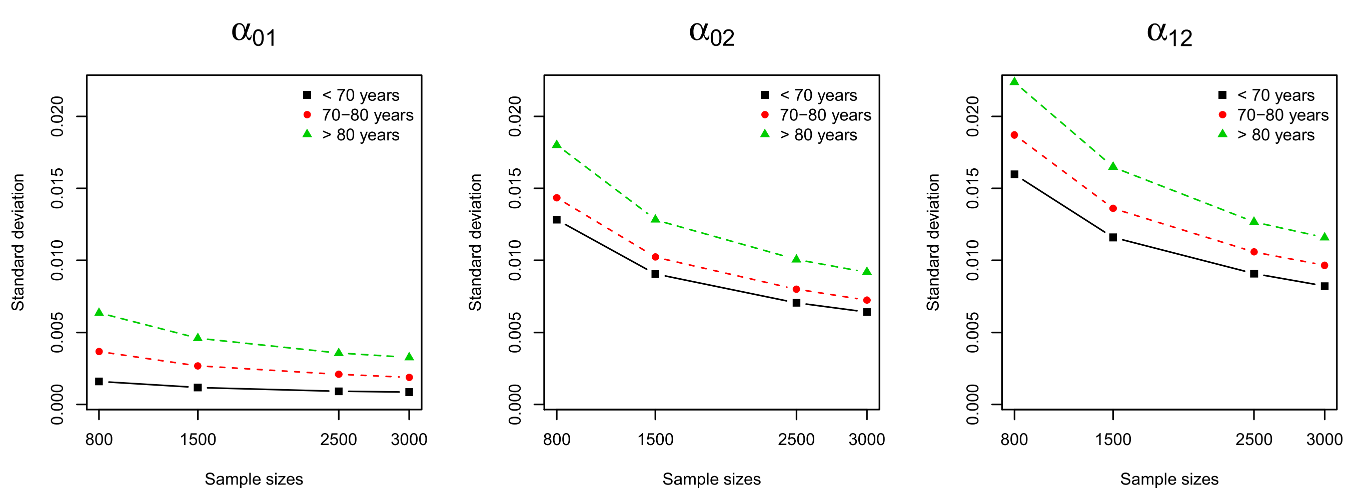



Figure 4: Mean standard deviation by sample size in the case of 5 visits. Source: Elaborated by authors.

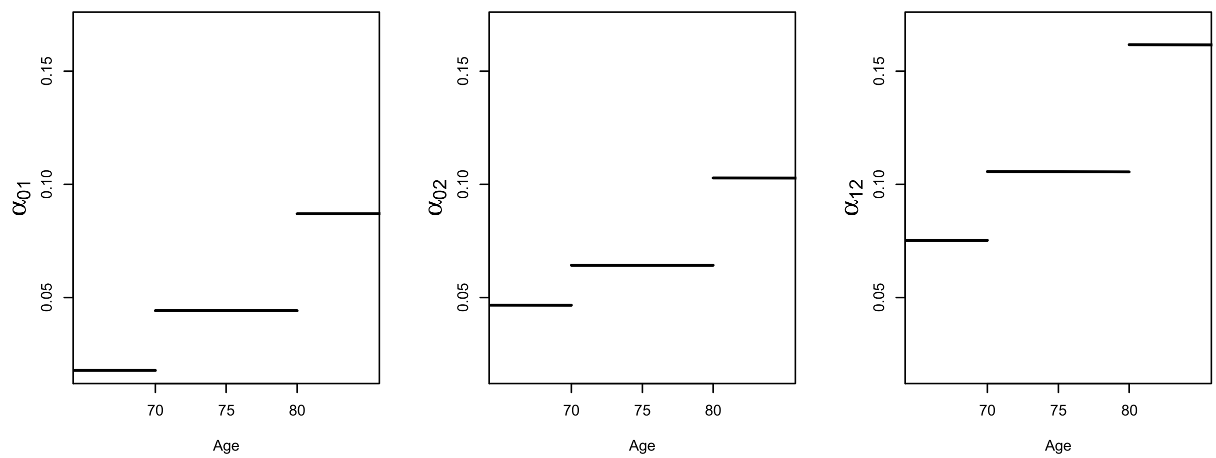



Figure 5: Estimate of intensity functions for a sample size of 1500 and a maximum number of visits of 4. Source: Elaborated by authors.

In order to illustrate the behavior of the estimated parameters with respect to the simulated ones, a coverage study was carried out to evaluate the ability of the estimation method to keep the probability of type I error small. This consists of the estimation of 1000 95 % confidence intervals and the objective is to know how many times the reference value is contained in those confidence intervals. For this, two types of intervals will be used: Wald type confidence intervals Van der Vaart (2000) and Lee type confidence intervals Lee (1992).

## 4.1. Coverage with Wald type confidence intervals

Wald type intervals are as follows:

$$\widehat{\alpha} \pm Z_{\alpha/2}\widehat{SE}\left(\widehat{\alpha}\right).$$

Once the confidence intervals were calculated, the proportion of times in which the parameter was contained in the estimated confidence interval was obtained. Table 1 shows the results for the three age groups, the four sample sizes and for the three intensities, where the coverage for the second intensity $\alpha_{02}$ in all age groups and sample sizes is 100 %, for the other intensities the coverings vary between 82 % and 98,6 %, although in the group of those over 80 years of age and five visits where the highest coverage for the first intensity is presented, it happens that as the sample size increases the coverage decreases until it reaches the lowest of all with 68,9 %. The presence of these values that are so far from the true level of significance can be attributed to a poor specification of the CI. For this reason, it was decided to try with the Lee type confidence intervals.

Table 1: Observed coverage and mean length (between parenthesis) of the Wald type interval.

| | Four visits | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $< 70$ years | | | 70-80 | | | Greater than 80 years | | |
| N | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ |
| 800 | 0,889 | 1,000 | 0,945 | 0,966 | 1,000 | 0,930 | 0,985 | 1,000 | 0,944 |
| | (0,006) | (0,050) | (0,063) | (0,014) | (0,056) | (0,073) | (0,025) | (0,071) | (0,088) |
| 1500 | 0,883 | 1,000 | 0,972 | 0,953 | 1,000 | 0,941 | 0,979 | 1,000 | 0,951 |
| | (0,005) | (0,036) | (0,046) | (0,011) | (0,040) | (0,053) | (0,018) | (0,050) | (0,065) |
| 2500 | 0,878 | 1,000 | 0,958 | 0,939 | 1,000 | 0,931 | 0,949 | 1,000 | 0,929 |
| | (0,004) | (0,028) | (0,035) | (0,008) | (0,032) | (0,041) | (0,014) | (0,040) | (0,050) |
| 3000 | 0,876 | 1,000 | 0,975 | 0,93 | 1,000 | 0,921 | 0,884 | 1,000 | 0,937 |
| | (0,003) | (0,025) | (0,032) | (0,007) | (0,029) | (0,038) | (0,013) | (0,036) | (0,045) |
| | Five visits | | | | | | | | |
| | $< 70$ years | | | 70-80 | | | Greater than 80 years | | |
| N | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ |
| 800 | 0,895 | 1,000 | 0,972 | 0,965 | 1,000 | 0,951 | 0,986 | 1,000 | 0,964 |
| | (0,007) | (0,044) | (0,057) | (0,014) | (0,050) | (0,050) | (0,024) | (0,062) | (0,083) |
| 1500 | 0,862 | 1,000 | 0,964 | 0,948 | 1,000 | 0,956 | 0,957 | 1,000 | 0,947 |
| | (0,005) | (0,032) | (0,042) | (0,010) | (0,035) | (0,051) | (0,017) | (0,045) | (0,060) |
| 2500 | 0,851 | 1,000 | 0,957 | 0,938 | 1,000 | 0,961 | 0,784 | 1,000 | 0,94 |
| | (0,004) | (0,025) | (0,032) | (0,008) | (0,028) | (0,039) | (0,013) | (0,035) | (0,047) |
| 3000 | 0,82 | 1,000 | 0,97 | 0,928 | 1,000 | 0,959 | 0,689 | 1,000 | 0,919 |
| | (0,003) | (0,023) | (0,029) | (0,007) | (0,025) | (0,036) | (0,012) | (0,032) | (0,043) |

## 4.2. Coverage with Lee type confidence intervals

Since in the simulation study we assumed a piecewise constant intensity, the estimation of the confidence intervals under this assumption and taking into account the right censoring mechanism, the Lee type intervals are as follows (Lee, 1992):

$$\widehat{\alpha} \pm \frac{Z_{\alpha/2}\,\widehat{\alpha}}{\sqrt{r-1}},$$

where $r$ is the number of right-censored observations.

The results are reported on Table 2, where the coverage values are better than those obtained using Wald type confidence intervals. It was found, as for the Wald-type confidence intervals, that the second intensity had a coverage of 100 % for all age groups and all sample sizes. In the groups between 70 and 80 years and those over 80 the coverage for the first intensity is 100 %, in both cases (4 and 5 visits). The range of coverage values that are not 100 % are between 84.7 % and 99.5 %. Note that the Lee type intervals have a better performance than the Wald type intervals, so it can be said that, in general, the method preserves the level of significance reasonably.

Table 2: Observed coverage and mean length (between parenthesis) of the Lee type interval.

| | Four visits | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $<70$ years | | | 70-80 | | | Greater than 80 years | | |
| N | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ |
| 800 | 0,994 | 1,000 | 0,847 | 1,000 | 1,000 | 0,887 | 1,000 | 1,000 | 0,957 |
| | (0,011) | (0,029) | (0,046) | (0,027) | (0,040) | (0,065) | (0,054) | (0,064) | (0,098) |
| 1500 | 1,000 | 1,000 | 0,964 | 1,000 | 1,000 | 0,954 | 1,000 | 1,000 | 0,992 |
| | (0,011) | (0,028) | (0,044) | (0,026) | (0,038) | (0,062) | (0,051) | (0,061) | (0,096) |
| 2500 | 0,995 | 1,000 | 0,862 | 1,000 | 1,000 | 0,885 | 1,000 | 1,000 | 0,951 |
| | (0,006) | (0,016) | (0,026) | (0,015) | (0,022) | (0,037) | (0,030) | (0,036) | (0,056) |
| 3000 | 0,995 | 1,000 | 0,885 | 1,000 | 1,000 | 0,872 | 1,000 | 1,000 | 0,957 |
| | (0,006) | (0,015) | (0,023) | (0,014) | (0,021) | (0,033) | (0,028) | (0,033) | (0,051) |
| | Five visits | | | | | | | | |
| | $<70$ years | | | 70-80 | | | Greater than 80 years | | |
| N | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{12}$ |
| 800 | 0,992 | 1,000 | 0,899 | 1,000 | 1,000 | 0,944 | 1,000 | 1,000 | 0,993 |
| | (0,011) | (0,029) | (0,046) | (0,028) | (0,040) | (0,069) | (0,053) | (0,064) | (0,105) |
| 1500 | 1,000 | 1,000 | 0,97 | 1,000 | 1,000 | 0,993 | 1,000 | 1,000 | 1,0001 |
| | (0,011) | (0,028) | (0,045) | (0,026) | (0,038) | (0,066) | (0,051) | (0,061) | (0,102) |
| 2500 | 0,988 | 1,000 | 0,906 | 1,000 | 1,000 | 0,961 | 1,000 | 1,000 | 0,993 |
| | (0,006) | (0,016) | (0,026) | (0,015) | (0,022) | (0,039) | (0,030) | (0,036) | (0,059) |
| 3000 | 0,990 | 1,000 | 0,919 | 1,000 | 1,000 | 0,960 | 1,000 | 1,000 | 0,984 |
| | (0,006) | (0,015) | (0,024) | (0,014) | (0,021) | (0,035) | (0,027) | (0,033) | (0,054) |

# 5. CONCLUSIONS

In this work the problem of the estimation of intensity functions in a three-state model with arbitrary censoring has been discussed. The formulation of the likelihood function is presented for a more general case than that proposed by Joly and Commengues (1999). Based on the simulation study, some advantages and disadvantages are evident. Among the advantages it is observed that by including different types of censoring, the loss of information is minimized, which makes the method realistic with applicability to studies where the response is categorical.

Another advantage of the method is that it provides estimates of the standard deviation, which facilitates the calculation of confidence intervals for the intensity functions. In the simulation study, the method showed a good performance in addition to providing a very good fit despite the fact that the function has no closed analytical form and it requires the use of numerical methods for optimization. Among the disadvantages, we can say that the method requires large amounts of data and that it has few flexibility in the specification of the covariance matrix, indeed it only allows to specify a first order Markov dependency. The CI obtained in the simulation study showed a very adequate coverage, which makes it possible to recommend their use in practice. The three-state model has been widely used in practice, so the present work represents an excellent alternative to obtain estimates of intensity rates in the presence of arbitrary censorship modes.

As a future direction, a new simulation study based on MCMC and obtaining bootstrap-based intervals can be considered.

**Acknowledgments**

To the School of Statistics of the Faculty of Sciences, Universidad Nacional de Colombia, Campus Medellín, for its continuous support to all our research initiatives.

## References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications, Inc.

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer - Verlag. New York.

Bhat, U. N. (1984). *Elements of applied stochastic processes*. 2nd ed. New York: John Wiley and sons.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* 32(**Series B**), 187-220.

Faddy, M. J. (1976). A note on the general time-dependent stochastic compartmental model. *Biometrics* 32, 443-448.

Frydman, H. (1992). A Nonparametric estimation procedure for a periodically observed three state Markov process, with application to AIDS. *Journal of Royal Statistical Society* 4(**Series B**), 853-866.

Hajihosseini, M., Kazemi, T. & Faradmal, J. (2016). Multistate models for survival analysis of cardiovascular disease process. *Revista Española de Cardiología* 69(**7**), 714-715.

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

Hedeker, H., Siddiqui, O. & Hu, F. B. (2000). Random-effects regression analysis of correlated group-time survival data. *Statistical Methods in Medical Research* 9, 161-179.

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.

Joly, P., Commenges, D. & Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics* 55(**1**), 185-194.

Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics* 42, 855-865.

Kryscio, R. J., Schmit, F. A., Salazar, J., Mendiondo, M. S. & Markesbery, W. R. (2016). Risk factors for transitions from normal to mild cognitive impairment and dementia. *Neurology* 66, 828-832.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Lee, E. T. (1992). *Statistical methods for survival data analysis*. 2nd Ed. John Wiley and Sons. New York.

Salazar, J. C., Palomino, R., Calvo, E., Rojas, A., Hincapié, M. E., Anaya, J. M. & Díaz, F. (2007). 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, J. C., Schmitt, F. A., Yu, L., Mendiondo, M. S. & Kryscio, R. J. (2005). Shared random effects analysis of multi-state Markov models: application to a longitudinal study of transitions to dementia. *Statistics in Medicine* 26(**3**), 568-580.

Salgado, K., Salazar-Uribe, J. C., Gallo-Villegas, J., Valencia, A., Espíndola-Fernández, D., Mesa, C., de la Calle, J., Montoya, Y. & Aristizábal, D. (2018). Evaluación de la costo-efectividad de un modelo integral de tratamiento ambulatorio en pacientes con síndrome coronario agudo: aplicación de un modelo de Markov probabilístico. *Pan American Journal of Public Health* 42, 1-9.

SAS Institute Inc. Base SAS 9.2 (2011). Procedures Guide. Cary, NC: SAS Institute Inc.

Tovar, R. & Salazar, J. C. (2009). A linear mixed model adapted to a markov chain with ordinal state space. Application to data about grade point average (GPA). *Revista Colombiana de Estadística* 32(**2**), 213-230.

Wei, S., Xu, L. & Kryscio, R. (2014). Markov transition model to dementia with death as a competing event. *Computational Statistics and Data Analysis* 80, 78-88.

Yeh H., Chan, W., Symanski, E. & Davis, B. (2010). Estimating transition probabilities for ignorable intermittent missing data in a discrete-time markov chain *Communications in Statistics: Simulation and Computation* 39(**2**), 433-448.