

# Resultados

Rafael Eduardo Diaz

14 de julio de 2019

## 1. Aplicación

A continuación, ilustramos todos los modelos descritos anteriormente aplicándolos a dos conjuntos de datos, para la serie anual del número de homicidios en Colombia de 1960 a 2018 se ajustaron varios PHMM, mientras que para la serie mensual de incendios forestales en Colombia, entre el 2002 y 2016 se ajustaron diversos ZIP-HMM. Antes de que se ajusten los modelos, se llevó a cabo un análisis exploratorio básico del conjunto de datos que aborda algunos problemas que generalmente se presentan al visualizar los datos de conteo. Al final de la sección, se comparan todos los modelos ajustados, tanto desde el enfoque clásico como Bayesiano, y se selecciona el mejor modelo a partir de las dos metodologías.

Para ambas series, la aplicación de modelos estándar como modelos auto regresivos de media móvil (ARMA) sería inapropiado, ya que estos modelos se basan en la distribución normal. En cambio, se propone un modelo usual para datos con conteos la distribución de Poisson, pero, como se demostrará más adelante, las series presenta una sobredispersión considerable con respecto a la distribución de Poisson, y fuerte dependencia serial positiva además de inflación en ceros en el caso de la serie de incendios. Por lo tanto, un modelo que consiste en variables aleatorias independientes de Poisson; sería por dos razones inadecuado. Primero que puede haber algunos períodos con una baja tasa de homicidios e incendios, y algunos con una tasa relativamente alta. Los HMMs, permiten que la distribución de probabilidad de cada observación dependa del estado no observado (*u oculto*) de una cadena de Markov, por lo tanto puede acomodar la sobredispersión y la dependencia serial al mismo tiempo.

### 1.1. Descripción de los datos

**Homicidios:** Esta tabla contiene las cifras actualizadas de homicidios en Colombia 1960-2018, con base en la Compilación de estadísticas históricas económicas y sociales, extraída del [departamento Nacional de Planeación](#) (DNP) se consultó específicamente el capítulo 8 indicadores de violencia, se complementó junto con las estadísticas delectivas de la [Policía Nacional](#) y Medicina Legal. Los datos publicados corresponden a consolidados de los Delitos de Impacto del país, así mismo la Actividad Operativa realizada por la Policía Nacional. Mientras que para la población total Colombiana se extrajo la información de la sección Estadísticas por tema, demografía y población. La serie es anual para un total de 59 observaciones y se expresa como el número de homicidios por cada 100.000 habitantes comúnmente conocida como *Tasa de homicidios*, para ser posible la modelación se redondeo la cifra al entero más cercano. Nota: La confiabilidad de los datos de la tasa de asesinatos puede variar, de acuerdo a la fuente.

31	31	31	32	31	32	30	29	31	19	21	23	23	23	24	24	25	25	27	26	27
29	36	34	30	30	40	48	52	63	65	68	78	76	74	70	66	68	62	58	61	
65	49	69	56	48	42	40	39	36	35	34	32	32	32	27	26	25	25	25		

Tabla 1: Número de homicidios por 100.000 habitantes en Colombia, 1960 - 2018.

**Incendios:** Los datos referentes a incendios forestales en Colombia, fueron recolectados de la página del IDEAM - Instituto de Hidrología, Meteorología y Estudios Ambientales que ha venido realizando una revisión histórica y consolidado de los datos reportados por las siguientes instituciones: entidades

del SINA, entidades del Sistema Nacional para la Prevención y atención de Desastres, la Defensa Civil, entre otras, y aunque se ha adoptado un Formulario Único de Captura (MAVDT & otros, 2002), con el fin de estandarizar la información, este no ha sido utilizado en su totalidad y existen otros formatos desarrollados por las distintas entidades, de acuerdo con sus particularidades técnicas e informáticas, lo que ha dificultado la estandarización en el flujo de información.

Las estadísticas sobre incendios en Colombia, permiten en términos generales, realizar análisis de su comportamiento bajo diferentes escenarios, esto es, por regiones, departamentos o municipios, con Niño o en condiciones climáticas normales, por cobertura vegetal afectada, por Corporación Autónoma Regional, por año o por mes, y de esta manera, poder ser utilizarlas para priorizar áreas, orientar acciones o sustentar la necesidad de realizar estudios más detallados. El Ideam ha venido realizando una revisión histórica de los datos reportados por las instituciones anteriormente mencionadas, con el fin de tener datos más confiables que permitan tener una mejor aproximación al tema. La variable de interés es el número de grandes incendios forestales (GIF), y se definen como aquellos incendios que superan las 500 hectáreas forestales afectadas. El número de observaciones es mensual iniciando en enero del 2002 y finalizando en diciembre del 2016, para un total 180 observaciones.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2002	10	8	12	2	0	1	0	6	1	1	0	0
2003	5	6	11	3	0	0	5	0	3	0	0	0
2004	7	14	10	1	0	0	3	12	4	1	0	3
2005	5	6	13	4	1	0	4	9	25	1	0	2
2006	2	3	0	0	0	0	6	6	10	0	0	0
2007	19	100	16	1	1	0	0	0	1	0	1	0
2008	3	4	3	1	0	0	0	0	0	0	0	0
2009	1	6	5	3	3	3	12	24	58	22	0	7
2010	103	95	37	3	0	0	0	0	0	0	0	0
2011	14	21	3	0	0	1	2	16	20	0	0	0
2012	16	27	14	1	3	3	31	36	45	4	3	3
2013	62	56	33	14	0	1	19	17	36	13	2	0
2014	15	32	18	17	2	0	48	38	47	3	1	0
2015	18	19	27	4	9	5	11	31	39	12	0	8
2016	40	60	58	12	0	0	5	22	18	1	0	2

Tabla 2: Número de Grandes Incendios Forestales (GIF) en Colombia, 2002 - 2016.

### Estadísticas de resumen

A continuación se muestran algunas estadísticas descriptivas, sobre la serie de homicidios Colombia para los años 1960-2018.

Tabla 3: Estadísticas de Resumen serie homicidios en Colombia.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Homicidios	59	14,189.170	8,012.999	3,908	5,969.5	12,626	20,907	28,837
Tasa	59	40.421	17.111	19.256	27.057	32.359	53.894	77.946

En la tabla el número mínimo de homicidios ocurrido en este período fue de 3908 con una Tasa de 19.26 homicidios por cada 100.000 habitantes, que corresponde al año 1969, mientras que el máximo número de homicidios registrados fue de 28.837 en el año 2002, sin embargo la Tasa más alta de homicidios fue en el año 1991 con casi 78 homicidios por cada 100.000 la más alta de la región para esta época según un estudio que presentó la CEPAL encontró que la tasa promedio homicidios en Latino-américa era de 20 por cada 100.000 habitantes. Algunas investigaciones sobre el tema como la de Franco (2006) y Pécaut (2003) han enfatizado ciertos aspectos coyunturales, tales como el problema del narcotráfico, la persistencia del conflicto armado interno, la debilidad del Estado, la corrupción y la inmadurez en el ejercicio de la ciudadanía pero aun son insuficientes los estudios y poco el consenso sobre las explicaciones

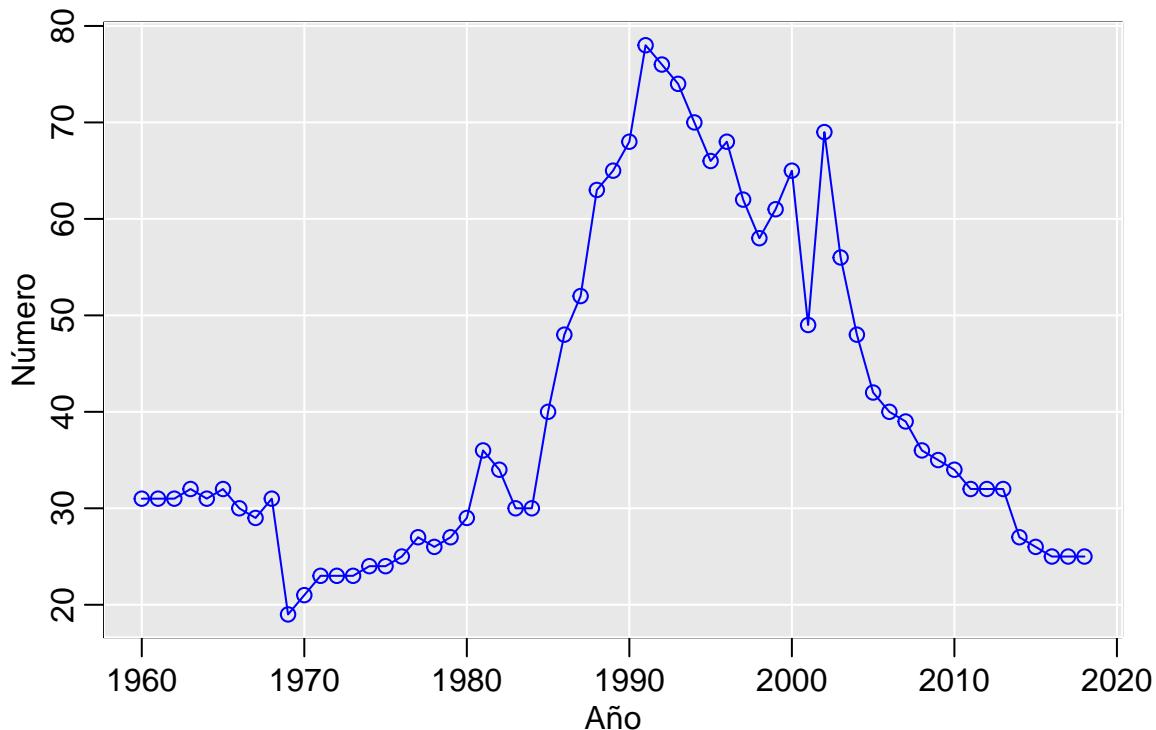


Figura 1: Serie de tiempo homicidios en Colombia desde el año 1960 hasta el año 2018.

de fondo de la situación de violencia que vive el país

En el conjunto de los países con conflictos armados en el mundo, Colombia presenta uno de los más altos índices de homicidio 40/100.000 en estas últimas seis décadas, con cifras comparables a las de países con guerra civil declarada. (Franco, 1980).

En la figura 2, se encuentra gráficamente la densidad de la serie Tasa de homicidios por 100.000 habitantes en Colombia, se deduce que utilizar modelo de regresión Poisson, sería inapropiado pues parece haber una mixtura entre dos distribuciones, ahora la pregunta que deberíamos hacernos es si estas dos distribuciones están correlacionadas, pues de no estarlo una opción para modelar esta serie sería utilizar una mixtura entre dos o más distribuciones independientes, como se muestra en Zucchini (2012, capítulo 1). Por otra parte parece haber una sobredispersión enorme pues mientras la media se sitúa en 40, la varianza es 292 es decir 7 veces la media, y recordemos que para la distribución Poisson  $\mu = \sigma^2 = \lambda$ .

Un primer período de incremento acelerado que va desde comienzos de los 80, en particular desde 1983, hasta 1991. Es la fase más crítica de violencia, en particular de violencia homicida, en los anales de la ciudad. Las tasas de homicidio en la ciudad llegaron a marcar la tendencia de la curva de homicidios a nivel nacional. Investigaciones anteriores **19-22** han tratado de explicar este incremento acelerado mediante la convergencia de los problemas acumulados de debilidad institucional, ausencias estatales, ciudadanía precaria, desempleo e inequidades crecientes, con la expansión del fenómeno del narcotráfico en la ciudad **23** y su confrontación armada estatal, con la intensificación de la presencia urbana del conflicto armado interno, en especial la actuación de las milicias afines a las organizaciones guerrilleras y la emergencia y acelerado desarrollo de organizaciones paramilitares **24,25**.

En la figura 3 se observa la función de autocorrelación muestral para la serie Tasa de homicidios hasta el rezago 30, como se evidencia existe una fuerte dependencia serial en los datos por lo que sería inapropiado utilizar un modelo de mixturas independientes (distribución Poisson), como alternativa surge la utilización de los modelos ocultos de Markov, en este caso se utilizará un PHMM.

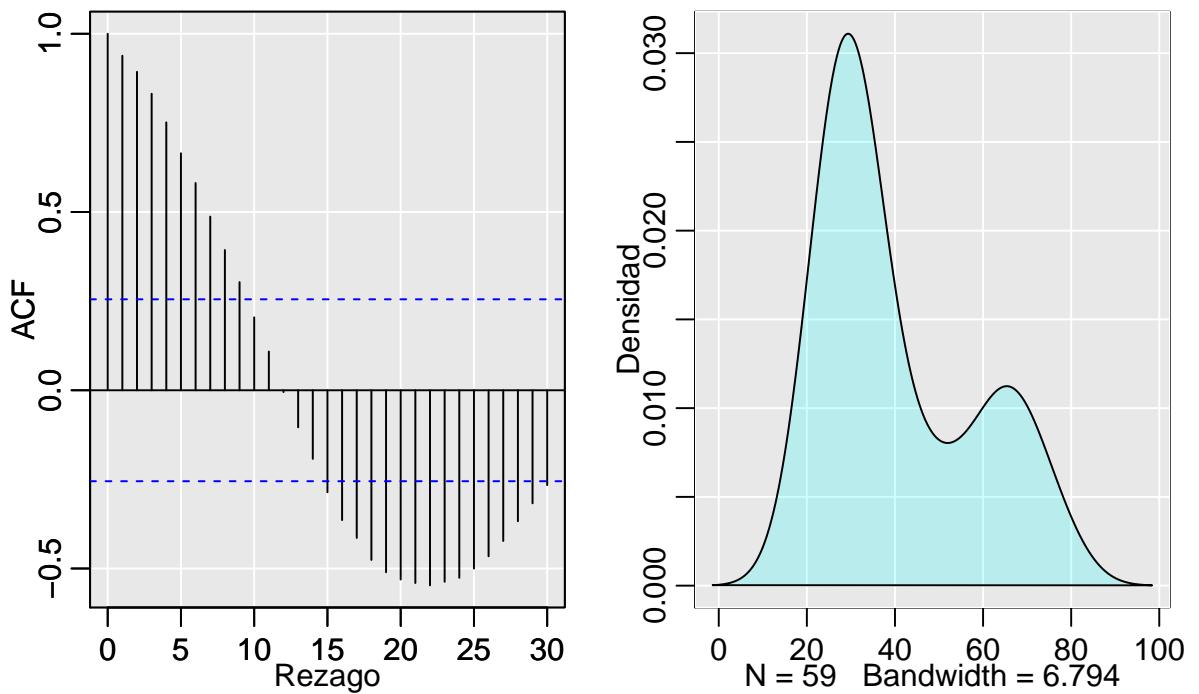


Figura 2: Función de autocorrelación muestral, y kernel de densidad para la serie homicidios en Colombia (1960-2018).

### Ajuste clásico PHMM

Primero ajustamos varios modelos Poisson ocultos de Markov con 1 a 5 estados, y tres modelos con mixturas independientes con 2, 3 y 4 componentes de la distribución Poisson utilizando el paquete **flexmix** de R. Por último registramos los siguientes valores en la Tabla 3, el número de parámetros estimados, la log-verosimilitud el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC). Con el fin de seleccionar el modelo más apropiado, el valor que minimiza el AIC es el PHMM de orden 3 con un valor de 404.02, mientras que el BIC indica que el modelo apropiado es un PHMM de orden 2, con un valor de 418.96. Tanto el BIC y AIC resuelven este problema mediante la introducción de un término de penalización para el número de parámetros en el modelo, el término de penalización es mayor en el BIC que en el AIC. El BIC generalmente penaliza parámetros libres con más fuerza que hace el criterio de información de Akaike, aunque depende del tamaño de  $n$  y la magnitud relativa de  $n$  y  $p$ . Como el tamaño de la muestra es relativamente grande  $n = 59$ , y la cantidad de parámetros que se estiman en un HMM es bastante utilizaremos el BIC en este caso en concreto, eligiendo por tanto el PHMM de orden 2.

	Modelo	p	logL	AIC	BIC
1	PHMM - 1 Estado	1	-356.91	715.81	717.89
2	PHMM - 2 Estados	4	-201.32	410.65	418.96
3	PHMM - 3 Estados	9	-193.01	404.02	422.71
4	PHMM - 4 Estados	16	-190.84	413.69	446.93
5	PHMM - 5 Estados	25	-190.29	430.58	482.51
6	mixtura indep. (2)	3	-229.38	464.75	470.98
7	mixtura indep. (3)	5	-228.11	466.21	476.60
8	mixtura indep. (4)	7	-228.11	472.69	487.23

Tabla 4: Datos homicidios: comparación de modelos ocultos de Markov (estacionarios) por AIC y BIC.

Varios comentarios surgen de la Tabla 4. En primer lugar, dada la dependencia en serie manifestada en la Figura 2, no es sorprendente que los modelos de mezcla independientes no tengan un buen desempeño en

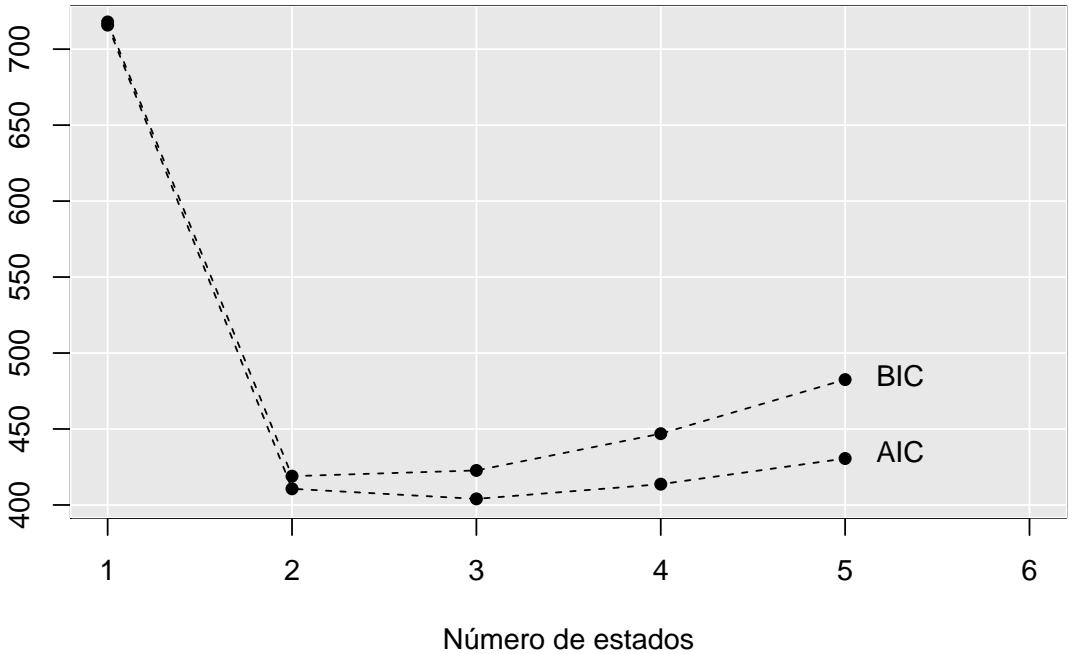


Figura 3: Serie homicidios: selección de modelos AIC y BIC.

relación con los HMM. En segundo lugar, aunque quizás sea obvio a priori que ni siquiera se debe intentar establecer un modelo con un máximo de 16 o 25 parámetros para 59 observaciones, y observaciones dependientes, es interesante explorar las funciones de verosimilitud en el caso de HMM con cuatro y cinco estados. La verosimilitud parece ser altamente multimodal en estos casos, y es fácil encontrar varios máximos locales utilizando diferentes valores de inicio. Una estrategia que parece tener éxito en estos casos es comenzar todas las probabilidades de transición fuera de la diagonal en valores pequeños (como 0.1 o 0.05), mientras que para los valores de las medias estado dependientes se pueden usar los valores de los deciles, calculados a partir de la variable de interés.

La estimaciones del PHMM de dos estados se muestran a continuación, primero la tpm  $A$ , además del vector de medias de los estados dependientes  $\lambda$  y los valores de la distribución estacionaria  $\pi$ .

$$A = \begin{pmatrix} 0.980 & 0.020 \\ 0.064 & 0.936 \end{pmatrix}$$

$$\lambda = (29.715, 62.812) \quad \pi = (0.764, 0.235)$$

Ahora miraremos otras metodologías alternativas a los criterios de información AIC y BIC, que determinan si el modelo tiene un buen ajuste. Entre estas es útil comparar las funciones de autocorrelación de los HMM con dos, tres, cuatro y cinco estados con la función de autocorrelación muestral (ACF). Los ACF de los modelos se pueden encontrar utilizando la función ‘Bayeshmmcts::pois.HMM.moments’ utilizando la ecuación de Zucchini, pág. 55. En forma tabular los ACF se muestran en la tabla 5:

En la Figura 6, de izquierda a derecha se muestran el ACF de las observaciones, la barra de color verde pertenece al modelo de dos estados y la azul al modelo de tres estados. Nos interesa ver como está yuxtapuesto los ACF de ambos modelos con respecto al ACF de las observaciones. Está claro que los ACF del modelo con tres estados corresponden bien con el ACF de las observaciones hasta aproximadamente el rezago 6, mientras que el modelo 2 estados coincide hasta el rezago 9. Sin embargo, se pueden aplicar diagnósticos más sistemáticos, como se mostrará a continuación.

	1	2	3	4	5	6	7	8	9	10	11	12
observaciones	0.94	0.89	0.83	0.75	0.66	0.58	0.49	0.39	0.30	0.20	0.11	-0.00
PHMM 2 Estados	0.77	0.71	0.65	0.59	0.54	0.50	0.46	0.42	0.38	0.35	0.32	0.29
PHMM 3 Estados	0.79	0.75	0.71	0.68	0.64	0.61	0.58	0.55	0.52	0.50	0.47	0.45
PHMM 4 Estados	0.80	0.76	0.72	0.69	0.65	0.62	0.58	0.55	0.52	0.50	0.47	0.44

Tabla 5: Datos homicidios: ACF y ACF de los cuatro modelos hasta el rezago 12.

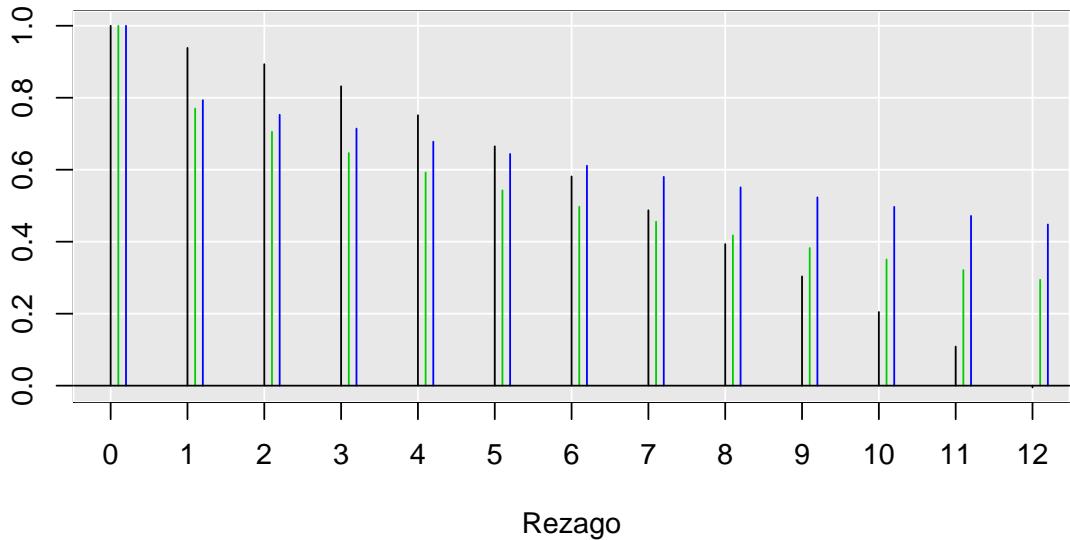


Figura 4: Datos homicidios: ACF y ACF de los PHMM con dos y tres estados.

### Verificación de supuestos del PHMM

En este caso hemos elegido el BIC como criterio para la selección del mejor modelo como mostramos anteriormente, sin embargo sigue existiendo el problema de decidir si el modelo es realmente adecuado; por lo tanto se necesitan herramientas para evaluar la bondad general del ajuste del modelo e identificar valores atípicos en relación con el modelo. En el contexto más simple como por ejemplo los modelos de regresión (teoría normal), el papel que juegan los residuales como herramienta para la verificación del supuesto del modelo está muy bien establecido, entre estos supuestos están la normalidad de los residuales, la homocedasticidad y la independencia de estos. Los pseudo-residuos (también conocidos como residuos quantílicos) que se ilustraron en la sección tres tienen la intención de cumplir esta función de manera mucho más general, y que son útiles en el contexto de los HMM.

En el gráfico 6, se muestra los pseudo residuales ordinarios del PHMM con 2 estados. La fila superior izquierda muestra el diagramas de índice de los pseudo-residuos normales, con líneas horizontales en 0,  $\pm 1.96$  y  $\pm 2.58$ . En la parte superior derecha se muestra los gráficos de cuantiles-cuantiles de los pseudo-residuos normales, con los cuantiles teóricos en el eje  $x$ . La última fila muestra en la parte izquierda el histograma de los pseudo residuales normales, y en la parte derecha la función de autocorrelación muestral de los pseudo-residuos normales. Efectivamente los pseudo-residuales parecen distribuirse normalmente, sin embargo realizamos la prueba de Shapiro-Wilks para verificar este supuesto, donde el p-valor es 0.7529, por lo tanto no podemos rechazar la hipótesis nula  $H_0$ , y concluimos que hay suficiente evidencia estadística para decir que los pseudo-residuos se distribuyen normalmente con un nivel de confianza del 95 %. Además todos los puntos están dentro de las bandas de confianza, sin embargo el histograma no parece acomodarse en todos sus puntos a la curva de la distribución normal, y el mayor problema es que los pseudo-residuales parecen estar correlacionados, hasta el rezago 3.

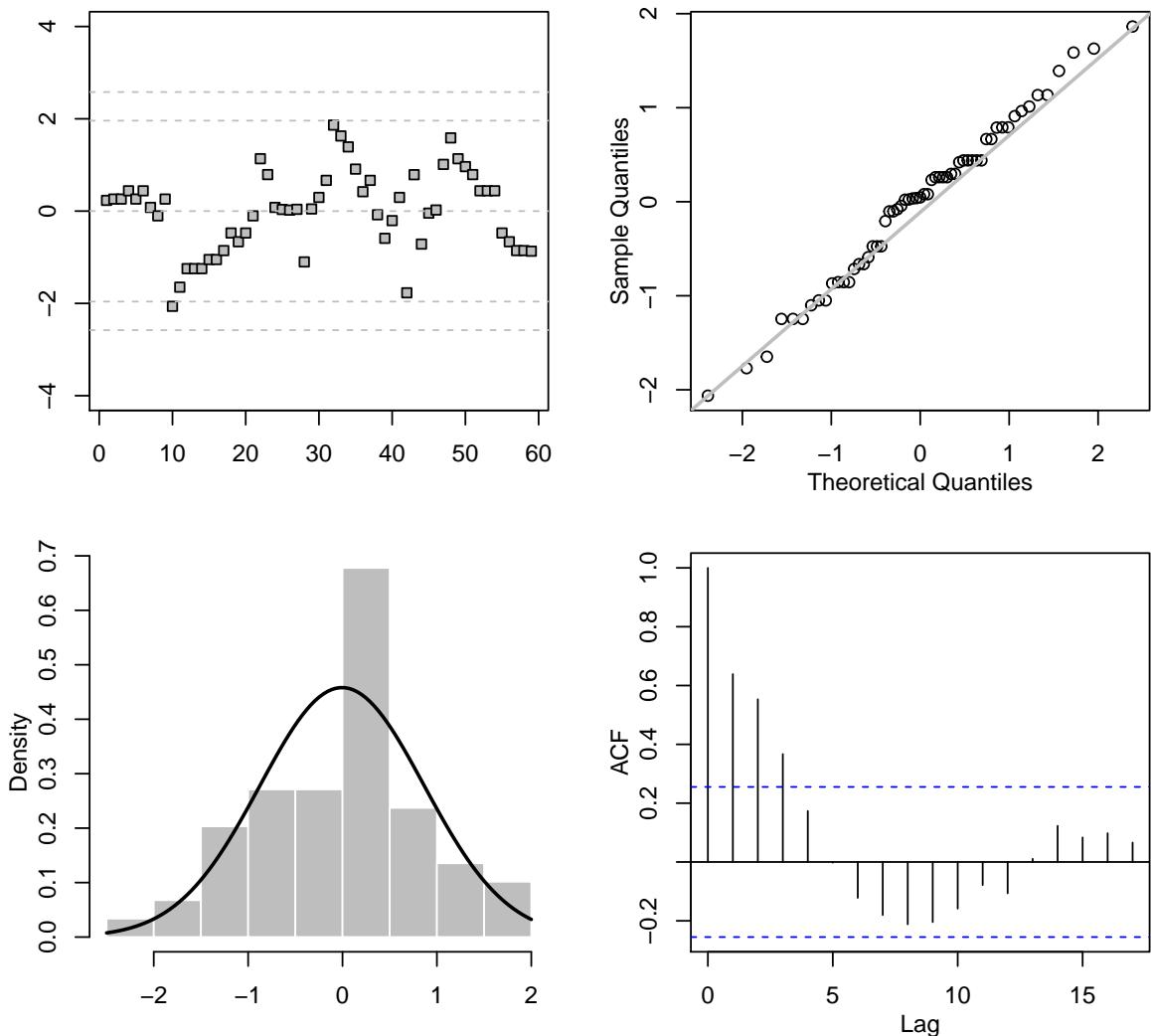


Figura 5: Grafico pseudo-residuales ordinarios para el PHMM de 2 estados.

### Algoritmo Viterbi

El algoritmo Viterbi, permite realizar la decodificación global de los estados clasificando a cada una de las observaciones en su correspondiente estado, indicando la secuencia más probable de los estados ocultos. Para la serie homicidios de 59 observaciones, el algoritmo Viterbi clasificó 40 observaciones en el estado 1 y 19 en el estado 2. En la grafica 5 se visualiza el algoritmo viterbi, y las distribuciones marginales para cada estado.

La decodificación global (algoritmo Viterbi) es el objetivo principal en muchas aplicaciones, especialmente cuando existen interpretaciones importantes para los estados. Sin embargo los estados no observados en el modelo, no siempre necesitan tener interpretaciones sustantiva, pues se consideran artefactos útiles para adaptarse a la heterogeneidad no explicada y la dependencia serial de los datos. En el caso de la serie homicidios no parece haber una interpretación clara de los estados.

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Tabla 6: Resultados de la decodificación global con el algoritmo Viterbi.

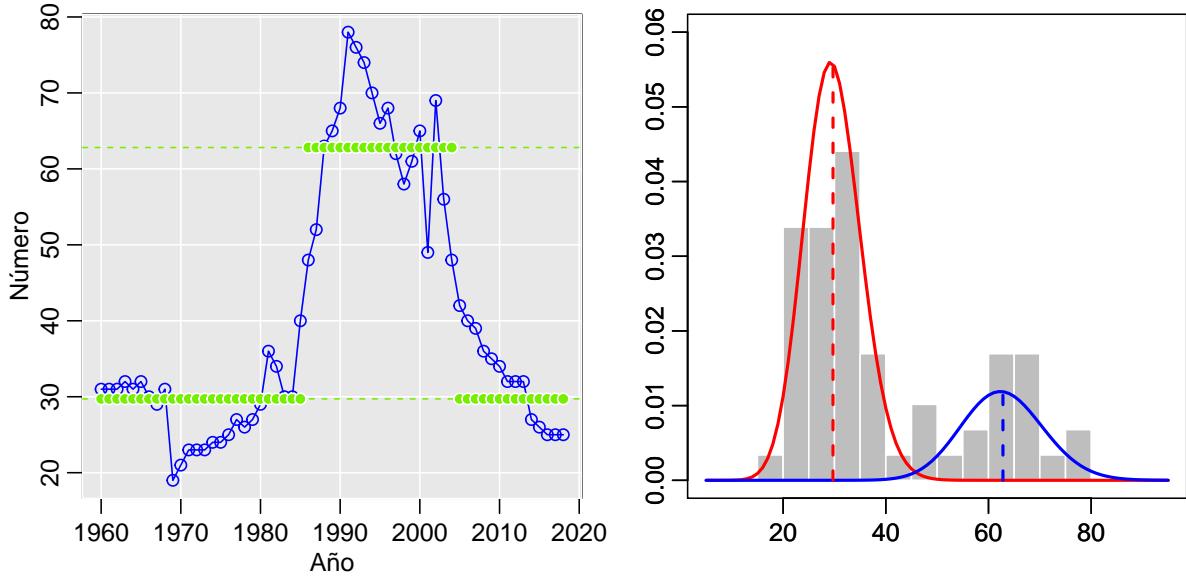


Figura 6: Algoritmo Viterbi aplicado a un PHMM de dos estados.

Se realiza la predicción de los estados más probables para los próximos 16 años, también podemos pronosticar la distribución para estos mismos años. Como se observa en la figura 8 a medida que el horizonte de pronóstico  $h$  aumenta, la distribución de pronóstico converge a la distribución marginal del HMM estacionario. En la tabla 7, se observa que el pronóstico de los estados, para los próximos 16 años es el 1, es decir que se espera una tasa de homicidios por cada 100.000 habitantes cercana a 29, la cual sigue siendo alta ya que según datos de la ONUDD (Oficina de Naciones Unidas contra la Droga y el Delito), en sur America la tasa se sitúa en 20/100.000 homicidios, lo que indica que la tasa de homicidios en Colombia está por encima de la región. Además según estadísticas de la ONUDD, Colombia se sitúa como uno de los países más violentos del mundo ubicándose en el top 20, las cifras de la fiscalía indican que después de haber disminuido la tasa de homicidios en los últimos años, a partir del 2018 hubo un incremento del 3.25 %, de este delito siendo caso críticos las ciudades de Medellín, bajo Cauca y Tumaco, mientras la capital sigue con tendencia a la baja.

Año	Estado 1	Estado 2	Estado
2019	0.9802	0.0198	1
2020	0.9621	0.0379	1
2021	0.9456	0.0544	1
2022	0.9304	0.0696	1
2023	0.9164	0.0836	1
2024	0.9037	0.0963	1
2025	0.8920	0.1080	1
2026	0.8813	0.1187	1
2027	0.8714	0.1286	1
2028	0.8624	0.1376	1
2029	0.8542	0.1458	1
2030	0.8467	0.1533	1
2031	0.8397	0.1603	1
2032	0.8334	0.1666	1
2033	0.8276	0.1724	1
2034	0.8223	0.1777	1

Tabla 7: Predicción para las probabilidades de los estados hasta un rezago  $h = 16$ .

En el siguiente apartado se muestran las estimaciones bayesianas realizadas a la serie homicidios

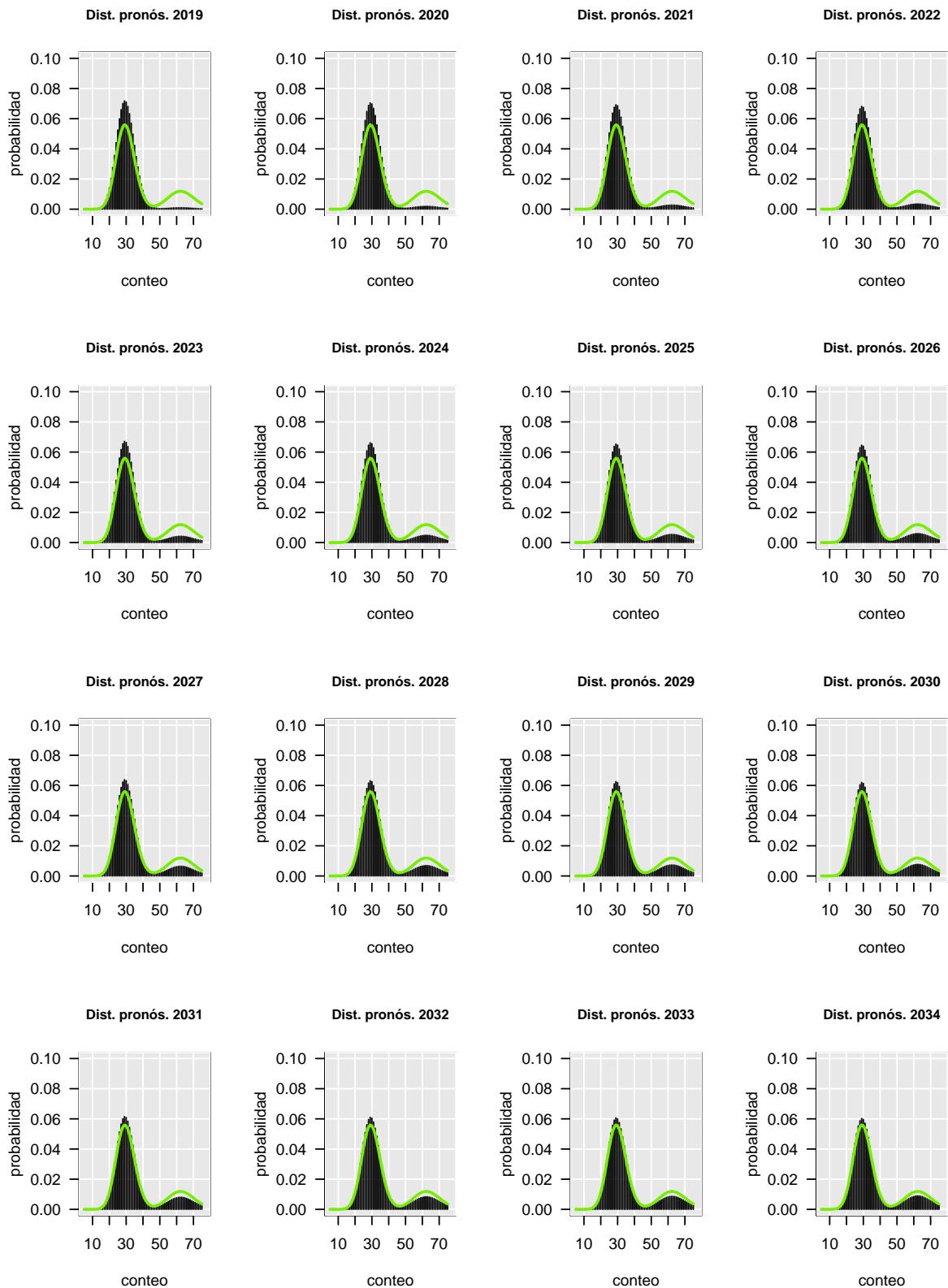


Figura 7: Pronostico de la distribución para los años 2019 a 2034.

## Estimación Bayesiana del PHMM

Primero se ajustaron cuatro modelos, con la función ‘Bayeshmmcts::bayes.PHMM’, para 2, 3, 4 y 5 estados, después, se estima la log - verosimilitud marginal, utilizando muestreo por puente como alternativa a las propuesta hecha por Newton y Raftery (1994) que sugiere utilizar la verosimilitud integrada, para hallar el estimador de la media armónica de los valores de la verosimilitud de una muestra obtenida desde la distribución posterior. Pero como se vio en la sección (4), aunque el estimador es consistente tiene un gran problema varianza infinita. Mientras que el estimador del muestreador por puente, no presenta ese problema además de su fácil implementación, pues esta metodología se puede ejecutar con la función ‘bridge sampler’ del paquete bridgesampling, del autor Gronau. El paquete **bridgesampling**, permite además calcular el error de la estimación para la verosimilitud marginal, obtenido vía muestreo por puente que en el caso del modelo con dos estados, el error es de 0.478 %.

Un factor de Bayes es la relación entre la probabilidad de una hipótesis particular y la probabilidad de otra. Puede interpretarse como una medida de la fuerza de la evidencia en favor de una teoría entre dos teorías en competencia. Esto se debe a que el factor de Bayes nos permite evaluar los datos a favor de una hipótesis nula y utilizar información externa para hacerlo. Nos dice cuál es el peso de la evidencia a favor de una hipótesis dada.

Cuando estamos comparando dos hipótesis,  $H_0$  (la hipótesis nula) y  $H_1$  (la hipótesis alternativa) y , el factor de Bayes a menudo se escribe como  $B_{01}$ . Se puede definir matemáticamente como

$$B_{01} = \frac{\text{verosimilitud de los datos dado } H_0}{\text{verosimilitud de los datos dado } H_1} = \frac{P(D|H_0)}{P(D|H_1)}$$

El factor de Bayes puede ser un número positivo, y una de las interpretaciones más comunes es esta: propuesta por primera vez por Harold Jeffreys(1961) y modificada ligeramente por Lee y Wagenmakers en 2013:

B01	Desición
>100	Evidencia extrema para $H_0$
30 - 100	Evidencia muy fuerte para $H_0$
10 - 30	Evidencia fuerte para $H_0$
3 - 10	Evidencia moderada para $H_0$
1 - 3	Evidencia apenas mencionable para $H_0$
1	No hay evidencia
1/3 - 1	Evidencia apenas mencionable para $H_1$
1/10 - 1/3	Evidencia moderada para $H_1$
1/30 - 1/3	Evidencia fuerte para $H_1$
1/100 - 1/30	Evidencia muy fuerte para $H_1$
< 1/100	Evidencia extrema para $H_1$

Tabla 8: Interpretación del factor de Bayes, Lee y Wagenmakers (2013).

Ahora utilizamos el factor de bayes para contrastar los modelos con m-estados de a parejas, y seleccionar el más adecuado, en la siguiente tabla ilustra el contraste de hipótesis, donde las filas indican  $P(D|H_0)$  y las columnas  $P(D|H_1)$ . Por ejemplo en el contraste de hipótesis entre el modelo de 3 estados vs el modelo de 4 estados, el valor obtenido fue  $B_{01} = 766.05$ , lo que indica evidencia extrema para  $H_0$ , es decir el modelo de 3 estados es más apropiado que el de 4 estados.

	mod 2 Estados	mod 3 Estados	mod 4 Estados	mod 5 Estados
mod 2 Estados		3.36	2545.85	390147608.00
mod 3 Estados			766.05	125542040.00
mod 4 Estados				128023.00

Tabla 9: Comparación resultados Factor de Bayes para los PHMM.

De la tabla 9, se concluye que el modelo apropiado es el de orden 2, lo cual coincide con el BIC. Se corrieron 5.000 iteraciones con 3 cadenas y las primeras 2.500 iteraciones de calentamiento adelgazando la cadena cada 3 iteraciones; con tasa de aceptación para la función objetivo en el metropolis de 0.99. A

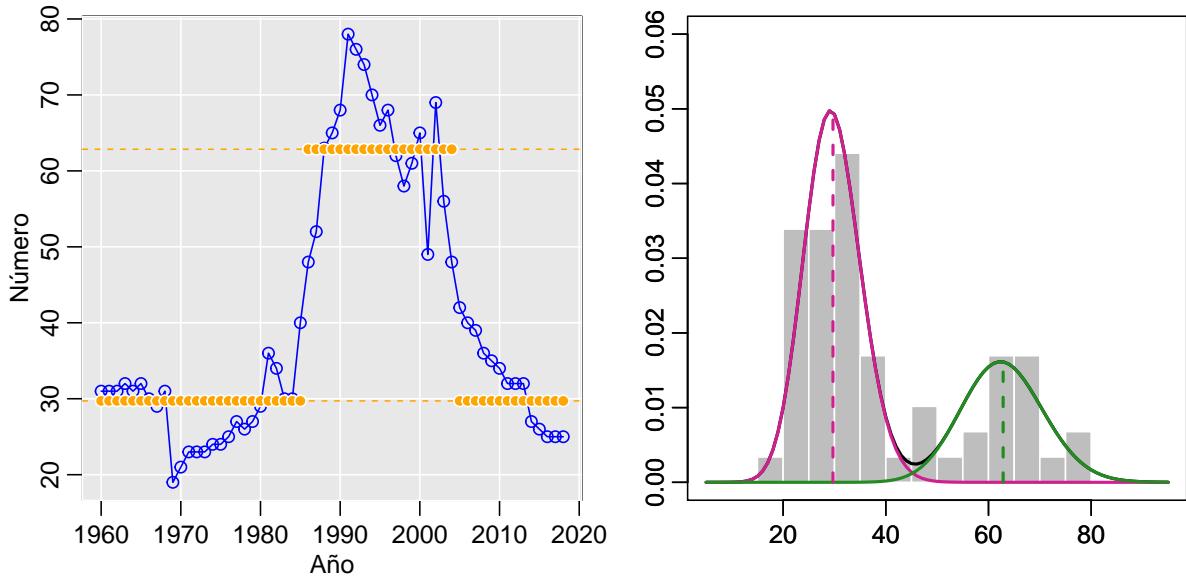


Figura 8: Algoritmo Viterbi aplicado a un PHMM de dos estados.

continuación mostramos las estimaciones bayesianas de la matriz de transición, y la media de los estados dependientes:

	Media	Err.Sta	Desv	2.5 %	25 %	50 %	75 %	97.5 %	n_eff	Rhat
$a_{11}$	0.953	0.001	0.032	0.873	0.935	0.961	0.977	0.994	2491.677	1.000
$a_{12}$	0.047	0.001	0.032	0.006	0.023	0.039	0.065	0.127	2491.677	1.000
$a_{21}$	0.099	0.001	0.065	0.014	0.051	0.084	0.133	0.257	2417.781	1.000
$a_{22}$	0.901	0.001	0.065	0.743	0.867	0.916	0.949	0.986	2417.781	1.000
$\lambda_1$	29.715	0.018	0.871	28.097	29.111	29.684	30.299	31.460	2456.451	1.001
$\lambda_2$	62.849	0.039	1.961	59.068	61.491	62.811	64.184	66.735	2560.484	1.000
lp	-210.558	0.030	1.426	-214.125	-211.268	-210.260	-209.512	-208.739	2200.204	1.002

Tabla 10: Estimación bayesiana de los parámetros para un PHMM.

Para cada parámetro estimado a partir de las muestras obtenidas por MCMC se calculo, la media de las tres cadenas fusionadas. También se calculo el error estándar es aquel el error debido a la estimación de la media poblacional a partir de las medias muestrales. La desviación estándar para este caso indica una dispersión muy pequeña tanto en la estimación de la tpm como del vector de medias de los estados dependientes. Se calculan los intervalos de credibilidad al 95 %, y la mediana de las estimaciones que distan muy poco de la media, lo que indica que en el proceso de muestreo no hubo valores atípicos o extremos. Stan define el logaritmo de la función de densidad de probabilidad de una distribución posterior hasta una constante aditiva desconocida. Usando a lp para representar las realizaciones de este log kernel en cada iteración (lp se trata como una incógnita en el resumen y el cálculo de la división  $\hat{R}$  y el tamaño de muestra efectivo).

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1

Tabla 11: Resultados de la decodificación global bayesiana para el PHMM orden 2, con el algoritmo Viterbi.

Hay dos estadísticas de diagnóstico realmente importantes ocultas en este resumen:

- $n\_eff$ : el tamaño efectivo de la muestra.

- *Rhat*: la “estadística de reducción de escala de potencial de Gelman y Rubin”.

*n\_eff* mide el tamaño de muestra efectivo de ese parámetro en particular. Recuerde que cada iteración de la HMC se basa en el valor del parámetro en la iteración anterior. Sin embargo, idealmente, si el algoritmo funciona correctamente, el parámetro elegido en la siguiente iteración será independiente de ese valor de parámetro inicial (esto es lo que el “delgazamiento” debe lograr en otros MCMC, aunque también puede hacerlo usando HMC). Sin embargo, si no está realizando un trabajo muy eficiente al muestrear el espacio de parámetros, es más probable que los valores de los parámetros en una iteración dada estén cerca de los valores de los parámetros en la última iteración. Esto significa que estos parámetros no son realmente independientes por ejemplo se tiene 1000 muestras obtenidas de la distribución posterior, es posible que no se tenga 1000 muestras independientes del parámetro, sino un número menor de muestras verdaderamente “independientes”.

Entonces, *n\_eff* es la suma de las iteraciones de muestreo efectivamente independientes en todas las cadenas. En este caso, tenemos 3 cadenas, con 5000 iteraciones, la mitad de las cuales son de calentamiento, lo que significa que muestreamos 2500 iteraciones en cada cadena, por lo que el máximo *n\_eff* posible en este caso es 7500.

La mayoría de nuestros parámetros tienen un *n\_eff* bastante alto, aunque vemos que algunos son un poco más bajos. ¿Cómo sabemos si un *n\_eff* como 2500 (de 7500 posibles) es demasiado bajo? El estadístico Rhat nos ayuda a saber si estos parámetros están tan mal muestreados que tenemos un problema. Más o menos Rhat le dice si cada una de las cadenas ha alcanzado o no una distribución posterior estable, a pesar de comenzar con diferentes valores iniciales. Gelman recomienda que Rhat para cada parámetro sea inferior a 1.1.

En la parte izquierda de la figura 8 y en la tabla 11, se muestra la decodificación global de la secuencia de estados más probables, para la serie homicidios en Colombia. Al igual que otros algoritmos de programación dinámica, Viterbi funciona de forma recursiva; encontrando el estado más probable al tomar el máximo sobre todas las posibles secuencias de estados anteriores. Dada la secuencia de observaciones sobre homicidios en Colombia y los modelos clásico con el bayesiano de orden dos, dan exactamente los mismos resultados. Mientras que la parte derecha de la figura 8 muestra las distribuciones marginales, utilizadas para hacer el pronóstico de las distribuciones para un rezago  $h$  dado.

	Estado 1	Estado 2	Estado
2019	0.9533	0.0467	1
2020	0.9134	0.0866	1
2021	0.8793	0.1207	1
2022	0.8502	0.1498	1
2023	0.8253	0.1747	1
2024	0.8040	0.1960	1
2025	0.7859	0.2141	1
2026	0.7703	0.2297	1
2027	0.7571	0.2429	1
2028	0.7457	0.2543	1
2029	0.7361	0.2639	1
2030	0.7278	0.2722	1
2031	0.7207	0.2793	1
2032	0.7147	0.2853	1
2033	0.7095	0.2905	1
2034	0.7051	0.2949	1

Tabla 12: Predicción bayesiana para las probabilidades de los estados hasta un rezago  $h = 16$ .

La tabla 12 muestra, las probabilidades correspondientes a la predicción de los rezagos para un  $h \in N$ . El error de la predicción aumenta a medida que crece el horizonte en el tiempo, por ejemplo para los dos próximos años 2019 y 2020, la probabilidad de estar en el estado 1 es mayor al 90% mientras que para los años 2033 y 2034, la probabilidad de estar en el estado 1 se reduce a un 70%, sin embargo en los próximos 16 años se espera que la tasa de homicidios esté alrededor 30 muertes por cada 100.000 habitantes.

## Diagnosticos de la cadena

En esta sección se verificará el diagnóstico de convergencia de las cadenas utilizadas en la extracción de las muestras. Para los métodos MCMC ajustados con **Stan**, ya sean Hamiltonian Monte (HMC) o No-U-Turn-Sampler (NUTS), el paquete *bayesplot* y *coda*, cuenta con una serie de herramientas gráficas y pueblas diagnosticas para después del ajuste de modelos bayesianos. En la figura No se muestra los histogramas univariados y diagramas de dispersión bivariados para los parámetros de la matriz de transición de probabilidad y para el vector de medias de los estados dependientes, especialmente útil para identificar la colinealidad entre variables (que se manifiesta como gráficos bivariados estrechos), así como la presencia de no-identificabilidad multiplicativa (formas tipo plátano).

En sentido estricto, la no identificabilidad significa que dos valores de los parámetros dan como resultado la misma distribución de probabilidad de los datos observados. Algunas veces también se usa para cubrir situaciones en las que no hay un máximo local único de la densidad posterior, ya sea porque hay múltiples máximos separados o porque hay una meseta donde un conjunto de puntos tiene la misma densidad posterior (estos pueden o pueden No ser identificable en sentido estricto).

En la figura 9 se observa que no parece haber problemas con la identificabilidad, es decir que no existen problemas que señalan divergencias, lo único que se observa es colinealidad entre los parámetros de las filas de la matriz de transición, sin embargo recordemos por definición que la suma de las filas de la tpm suman 1, por lo tanto están de por si correlacionadas. Por lo tanto como el modelo es identificable, no estamos asegurando que las inferencias no estén sesgadas.

El gráfico de traza, muestra por cada una de las iteraciones los valores muestreados correspondiente a una o más cadenas de Markov, separado por parámetro. Las cadenas proporcionan una forma visual para inspeccionar el comportamiento de muestreo y evaluar la mezcla a través de las cadenas y la convergencia, como vemos se comportan bastante bien, pues hay un mínimo de muestras divergentes.

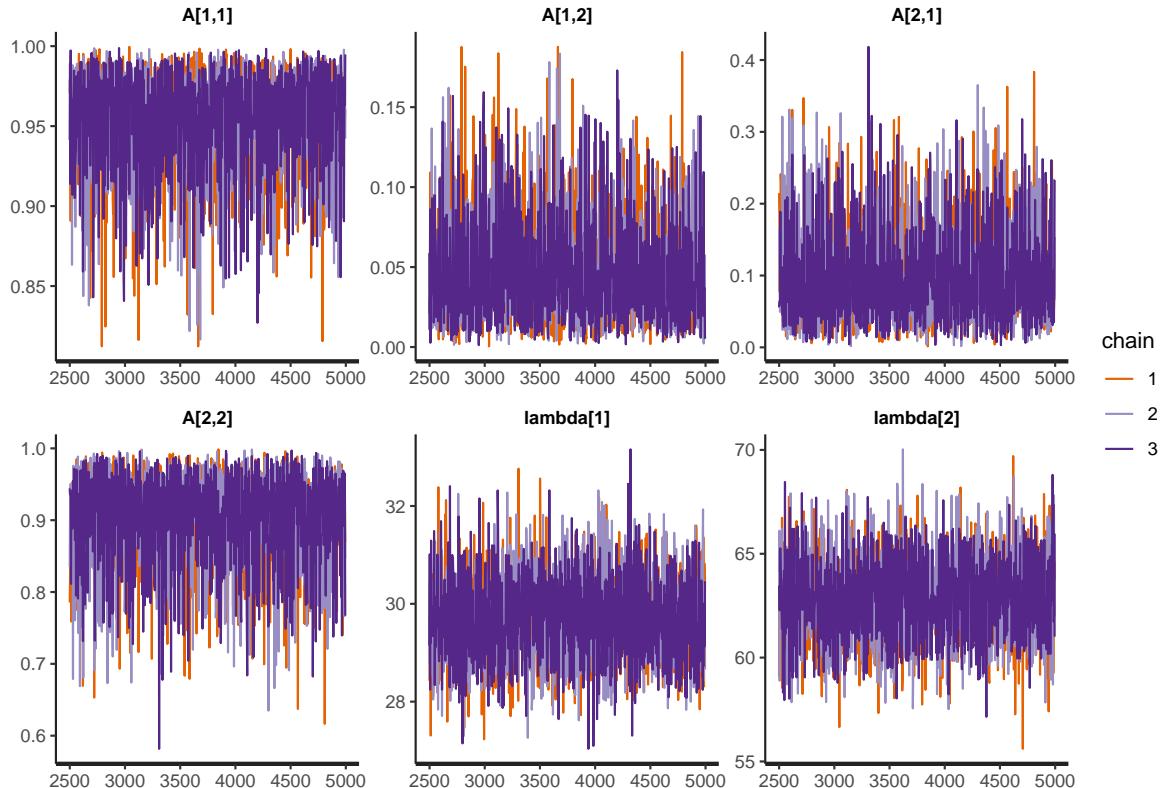


Figura 9: Gráfico de trazas de las cadenas, para cada iteración y por cadena.

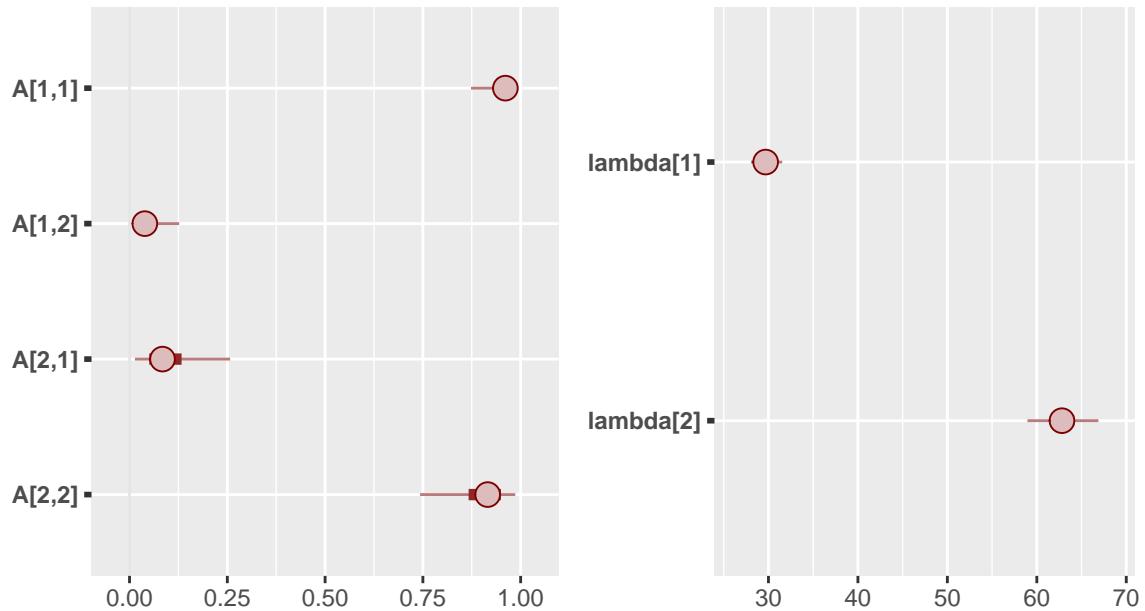


Figura 10: Intervalos de credibilidad al 0.95 PHMM.

Los intervalos de credibilidad, para los parámetros calculados a partir de las muestras posteriores con todas las cadenas fusionadas. Los resultados indican valores consistentes en las estimaciones de los parámetros, pues la longitud del intervalo es bastante pequeña como se mostrara más adelante.

La prueba de convergencia utiliza la estadística de Cramer-von-Mises para probar la hipótesis nula de que los valores muestrados provienen de una distribución estacionaria. La prueba se aplica sucesivamente, primero a toda la cadena, luego, después de descartar el primer 10 %, 20 %, ... de la cadena hasta que se acepte la hipótesis nula, o se haya descartado el 50 % de la cadena. El último resultado constituye un *fallo* de la prueba de estacionariedad e indica que se necesita una ejecución MCMC más larga. Si se pasa la prueba de estacionariedad, se informa el número de iteraciones a mantener y el número a descartar.

La prueba de medio ancho calcula un intervalo de confianza del 95 % para la media, utilizando la parte de la cadena que pasó la prueba de estacionariedad. La mitad del ancho de este intervalo se compara con la estimación de la media. Si la relación entre la mitad del ancho y la media es menor que *eps*, se pasa la prueba de la mitad del ancho. De lo contrario, la longitud de la muestra no se considera lo suficientemente larga como para estimar la media con suficiente precisión.

	P. Estacionariedad	Valor p	Prueba	Media	Medio.Ancho
$a_{11}$	paso	0.396	paso	0.953	0.001
$a_{21}$	paso	0.978	paso	0.099	0.002
$a_{12}$	paso	0.396	paso	0.047	0.001
$a_{22}$	paso	0.978	paso	0.901	0.002
$\lambda_1$	paso	0.569	paso	29.701	0.034
$\lambda_2$	paso	0.862	paso	62.742	0.079
$lp$	paso	0.440	paso	-210.525	0.062

Tabla 13: Prueba de estacionariedad, usando el estadístico de Cramer-von-Mises para la convergencia de la cadena y prueba de medio ancho para la media calculando el intervalo de confianza al 0.95.

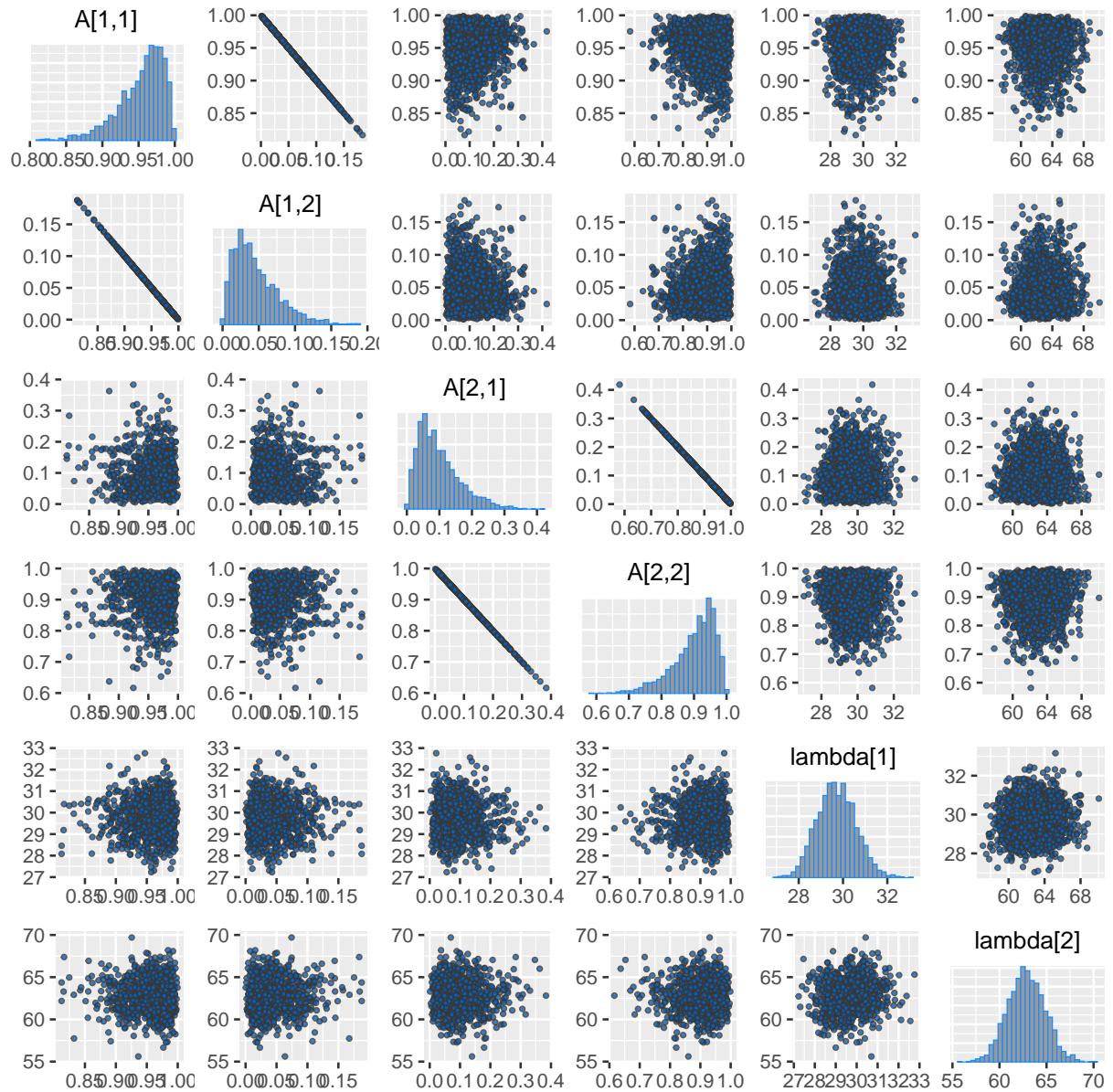


Figura 11: Gráfico de dispersión para las muestras MCMC.

### 1.1.1. Comparación PHMM clásico vs Bayesiano

La inferencia para los parámetros bajo el enfoque clásico se realizó utilizando bootstrap (Zucchini (2016)). El método bootstrap es una técnica de remuestreo diseñada para aproximar la función de distribución de probabilidad de los datos mediante una función empírica de una muestra finita. El método bootstrap se puede usar para estimar los intervalos de confianza directamente. Se utilizó el “método de percentil” (Efron y Tibshirani, 1993) para estimar los intervalos, se generaron 250 muestras independientes a partir del PHMM de orden 2 de longitud 59 igual a la serie homicidios en Colombia. Los valores iniciales usados fueron los estimados por PHMM de 2 estados con el fin de evitar inestabilidad numérica o problemas de convergencia. Los intervalos de credibilidad fueron calculados a partir de las distribuciones posteriores de los parámetros de las muestras generadas por MCMC. El nivel y la probabilidad de los intervalos de confianza y credibilidad, respectivamente, se fijaron en 0.95.

Parámetros	Intervalos de Credibilidad				Intervalos de Confianza			
	Media	2.5	97.5	Ancho	Media	2.5	97.5	Ancho
$a_{11}$	0.953	0.873	0.994	0.120	0.980	0.844	1.000	0.156
$a_{21}$	0.099	0.014	0.257	0.244	0.064	0.015	1.000	0.985
$a_{12}$	0.047	0.006	0.127	0.120	0.020	0.000	0.156	0.156
$a_{22}$	0.901	0.743	0.986	0.244	0.936	0.000	0.985	0.985
$\lambda_1$	29.715	28.097	31.460	3.363	29.716	27.689	31.648	3.959
$\lambda_2$	62.849	59.068	66.735	7.667	62.813	30.140	68.497	38.357

Tabla 14: Intervalos de Credibilidad y Confianza para el PHMM de orden 2.

Al calcular los intervalos de confianza y de credibilidad, es importante determinar cuál de estos métodos son más eficaces. Para determinar el comportamiento de los intervalos propuestos, usualmente se utiliza, la longitud del intervalo, su probabilidad de cobertura el valor esperado y la varianza de su longitud. Un buen método debe tener valores pequeños en la longitud del intervalo, en su valor esperado y en la varianza de su longitud; con probabilidades de cobertura cercanas a los niveles de confianza nominal. La longitud del intervalo, que indica su precisión, se muestran en la tabla 14, junto con la media de las estimaciones en el caso Bayesiano y el estimador de máxima verosimilitud para el caso clásico. Tanto para los parámetros de la matriz de transición como para el vector de medias de los estados dependientes, los intervalos de credibilidad indican una longitud menor es decir mayor precisión. Por lo que en este caso podríamos decir que las estimaciones bayesianas son más precisas y por lo tanto el enfoque bayesiano parece ser el más apropiado.

Finalmente, aunque el intervalo de credibilidad difiere de la interpretación del intervalo de confianza, permite juzgar la incertidumbre estadística para la tasa de homicidios suponiendo el PHMM subyacente válido. Mientras el intervalo de confianza indica que el 95 % de los intervalos de confianza generado por un mismo procedimiento incluyen el verdadero valor del parámetro. El intervalo de credibilidad representa con una probabilidad del 95 % que el intervalo incluya el verdadero valor de la población objetivo siempre que el modelo adoptado sea válido.

## 1.2. Modelo Poisson Cero inflado - Oculto de Markov

En esta sección utilizaremos, los datos de incendios forestales en Colombia, desde enero del 2001 hasta diciembre del 2016. La variable de interés es el número de grandes incendios forestales (GIF), que son aquellos incendios que superan las 500 hectáreas forestales afectadas. La periodicidad de los datos es mensual con un total 180 observaciones, en la Tabla 11 se muestran los primeros 12 registros, mientras que la Tabla 12 indica la frecuencia. Allí observamos que hay una alta proporción de ceros en los datos, pues de las 180 observaciones 124 son cero, es decir el 68.9 % de los registros. Por otra parte el número máximo de GIF ocurridos en un mes en Colombia fue 23 en febrero del 2017, lo cual es preocupante; pues aunque los incendios forestales naturales han ocurrido desde siempre como un elemento normal en el funcionamiento de los ecosistemas. El fuego ha permitido la regeneración de diversos ecosistemas y la producción de una serie de hábitats en los que distintos organismos pueden prosperar. No obstante notemos que el promedio de GIF se ubico en  $1.3 \pm 3.5$  incendio por mes, haciendo que la enorme proliferación de los incendios a causa de la actividad humana en estas últimas décadas sobrepasa la capacidad de recuperación natural.

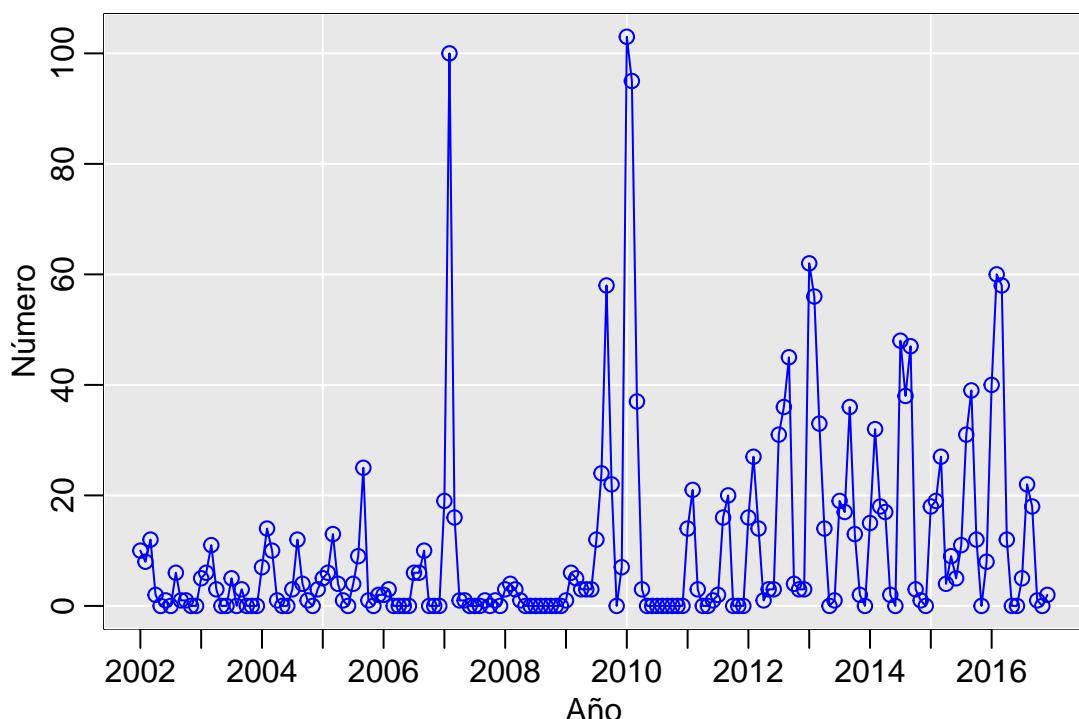


Figura 12: Serie de tiempo Grandes Incendios Forestales en Colombia desde el año 2002 hasta el año 2016.

En la figura 12, se observa dos picos altos en el 2007 y el 2010. Después del año 2011 la cero inflación disminuye considerablemente y el número de incendios en gran parte de los meses parece estar por encima de 5, este fenómeno se presenta de manera recurrente en gran parte del país, en especial durante los períodos secos prolongados, durante los cuales los ecosistemas tropicales húmedos y muy húmedos pierden parte de los contenidos de humedad superficial e interior, incrementando sus niveles de susceptibilidad y amenaza hacia la combustión de la biomasa vegetal que los compone. En la tabla 2 se encuentran todos los datos de GIF en Colombia.

Para determinar si existe correlación entre los GIF de cada mes, se calcula la función de autocorrelación muestral, la figura 14 indica no solo la existencia de la dependencia serial sino una estructura estacional.

Como se vio en la figura 14, existe dependencia serial entre los GIF mensuales ocurridos de Colombia, además parece haber una estructura estacional entre los meses donde ocurrieron estos incendios. Además la densidad estimada por kernel, muestra diferentes picos concentrándose los valores principalmente en cero.

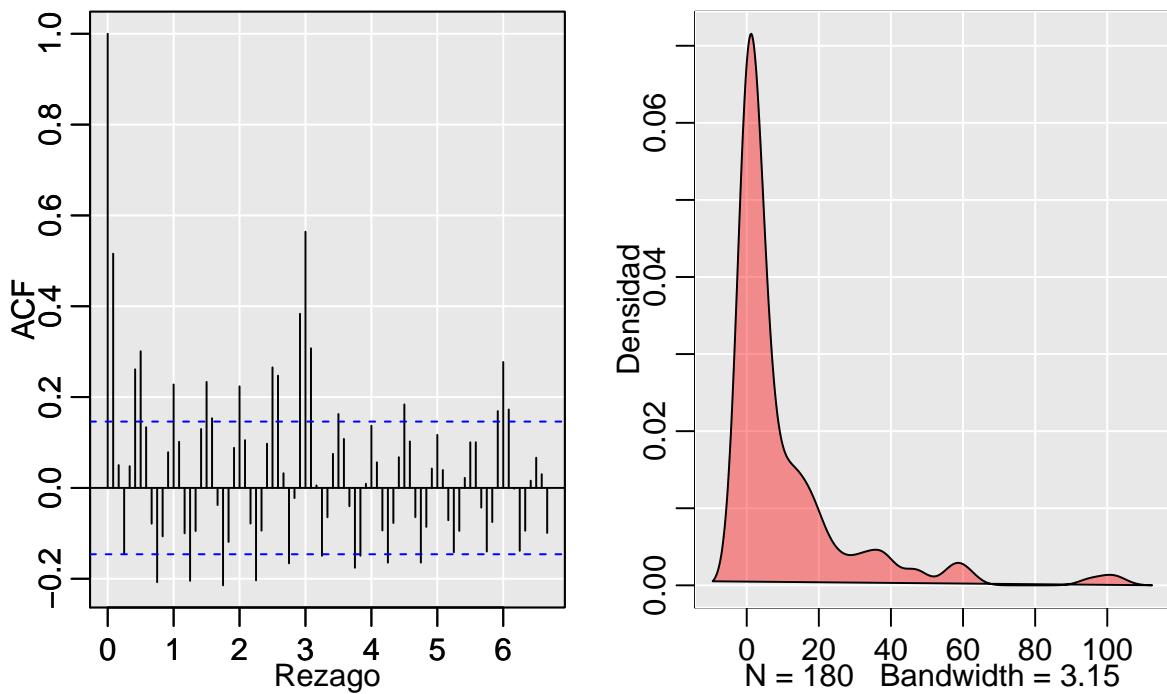


Figura 13: Función de autocorrelación muestral, y kernel de densidad para la serie Grandes Incendios Forestales en Colombia (2002-2016).

### 1.2.1. Ajuste del ZIP-HMM

Se ajustaron seis modelos ZIP-HMM con 2 a 6 estados, utilizando el paquete **ziphsmm** creado por Zekun Xu, que permite ajustar modelos Poisson cero Inflados - Ocultos de Markov, estimando los parámetros vía directa minimización de la función log verosimilitud usando el algoritmo descenso del gradiente. Se utilizó el método de Nelder-Mead con 1.000 iteraciones con el fin de evitar máximos locales. En la tabla 15 se registró para cada modelo el número de parámetros estimados, la log-verosimilitud el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC).

Para seleccionar el modelo más apropiado, se debe escoger el valor que minimiza el AIC, en este caso es el ZIP HMM de 6 estados con un Akaike de 1101, sin embargo el criterio de Schwarz con un valor de 1176 indica que el modelo más apropiado es el de orden 4, esta es una dicotomía que puede presentarse en ocasiones. Sin embargo decidimos escoger como criterio el BIC para la selección del modelo por dos razones principalmente. Primero porque el BIC generalmente penaliza parámetros libres con más fuerza de lo que lo hace el Akaike, y segundo porque para calcular el modelo de 6 estados es necesario calcular el doble de parámetros con respecto al de 4, haciéndolo más costoso computacionalmente, pues con cada estado adicional el número de parámetros a estimar crece de manera sustancial.

	Modelo	p	logL	AIC	BIC
1	ZIP HMM - 2 Estados	6	764.61	1541.23	1560.39
2	ZIP HMM - 3 Estados	12	592.81	1209.62	1247.94
3	ZIP HMM - 4 Estados	20	536.07	1112.15	1176.01
4	ZIP HMM - 5 Estados	30	521.87	1103.74	1199.53
5	ZIP HMM - 6 Estados	42	508.60	1101.20	1235.31
6	ZIP HMM - 7 Estados	56	510.79	1133.58	1312.39

Tabla 15: Datos incendios: comparación de modelos ocultos de Markov (Cero inflados) por AIC y BIC.

En la figura 14 se puede visualizar de una manera más clara el cambio en las magnitudes de los criterios de información, para los modelos ZIP HMM con diferentes estados. Para el Akaike no parece

haber diferencias importantes entre los modelos de 4, 5 o 6 estados, mientras que el BIC sugiere que los modelos de 4 o 5 estados serían los más apropiados.

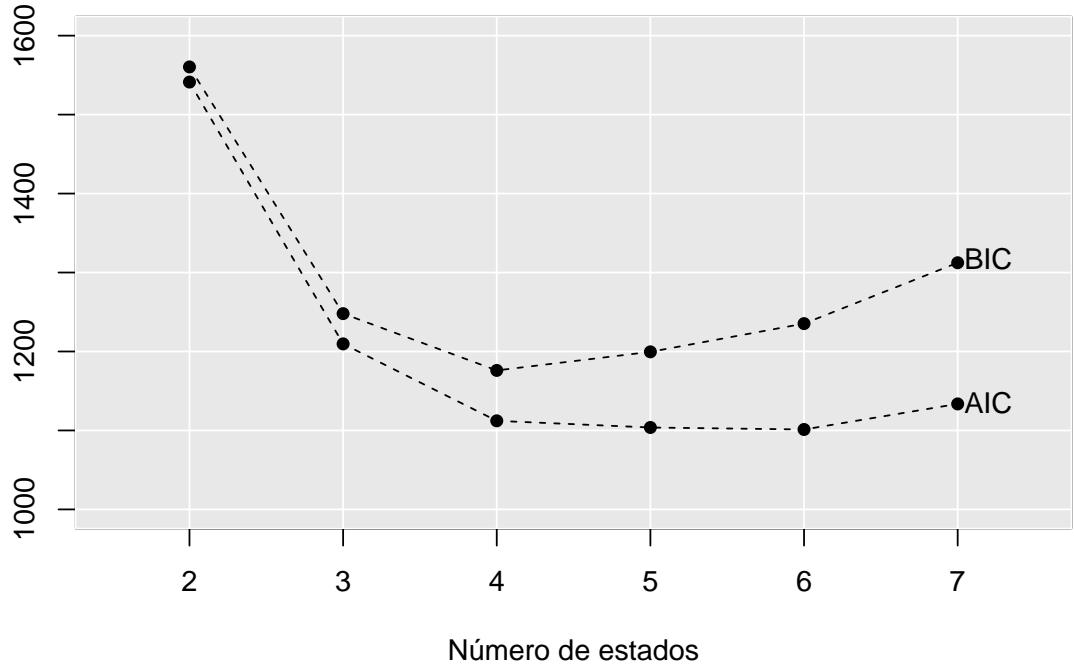


Figura 14: Serie incendios: selección de modelos AIC y BIC.

A continuación se muestran las estimaciones de la matriz de transición de probabilidad  $A$  para el ZIP HMM de orden 4, junto con el vector de medias de los estados dependientes  $\lambda$ , la distribución estacionaria  $\pi$  y el parámetro de proporción de cero inflación  $\theta$ . De estos resultados observamos que al si se está en el estado 1 y lo más probable es seguir en este mismo estado con un valor de 82 %, mientras lo más improbable es pasar del estado 1 al 4 con apenas un 0.001 %. Para el ZIP HMM estacionario de orden 4, lo más factible es iniciar en el estado 2 con un 97 %, y como indica la tpm después pasar al estado 2 con un 48 % de probabilidad. Asumiendo solo cero inflación en el estado 1,  $\theta$  indica la proporción de cero GIF es de 44 %.

$$A = \begin{pmatrix} 0.820 & 0.154 & 0.025 & 0.001 \\ 0.483 & 0.335 & 0.131 & 0.050 \\ 0.168 & 0.329 & 0.499 & 0.004 \\ 0.004 & 0.329 & 0.346 & 0.320 \end{pmatrix}$$

$$\lambda = (2.763, 15.114, 43.147, 99.306) \quad \pi = (0.002, 0.997, 0.001, 0.000) \quad \theta = 0.4440$$

2	2	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	
1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	4	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	3	2	1	2	4	4	3	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	1	2	2	1	1	1	1	1
2	2	2	1	1	1	3	3	3	1	1	1	3	3	3	2	1	1	2	2	3	2	1	1	2	3	2	2	1	1	1	1	1	1
3	3	3	1	1	1	2	2	2	1	2	1	2	3	3	2	1	2	3	3	2	1	1	2	2	1	1	1	2	2	1	1	1	1

Tabla 16: Resultados de la decodificación global con el algoritmo Viterbi, para el ZIP HMM.

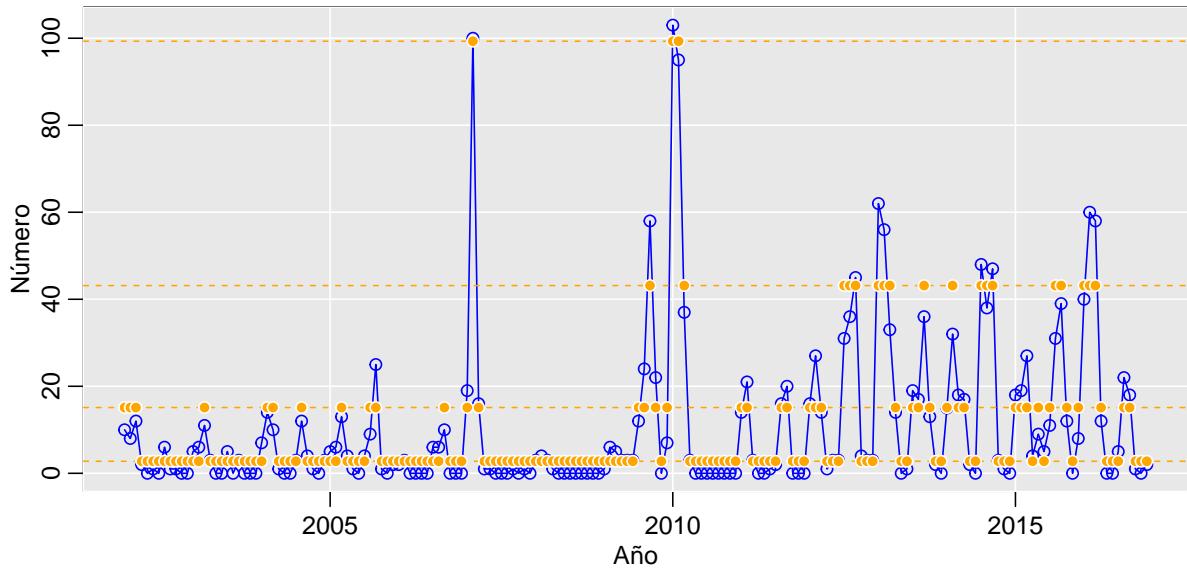


Figura 15: Algoritmo Viterbi aplicado al ZIPHMM de tres estados.

La decodificación global se muestra en la tabla 16 y figura 15. Se observa que el estado 4 con una media de 99, sol hay 3 observaciones correspondientes a los GIF de febrero del 2007, enero del 2010 y febrero del 2010, mientras que el estado 3 con una media de 43 indica que estos 18 incendios la mayoría ocurrieron en tercer semestre, es decir en los meses julio, agosto y septiembre. Para el estado 2 con una media de 15 la mayoría de estos incendios ocurrieron en el primer semestre, finalmente para el estado 1 con una media de 2.8, se observa que en el último trimestre fue donde en su mayoría no ocurrieron GIF. Es así como el algoritmo Viterbi nos indica efectivamente que la serie tiene períodos estacionales.

### Estimación Bayesiana del ZIP HMM

Analogamente al caso de la serie homicidios a la cual se ajustó un PHMM, para la serie de homicidios se determinará el modelo más apropiado utilizando el factor de bayes. Para esto se ajustaron un total de seis modelos con estados 2, 3, 4, 5, 6 y 7. Como vimos anteriormente el factor de Bayes permite evaluar los datos a favor de una hipótesis nula y utilizar información externa para hacerlo. Dando peso de la evidencia a favor de una hipótesis dada.

Utilizando la notación del factor de Bayes que notamos como  $B_{01}$  al contrastar dos hipótesis,  $H_0$  (la hipótesis nula) y  $H_1$  (la hipótesis alternativa), que se definió matemáticamente como:

$$B_{01} = \frac{\text{verosimilitud de los datos dado } H_0}{\text{verosimilitud de los datos dado } H_1} = \frac{P(D|H_0)}{P(D|H_1)}$$

Utilizando otra vez la interpretación propuesta por primera vez por Harold Jeffreys(1961) y modificada ligeramente por Lee y Wagenmakers en 2013, que está en la tabla 8, se deduce lo siguiente.

	mod 2 Est.	mod 3 Est.	mod 4 Est.	mod 5 Est.	mod 6 Est.	mod 7 Est.
mod 2 Est.		0.00	0.00	0.00	0.00	0.00
mod 3 Est.			0.00	0.00	0.00	0.00
mod 4 Est.				0.02	1.78	1513518.00
mod 5 Est.					81.23	84986740.00
mod 6 Est.						956226.00

Tabla 17: Comparación resultados Factor de Bayes para los ZIP HMM.

Recordemos que las filas de la tabla x corresponden a  $H_0$  y las columnas son  $H_1$ . Los resultados

indican que tanto el modelo de 2 como el de 3 estados no son apropiados, pues el valor de 0 indica evidencia extrema para  $H_1$ . Por otra parte los resultados para el modelo de 4 estados, indican que es más apropiado que el de 6 y 7 estados, más no que el de 5 estados. Finalmente los modelos de 5 y 6 estados resultaron vencedores en sus contrastes. Lo que finalmente se concluye que el ZIP HMM más apropiado es el de 5 estados y en segundo lugar el ZIP HMM de orden 4.

Con el fin de comparar los resultados obtenidas de las estimaciones del ZIP HMM clásico vs el bayesiano, se decide ajustar el modelo con 4 estados. La salida que arroja Stan, en primer lugar la media de las estimaciones para los 21 parámetros más lp, el error y la desviación estándar que para este caso son bastante pequeños, seguidos de los intervalos de credibilidad alrededor de la media y a mediana que es casi identica a los valores de la media especialmnte para  $\theta$  el parámetro de cero inflación y para el vector de medias, en la tpm varia ligeramente en algunos de los parámetros. Para este modelo se ajustaron 2.000 iteraciones con tres cadenas, la mitad de ellas se queman como calentamiento es decir que el número máximo de muestras efectivas debiera ser de 3.000, sin embargo ocurre algo extraño en varios de los parámetros  $n\_eff > N$ . Según el manual de Stan esto significa que las muestras que prproduce Stan son mejores que las muestras independientes para esos parámetros, o en otras palabras el muestreo realizado por NUTS es súper eficiente, antitético y con sobre relajación (Geyer, 2011), esto ocurre porque en cada iteración es eliminada la correlación entre las muestras lo cual ocurre en casos extremadamente raros, para más información revise el manual o foro de Stan. Finalmente Gelman indica que el tamaño de muestra efectivo utilizados es el adecuado si Rhat es menor a 1.1, es decir que en este caso el muestreo fue óptimo.

	Media	Err.Sta	Desv	2.5 %	25 %	50 %	75 %	97.5 %	n.eff	Rhat
$\theta$	0.449	0.001	0.050	0.351	0.414	0.449	0.483	0.545	5219.355	0.999
$\lambda_1$	2.868	0.004	0.266	2.371	2.684	2.866	3.047	3.408	4810.999	0.999
$\lambda_2$	15.369	0.012	0.775	13.859	14.830	15.361	15.874	16.877	4492.668	0.999
$\lambda_3$	43.153	0.040	1.807	39.600	42.040	43.143	44.300	46.658	2036.872	1.001
$\lambda_4$	99.235	0.145	6.120	87.428	95.445	99.219	103.236	110.912	1790.696	1.001
$a_{11}$	0.795	0.001	0.037	0.719	0.772	0.797	0.821	0.861	4587.670	0.999
$a_{12}$	0.158	0.000	0.034	0.098	0.134	0.156	0.180	0.231	4676.747	0.999
$a_{13}$	0.034	0.000	0.017	0.009	0.022	0.032	0.044	0.076	4870.506	1.000
$a_{14}$	0.013	0.000	0.011	0.000	0.005	0.010	0.018	0.040	3815.711	0.999
$a_{21}$	0.470	0.001	0.079	0.318	0.415	0.469	0.523	0.630	5238.938	0.999
$a_{22}$	0.328	0.001	0.077	0.192	0.272	0.325	0.380	0.482	5096.108	0.999
$a_{23}$	0.146	0.001	0.058	0.053	0.104	0.140	0.183	0.274	4306.099	1.000
$a_{24}$	0.056	0.001	0.037	0.008	0.029	0.049	0.076	0.147	4611.951	1.000
$a_{31}$	0.188	0.001	0.082	0.056	0.127	0.179	0.238	0.376	5736.173	1.000
$a_{32}$	0.331	0.001	0.099	0.151	0.261	0.325	0.396	0.537	5985.893	0.999
$a_{33}$	0.436	0.001	0.100	0.249	0.367	0.434	0.504	0.634	5179.787	1.000
$a_{34}$	0.045	0.001	0.043	0.001	0.013	0.032	0.063	0.161	3481.429	1.000
$a_{41}$	0.137	0.002	0.119	0.004	0.045	0.102	0.200	0.435	4984.285	0.999
$a_{42}$	0.309	0.002	0.163	0.055	0.182	0.287	0.421	0.658	5002.104	0.999
$a_{43}$	0.280	0.002	0.162	0.041	0.155	0.253	0.388	0.639	4855.509	1.000
$a_{44}$	0.274	0.002	0.157	0.037	0.152	0.252	0.376	0.630	4757.794	0.999
lp	-565.076	0.096	3.098	-572.077	-566.969	-564.699	-562.881	-559.905	1032.526	1.007

Tabla 18: Estimación bayesiana de los parámetros para un ZIPH-MM.

El gráfico de traza proporcionan una forma visual para inspeccionar el comportamiento de muestreo en cada uno de los parámetros de forma independiente y evaluar la mezcla a través de las cadenas y la convergencia, los resultados obtenidos para los 21 parámetros indican que se comportan de forma estable, ya que los valores muestreados en su mayoría se encuentran n un rango dee valores alrededor de la media. Dado el caso de la no convergencia el algunos casos la solución será aumentar el número de muestras.

En la estadísticas bayesianas, un intervalo creíble es un intervalo dentro del cual un valor de parámetro no observado cae con una probabilidad subjetiva particular. Es un intervalo en el dominio de una distribución de probabilidad posterior o una distribución predictiva. El intervalo e credibilidad es el equivalentebayesiano del intervalo de confianza no obstante, este depende de una distribución prior. Otra diferencia importante es que mientras en el intervalo de confianza se trata el parámetro como un valor fijo y los límites son variables aleatorias; en los intervalos creíbles, el parámetro estimado se trata como

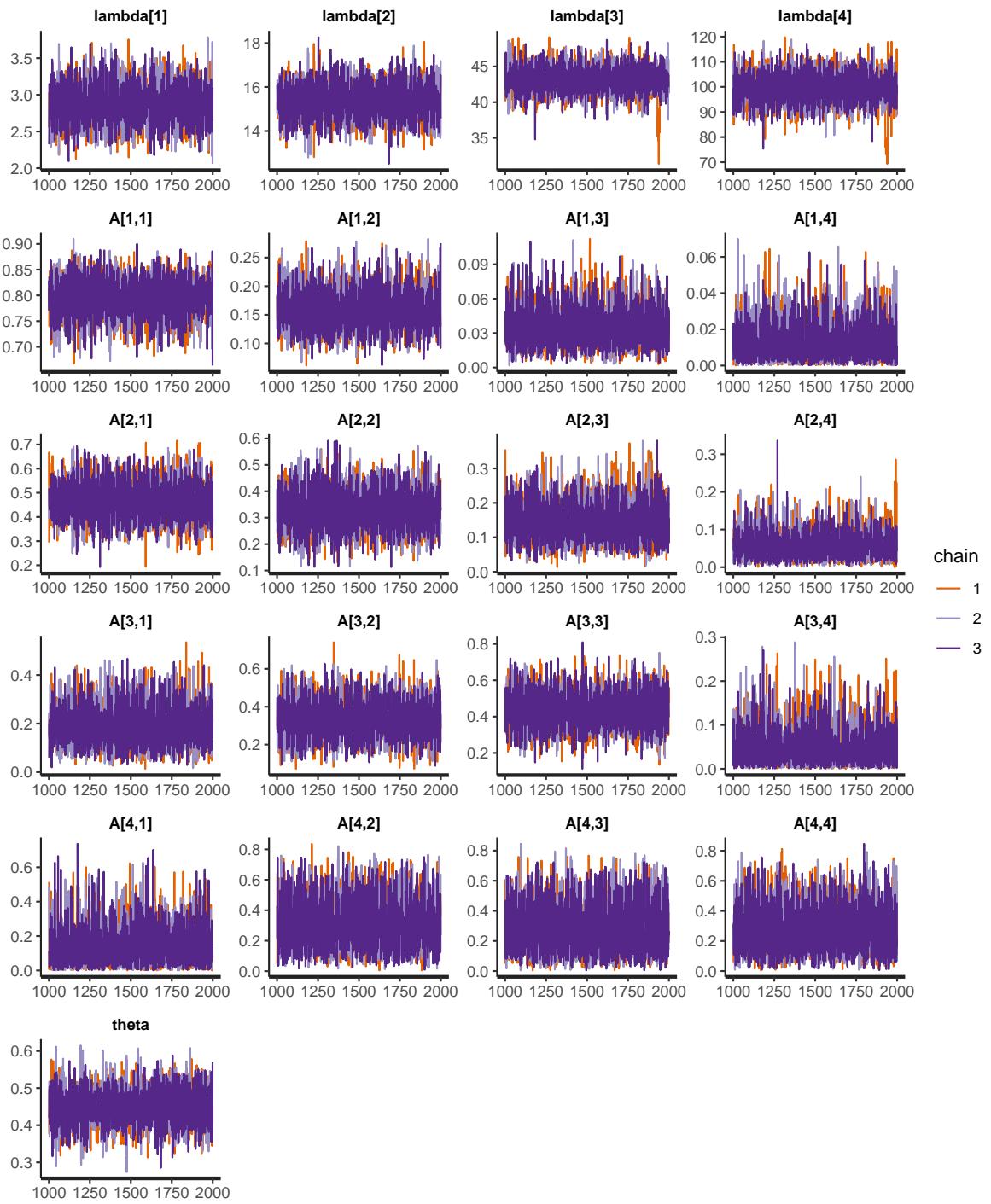


Figura 16: Gráfico de trazas de las cadenas, para cada iteración y por cadena.

una variable aleatoria mientras que los límites se consideran fijos.

Los intervalos de credibilidad al 95 % calculados para la estimación de los parámetros del ZIP HMM, se encuentran en la figura 18. En la gráfica de la izquierda se encuentra el parámetro de cero inflación  $\theta$ , junto con las entradas de la matriz de transición  $\gamma_{ij}$ , ya que todos estos se encuentran en la misma escala, es decir se mueven entre cero y uno, mientras que el vector de medias  $\lambda > 0$ , se dibuja en el grafico de la derecha. Se encontraron intervalos bastante compactos, es decir que la longitud del intervalo es pequeña en la mayoría de los casos, exceptuando las fila tres y cuatro de la tpm, que presentan una asimetría y una dispersión considerable. Los intervalos nos permiten un uso práctico de que tan precisas son las estimaciones.

El paquete bayesplot, proporciona la función MCMC-intervals basada en el método cuantil, que estima a partir de las muestras posteriores los intervalos de credibilidad con un nivel de probabilidad fijado por el usuario. Por lo tanto su implementación es bastante sencilla.

En la figura 17 se graficaron los histogramas univariados y los diagramas de dispersión bivariados para el vector de medias de los estados dependientes y el parámetro de cero inflación, el gráfico para los parámetros restantes se adjunta en anexos. No se evidencian problemas de colinealidad, ni la presencia de no-identificabilidad multiplicativa (formas tipo plátano), o en términos más simples problemas de divergencias al momento de aplicar el No-U-Turn-Sampler (NUTS), asegurandonos que las inferencias sean apropiadas. Mientras que para la matriz de transición de probabilidad, parece haber problemas de colinealidad entre  $\gamma_{11}$  con  $\gamma_{12}$ , y una ligera colinealidad entre  $\gamma_{21}$  con  $\gamma_{22}$  más no parecen haber problemas de no identificabilidad. Además se debe tener en cuenta que dada la restricción  $\sum_{j=1}^m \gamma_{ij} = 1$ , esto hace que los parámetros por fila de la tpm sean dependientes entre sí, sin embargo a continuación se realizan pruebas más avanzadas como el Test de Heidel con el fin de determinar que los valores muestreados sean apropiados.

	P. Estacionariedad	Valor p	Prueba	Media	Medio.Ancho
$\theta$	paso	0.928	paso	0.449	0.001
$\lambda_1$	paso	0.645	paso	2.868	0.008
$\lambda_2$	paso	0.689	paso	15.369	0.023
$\lambda_3$	paso	0.477	paso	43.153	0.080
$\lambda_4$	paso	0.594	paso	99.464	0.194
$a_{11}$	paso	0.601	paso	0.795	0.001
$a_{21}$	paso	0.504	paso	0.470	0.002
$a_{31}$	paso	0.257	paso	0.188	0.002
$a_{41}$	paso	0.611	paso	0.137	0.003
$a_{12}$	paso	0.444	paso	0.158	0.001
$a_{22}$	paso	0.356	paso	0.328	0.002
$a_{32}$	paso	0.620	paso	0.331	0.002
$a_{42}$	paso	0.992	paso	0.309	0.004
$a_{13}$	paso	0.632	paso	0.034	0.000
$a_{23}$	paso	0.786	paso	0.146	0.002
$a_{33}$	paso	0.091	paso	0.436	0.003
$a_{43}$	paso	0.462	paso	0.280	0.004
$a_{14}$	paso	0.765	paso	0.012	0.000
$a_{24}$	paso	0.234	paso	0.056	0.001
$a_{34}$	paso	0.268	paso	0.045	0.001
$a_{44}$	paso	0.583	paso	0.274	0.004
lp	paso	0.175	paso	-564.870	0.207

Tabla 19: Prueba de estacionariedad, usando el estadístico de Cramer-von-Mises para la convergencia de la cadena y prueba de medio ancho para la media calculando el intervalo de confianza al 0.95.

Con el fin de establecer técnicas más sofisticadas para determinar si el proceso de muestreo por NUTS, fue exitoso existen diferentes metodologías, tanto gráficas como basadas en hipótesis. La gráfica de traza en la figura 16 parece consistente, sin embargo la prueba de convergencia de Heidel, permite determinar si los valores muestreados provienen de una distribución estacionaria. Por lo tanto se aplica esta prueba para cada uno de los parámetros obtenidos por el ZIP HMM y se contrasta contrastan con la estadística

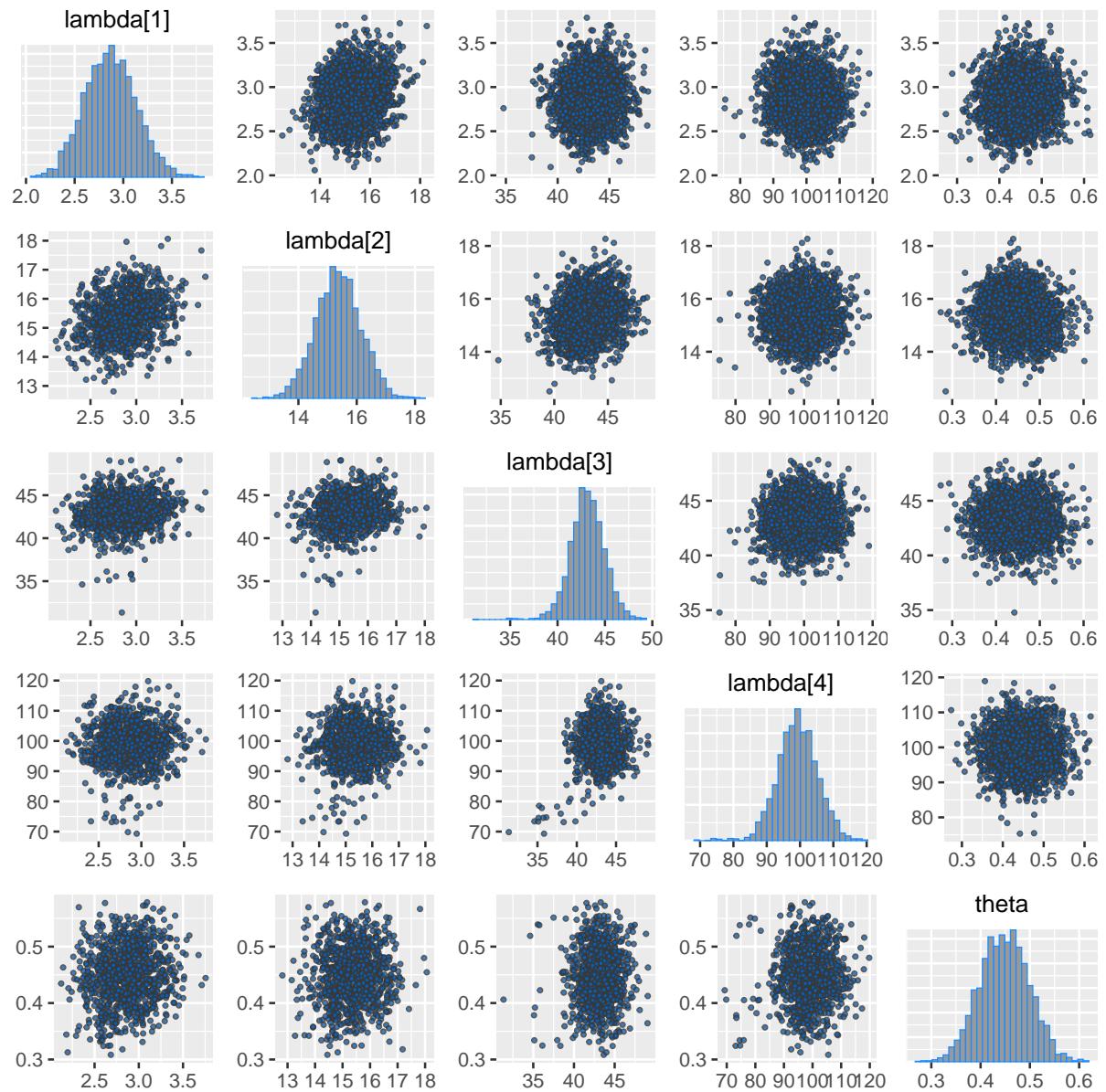


Figura 17: Gráfico de dispersión para las muestras MCMC del ZIP HMM.

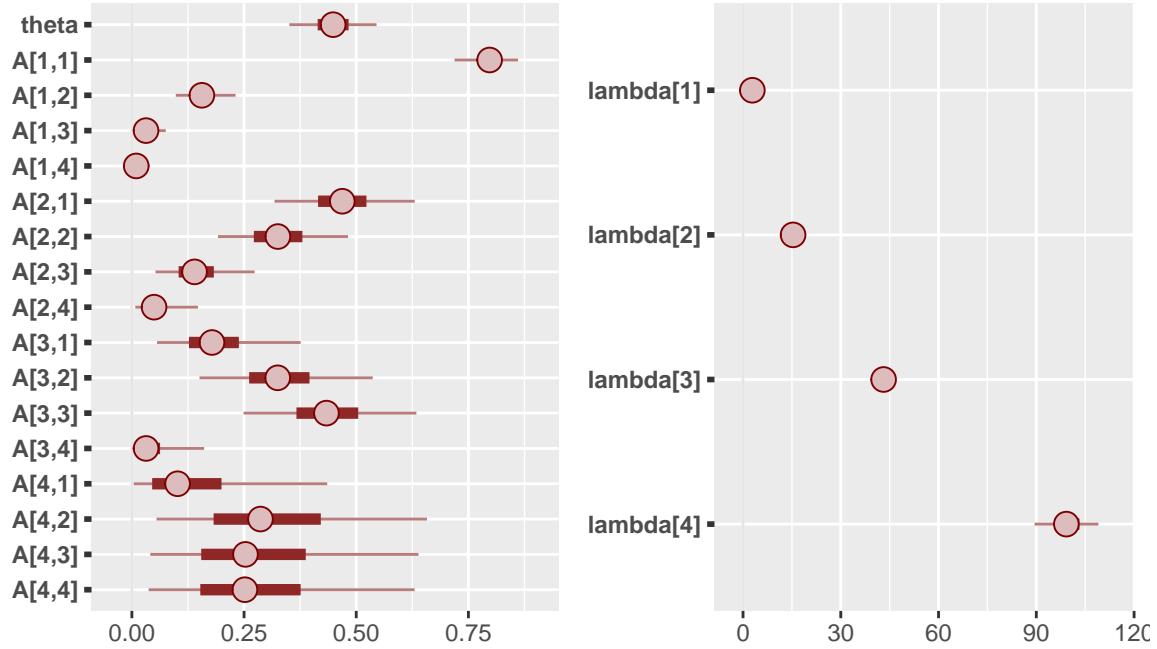


Figura 18: Intervalos de crédibilidad al 0.95 ZIP HMM.

de Cramer-von-Mises para un nivel de significancia  $\alpha = 0.05$  fijo, donde la hipótesis nula es que la cadena es estacionaria. En todos los casos el valor  $p$  fue mayor que 0.05 lo que indica que existe suficiente evidencia estadística para no rechazar la hipótesis nula, lo cual indica que los valores muestrados ofrecen una estimación confiable al provenir de una distribución estacionaria. También se utilizó la prueba de medio ancho, que calcula un intervalo de confianza del 95 % para la media, y utiliza la parte de la cadena que pasó la prueba de estacionariedad, para determinar si la muestra fue lo suficientemente grande para estimar la media con precisión. Los resultados indican que efectivamente cada uno de los parámetros aprobó la prueba de medio ancho.

### 1.2.2. Comparación ZIP HMM clásico vs Bayesiano

Igual que en el caso del PHMM, la estimación bajo el enfoque clásico se realizó utilizando bootstrap, generando 100 muestras independientes a partir del ZIP HMM de orden 4 de longitud 180 igual a la serie GIF en Colombia. Los valores iniciales usados fueron los estimados por ZIP HMM de 4 estados, que permiten la convergencia del algoritmo en pocas iteraciones. La probabilidad y el nivel de confianza se fijó al 0.95, para los intervalos de credibilidad y confianza, respectivamente, en la tabla 20 se encuentran registrados los resultados obtenidos.

Ambos intervalos se relacionan con la precisión de nuestra estimación. La forma más común para saber el desempeño del método, es calculando la longitud de este, donde se espera que el ancho sea lo más pequeño posible. Los resultados se muestran en la tabla 20, junto con la media de las estimaciones, tanto para el ZIP HMM frecuentista como para el bayesiano. No parece haber un ganador indiscutible, si embargo haciendo un análisis más detallado por parámetros se puede decir lo siguiente. Para  $\theta$  el parámetro de cero inflación es más pequeña la longitud en el caso bayesiano, para el vector de medias  $\lambda$  es un empate 2 y 2 para cada uno, sin embargo preocupa que para  $\lambda_4$  en el caso clásico el ancho es extremadamente grande. Finalmente para los valores de la tpm en 9 de los 16 intervalos estimados, el enfoque bayesiano nuevamente es el vencedor. En conclusión aunque los intervalos de credibilidad tienen una longitud en la mayoría de los casos, no parece haber un método que sea evidentemente el mejor.

Parámetros	Intervalos de Credibilidad				Intervalos de Confianza			
	Media	2.5	97.5	Ancho	Media	2.5	97.5	Ancho
$\theta$	0.449	0.351	0.545	0.194	0.444	0.343	0.543	0.200
$\lambda_1$	2.868	2.371	3.408	1.037	2.763	2.218	3.181	0.963
$\lambda_2$	15.369	13.859	16.877	3.017	15.115	13.853	16.409	2.556
$\lambda_3$	43.153	39.600	46.658	7.059	43.148	40.002	46.844	6.842
$\lambda_4$	99.235	87.428	110.912	23.485	99.306	38.082	109.076	70.994
$a_{11}$	0.795	0.719	0.861	0.142	0.820	0.735	0.889	0.154
$a_{21}$	0.470	0.318	0.630	0.313	0.483	0.317	0.680	0.363
$a_{31}$	0.188	0.056	0.376	0.320	0.168	0.019	0.366	0.347
$a_{41}$	0.137	0.004	0.435	0.431	0.004	0.000	0.006	0.006
$a_{12}$	0.158	0.098	0.231	0.133	0.154	0.089	0.221	0.131
$a_{22}$	0.328	0.192	0.482	0.290	0.335	0.184	0.500	0.316
$a_{32}$	0.331	0.151	0.537	0.386	0.329	0.128	0.643	0.514
$a_{42}$	0.309	0.055	0.658	0.603	0.329	0.000	0.996	0.996
$a_{13}$	0.034	0.009	0.076	0.067	0.025	0.001	0.065	0.064
$a_{23}$	0.146	0.053	0.274	0.221	0.131	0.051	0.257	0.206
$a_{33}$	0.436	0.249	0.634	0.385	0.499	0.210	0.665	0.455
$a_{43}$	0.280	0.041	0.639	0.598	0.346	0.000	0.998	0.997
$a_{14}$	0.013	0.000	0.040	0.039	0.001	0.000	0.018	0.018
$a_{24}$	0.056	0.008	0.147	0.140	0.050	0.004	0.138	0.134
$a_{34}$	0.045	0.001	0.161	0.160	0.004	0.000	0.099	0.099
$a_{44}$	0.274	0.037	0.630	0.593	0.320	0.000	0.665	0.665

Tabla 20: Intervalos de Credibilidad y Confianza para el ZIP HMM de orden 4.

## 2. Anexo Códigos

A continuación se anexa el código utilizado para el desarrollo de esta tesis, en la aplicación del PHMM a la base homicidios en Colombia y el ajuste del ZIP-HMM a los Grandes Incendios Forestales (GIF) en Colombia.

```
rm(list = ls())
#####
# Packages #
library(Bayeshmmcts)
library(bridgesampling)
library(rstan)
library(bayesplot)
library(coda)
library(ziphsmm)

#####
# Data homicidios #
data("homicides")

#####
# Poisson - Hidden Markov Model #
homicidios <- homicides
colnames(homicidios) <- c("Año", "Homicidios", "Población", "Tasa")
Homicidios <- ts(data = round(homicidios$Tasa), start = 1960)

# modelo clasico
mod2 <- pois.HMM.mle(o = Homicidios, m = 2, lambda0 = c(30, 63), A0 = matrix(c(0.9,
  0.1, 0.1, 0.9), 2, 2, byrow = TRUE), stationary = TRUE)

# Algoritmo viterbi (decodificación Global)
viterbi <- pois.HMM.viterbi(o = Homicidios, mod = mod2)

# Verificación de supuestos
residuales <- pois.HMM.pseudo_residuals(o = Homicidios, mod = mod2)
pois.HMM.plot.residuals(residuales)

# Predicción de los estados
año <- homicidios$Año
Estad_pred <- data.frame(Año = año[59] + 1:16, pois.HMM.state_prediction(h = 16,
  o = Homicidios, mod = mod2))
colnames(Estad_pred) <- c("Año", "Estado 1", "Estado 2")
Estad_pred$Estado <- apply(Estad_pred, 1, which.max)

# Distribución de pronóstico
delta <- pois.HMM.stadist(mod2)
h <- 16
xf <- 5:75
año <- homicidios[, 1]
forecasts <- pois.HMM.forecast(xf, h, Homicidios, mod2)
par(mfrow = c(4, 4), las = 1)
for (i in 1:h) {
  fc <- forecasts[, i]
  plot(xf, fc, type = "h", main = paste("Dist. pronós.", año[59] + i), xlim = c(5,
    max(xf + 2)), ylim = c(0, 0.1), cex.main = 0.85, xlab = "conteo", ylab = "probabilidad",
    lwd = 1)
  rect(par("usr")[1], par("usr")[3], par("usr")[2], par("usr")[4], col = gray(0.9,
    0.9), border = "white")
```

```

grid(lty = 1, col = "white")
lines(xf, fc, type = "h", lwd = 1)
dstat <- numeric(length(xf))
for (j in 1:mod2$m) dstat <- dstat + delta[j] * dpois(xf, mod2$lambda[j])
lines(xf, dstat, col = "chartreuse2", lwd = 2)
}

# Modelo bayesiano de 2 estados
PHMM_2states <- bayes.PHMM(y = Homicidios, m = 2, chains = 3, iter = 2000, control = list(adapt_delta = 0.001))
print(PHMM_2states, digits = 3)

# Modelo bayesiano de 3 estados
PHMM_3states <- bayes.PHMM(y = Homicidios, m = 3, chains = 3, iter = 2000, control = list(adapt_delta = 0.001))

# estimates of the log marginal likelihoods
bridge_H0 <- bridge_sampler(samples = PHMM_2states)
bridge_H1 <- bridge_sampler(samples = PHMM_3states)

error_measures(bridge_H0)$percentage
error_measures(bridge_H1)$percentage

# The Bayes factor in favor of H0 over H1 can then be obtained as follows:
bridge_H0$logml
bridge_H1$logml

# Factor de bayes
bf(bridge_H0, bridge_H1)

# Posterior
posterior <- as.array(PHMM_2states)
lp_cp <- log_posterior(PHMM_2states)
np_cp <- nuts_params(PHMM_2states)
rstan::traceplot(PHMM_2states)

# Gráfica de intervalos de credibilidad
color_scheme_set("red")
mcmc_intervals(posterior, prob_outer = 0.95, pars = c("A[1,1]", "A[1,2]", "A[2,1]", "A[2,2]"))

# Histogramas univariados y gráfico de dispersión bivariado
color_scheme_set("mix-brightblue-gray")
mcmc_pairs(posterior, np = np_cp, pars = c("A[1,1]", "A[1,2]", "A[2,1]", "A[2,2]", "lambda[1]", "lambda[2]"), off_diag_args = list(size = 0.75))

# Prueba de Heibelberg y test medio ancho
PHMM_mcmc <- as.mcmc(as.matrix(PHMM_2states))
Test_HyW <- heidel.diag(PHMM_mcmc)

# Intervalos de Confianza y de credibilidad
intervalos_cred <- mcmc_intervals_data(posterior, prob_outer = 0.95, point_est = "mean")
intervalos_conf <- pois.HMM.confint(mod = mod2, n = 59, B = 250)

##### Zero Inflated Poisson - Hidden Markov Model #

```

```

rm(list = ls())

##### Data GIF #
data("wildfires")
incendios <- wildfires
colnames(incendios) <- c("Fecha", "GIF")
GIF <- ts(data = incendios$GIF, start = c(2002, 1), frequency = 12)

#### ZIP HMM clásico de 2 estados
ZIPHMM_2states <- hmmfit(y = incendios$GIF, M = 2, prior_init = c(0.6, 0.4),
                           tpm_init = matrix(c(0.9, 0.1, 0.5, 0.5), 2, 2, byrow = TRUE), emit_init = c(7,
                                                                                         45), zero_init = c(0.4, 0), method = "Nelder-Mead", hessian = TRUE,
                           control = list(maxit = 1000, trace = 1))

#### ZIP HMM clásico de 4 estados
ZIPHMM_4states <- hmmfit(y = incendios$GIF, M = 4, prior_init = c(0.5, 0.2,
                                                                    0.2, 0.1), tpm_init = matrix(c(0.8, 0.15, 0.04, 0.01, 0.5, 0.3, 0.15, 0.05,
                                                                    0.15, 0.35, 0.45, 0.05, 0.15, 0.35, 0.25, 0.25), 4, 4, byrow = TRUE), emit_init = c(3,
                                                                                         15, 43, 100), zero_init = c(0.45, 0, 0, 0), method = "Nelder-Mead", hessian = TRUE,
                           control = list(maxit = 1000, trace = 1))

# Algoritmo Viterbi para el ZIP HMM
ZIP.viterbi <- hmmviterbi(y = incendios$GIF, ntimes = length(incendios$GIF),
                           M = 4, prior_init = ZIPHMM_4states$prior, tpm_init = ZIPHMM_4states$tpm,
                           emit_init = ZIPHMM_4states$emit_parm, zero_init = ZIPHMM_4states$zeroprop)

# Gráfica ddecodificación algoritmo Viterbi
par(mfrow = c(1, 1))
par(mar = c(2, 2, 1, 0.5) + 0.5, mgp = c(1.6, 0.6, 0))
### Plot 1
plot(GIF, xlab = "Año", type = "o", col = 4, ylab = "Número")
rect(par("usr")[1], par("usr")[3], par("usr")[2], par("usr")[4], col = gray(0.9,
                                                                      0.9), border = "white")
grid(lty = 1, col = "white")
lines(GIF, type = "o", col = 4)
abline(h = ZIPHMM_4states$emit_parm, col = "orange", lty = 2)
points(x = time(GIF), y = ifelse(ZIP.viterbi == 1, ZIPHMM_4states$emit_parm[1],
                                   ifelse(ZIP.viterbi == 2, ZIPHMM_4states$emit_parm[2], ifelse(ZIP.viterbi ==
                                   3, ZIPHMM_4states$emit_parm[3], ZIPHMM_4states$emit_parm[4])), pch = 21,
                                   bg = "orange", col = "white")

# Bayes ZIP HMM 2 y 4 Estaddos
Bayes_ZIPHMM1_2S <- bayes.ZIPHMM1(y = GIF, m = 2, chains = 4, iter = 2000)
Bayes_ZIPHMM1_4S <- bayes.ZIPHMM1(y = GIF, m = 4, chains = 4, iter = 2000)

# Factor de bayes, extrayendo la log verosimilitud marginal, utilizando
# muestreador por puente
set.seed(1)
bridge_H0 <- bridge_sampler(samples = Bayes_ZIPHMM1_2S)
bridge_H2 <- bridge_sampler(samples = Bayes_ZIPHMM1_4S)
bf(bridge_H0, bridge_H2) # Evidencia extrema para H2

# Gráfica de las cadenas
posteriorZ <- as.array(Bayes_ZIPHMM1_4S)

```

```

lp_cpZ <- log_posterior(Bayes_ZIPHMM1_4S)
np_cpZ <- nuts_params(Bayes_ZIPHMM1_4S)
rstan::traceplot(Bayes_ZIPHMM1_4S, pars = c("lambda[1]", "lambda[2]", "lambda[3]",
  "lambda[4]", "A[1,1]", "A[1,2]", "A[1,3]", "A[1,4]", "A[2,1]", "A[2,2]",
  "A[2,3]", "A[2,4]", "A[3,1]", "A[3,2]", "A[3,3]", "A[3,4]", "A[4,1]", "A[4,2]",
  "A[4,3]", "A[4,4]", "theta"), ncol = 4)

# Prueba de convergencia Heidelberg y Welch
ZIPHMM_mcmc <- as.mcmc(as.matrix(Bayes_ZIPHMM1_4S))
Test_HyW <- heidel.diag(ZIPHMM_mcmc)

# Intervalos de Credibilidad y de Confianza
intervalos_credZ <- as.data.frame(mcmc_intervals_data(posteriorZ, prob_outer = 0.95,
  point_est = "mean"))
intervalos_credZ$Ancho <- intervalos_credZ$hh - intervalos_credZ$l1
intervalos_confZIP <- ZIP.HMM.confint(mod = ZIPHMM_4states, n = length(GIF),
  B = 100)

```