

Resultados

Rafael Eduardo Diaz

3 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 llevo 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 consulto específicamente el capítulo 8 indicadores de violencia, se complemento 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 comunmente 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	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

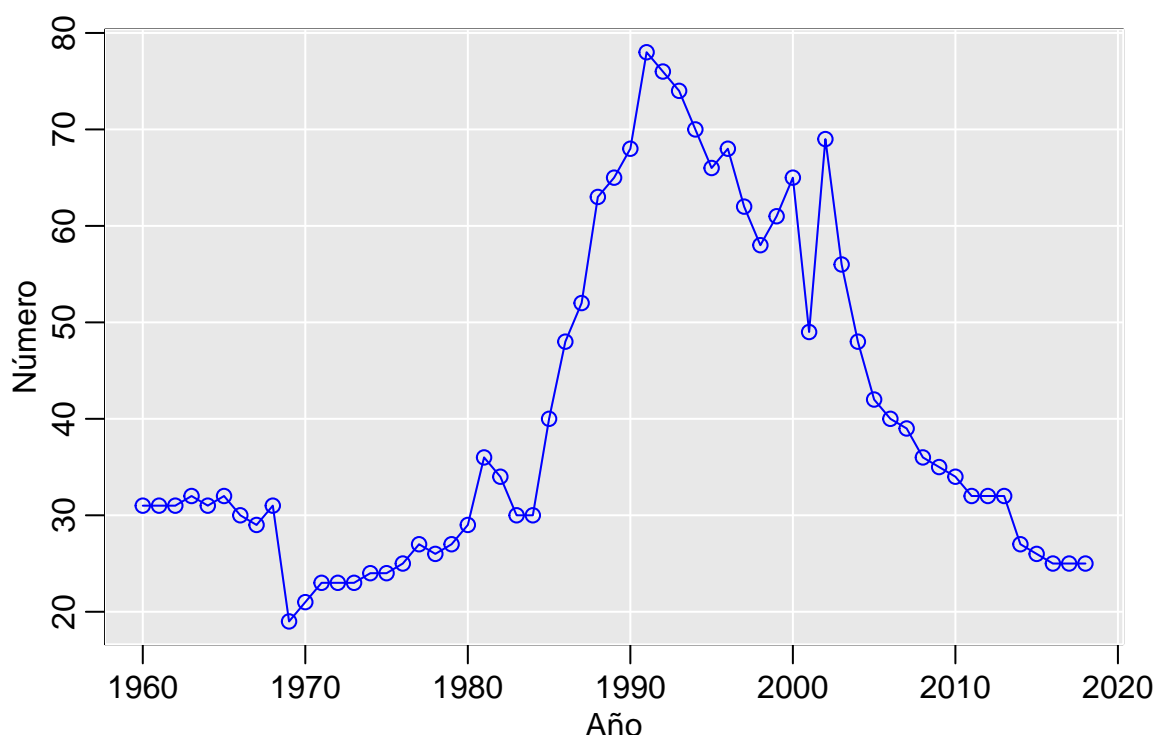


Figura 1: Serie de tiempo homicidios en Colombia desde el año 1960 hasta el año 2018.

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 utilizara un PHMM.

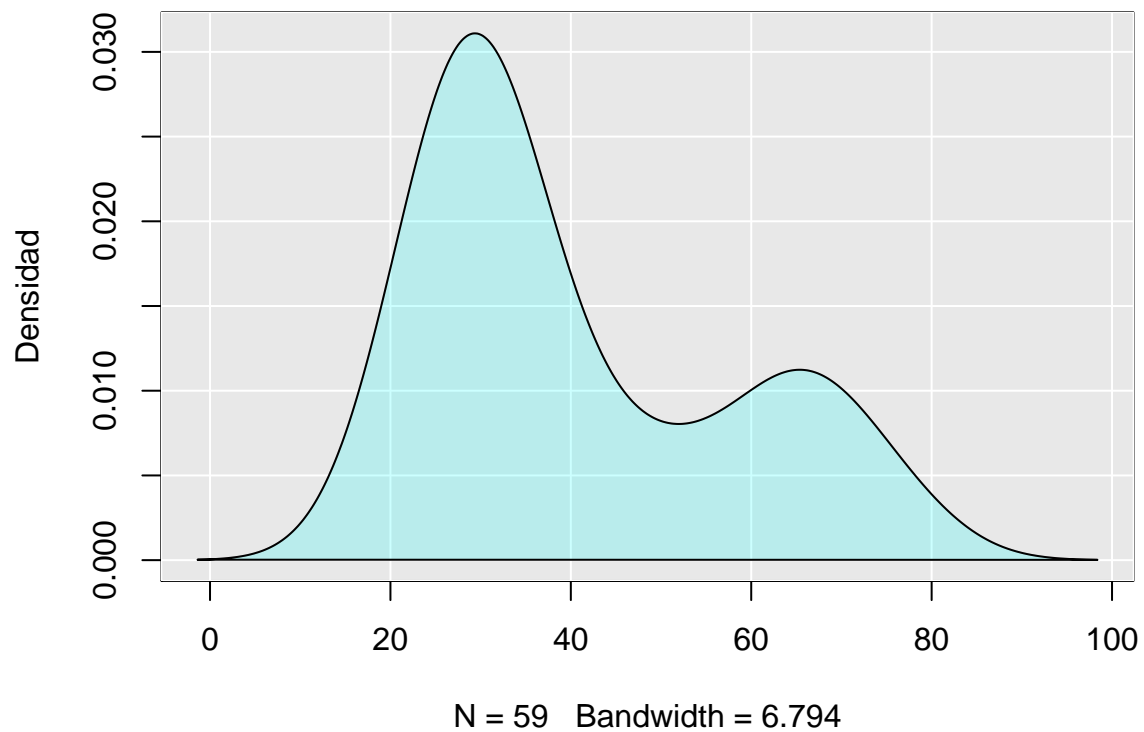


Figura 2: Kernel Densidad de homicidios en Colombia (1960-2018).

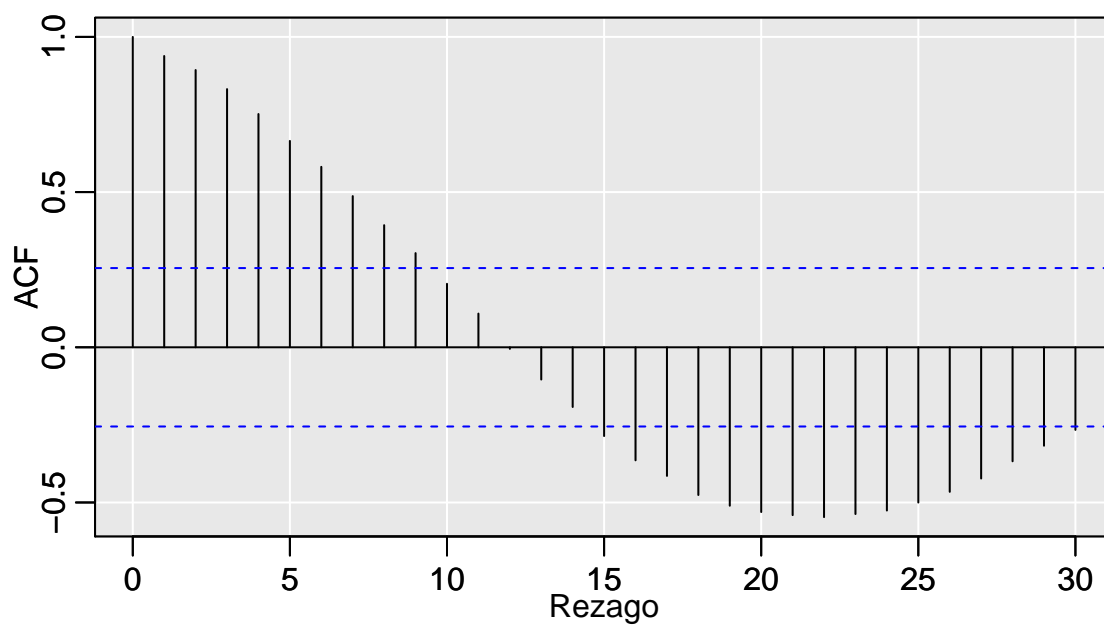


Figura 3: Función de autocorrelación muestral, para la serie de homicidios.

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 - Estados 1	1.00	-356.91	715.81	717.89
2	PHMM - Estados 2	4.00	-201.32	410.65	418.96
3	PHMM - Estados 3	9.00	-193.01	404.02	422.71
4	PHMM - Estados 4	16.00	-190.84	413.69	446.93
5	PHMM - Estados 5	25.00	-190.29	430.58	482.51
6	mixtura indep. (2)	3.00	-229.38	464.75	470.98
7	mixtura indep. (3)	5.00	-228.11	466.21	476.60
8	mixtura indep. (4)	7.00	-228.11	470.21	484.76

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

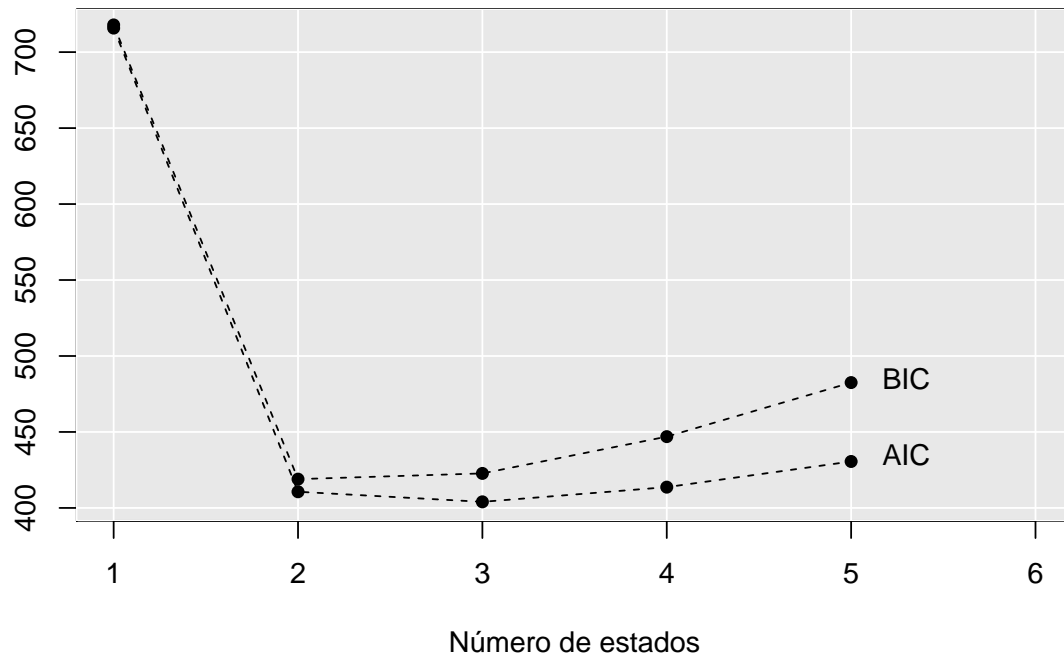


Figura 4: Serie homicidios: selección de modelos AIC y BIC.

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 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 λ y los valores de la distribución estacionaria π .

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

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

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:

	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.

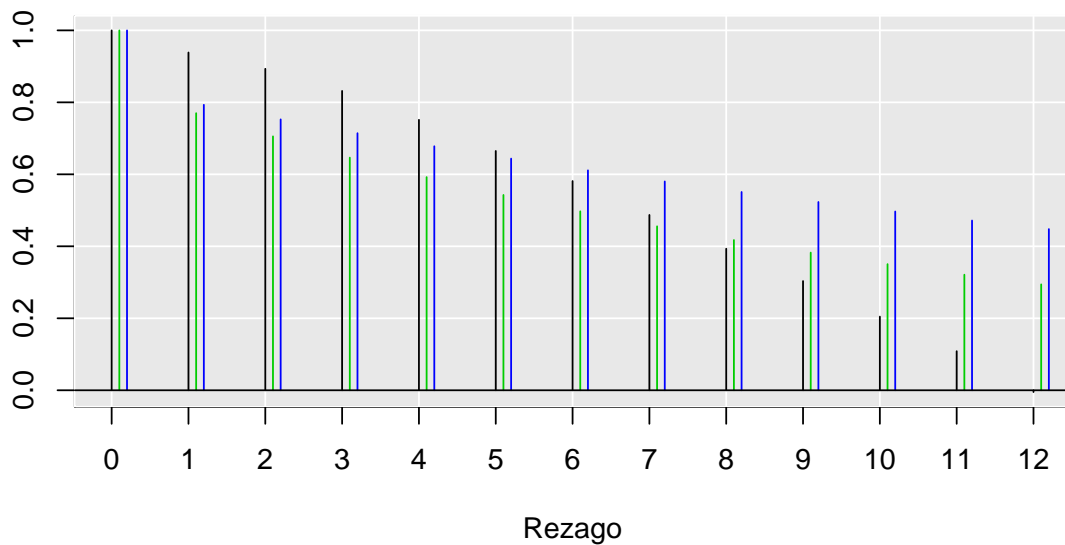


Figura 5: Datos homicidios: ACF y ACF de los PHMM con dos y tres estados.

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.

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 cuántílicos) 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.

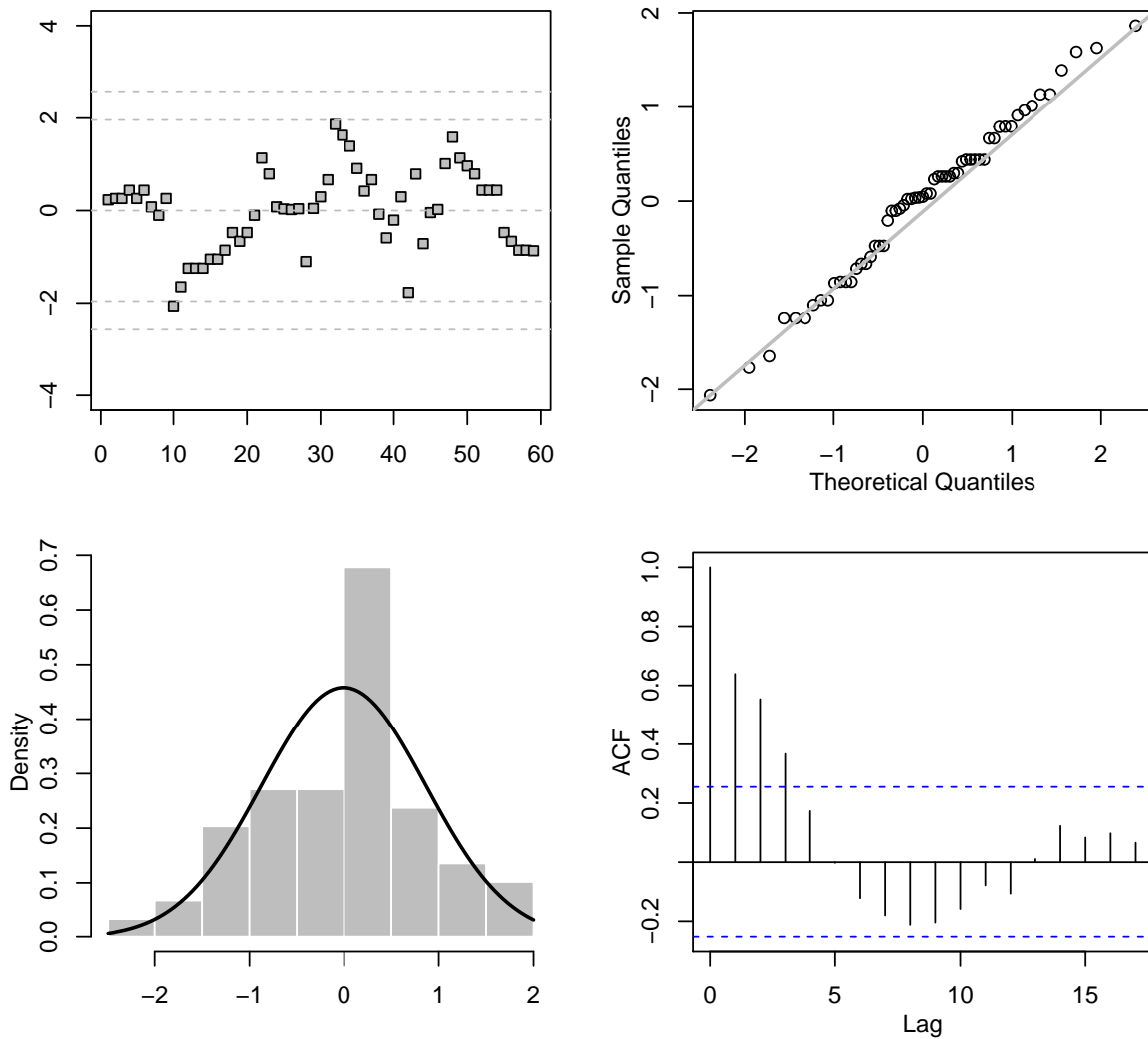


Figura 6: Grafico pseudo-residuales ordinarios para el PHMM de 2 estados.

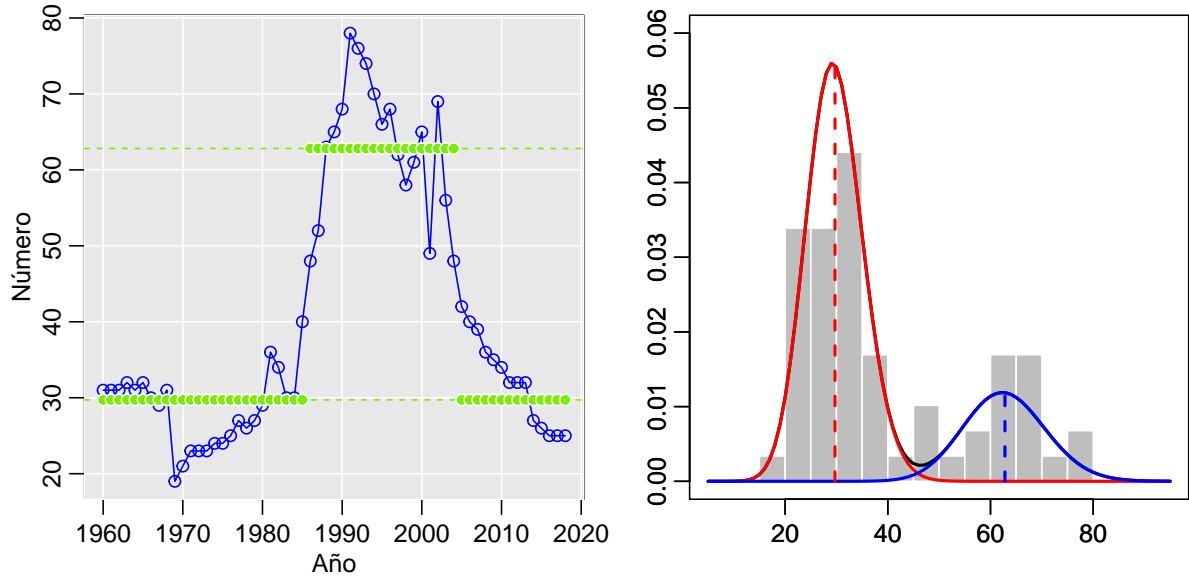


Figura 7: Algoritmo Viterbi aplicado a un PHMM de dos estados.

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, ± 1.96 y ± 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.

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 clasifico 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	1	2	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	1

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

Se realiza la predicción de los estados más probables para los proximos 16 años, tambien podemos pronosticar la distribución para estos mismos años. Como se observa en la figura 8 a medida que el

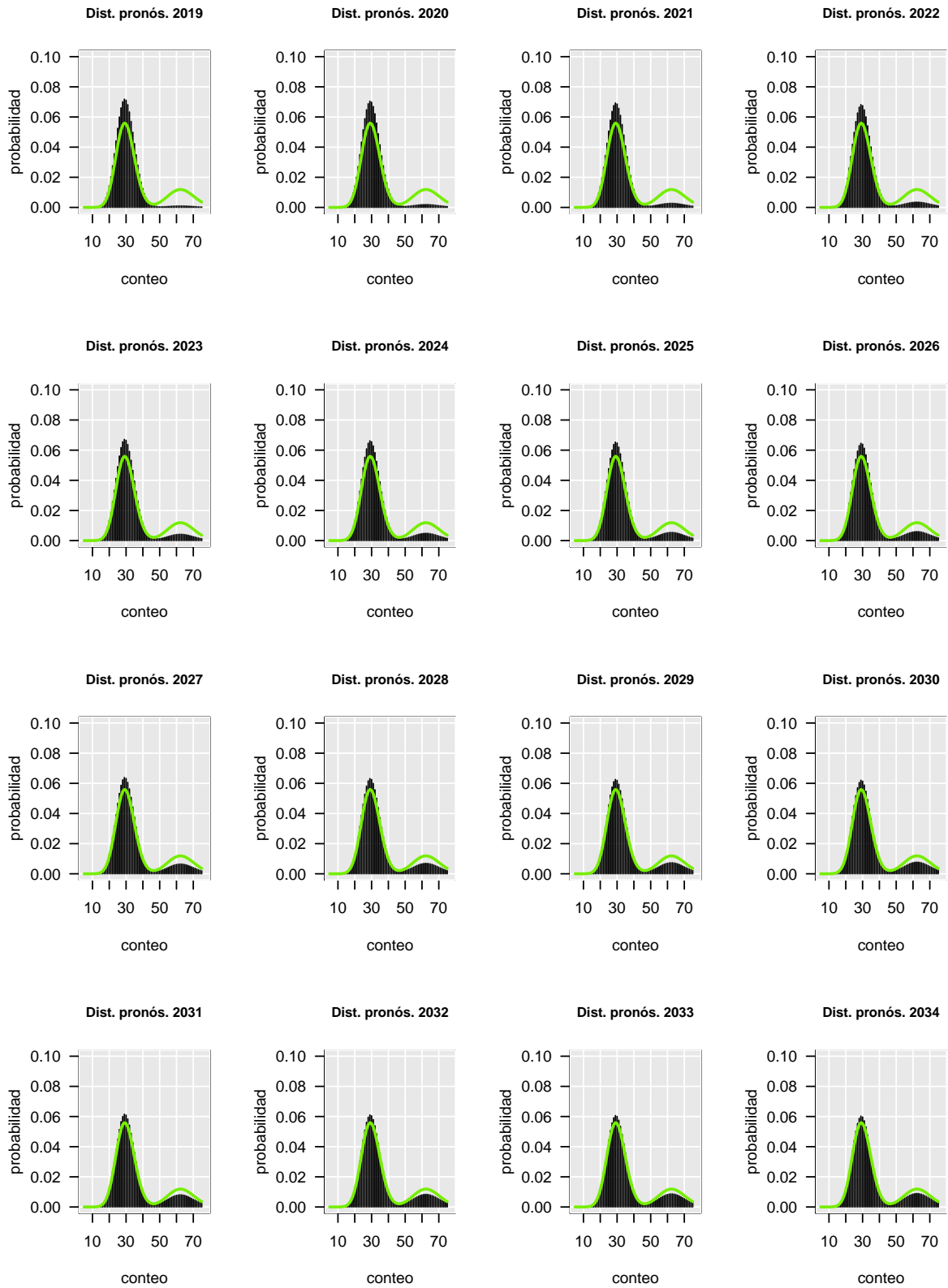


Figura 8: Pronostico de la distribución para los años 2019 a 2034.

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 situa 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 situa como uno de los países más violentos del mundo ubicandose en el top 20, las cifras de la fiscalia indican que despues 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

Estimación Bayesiana del PHMM

Primero se ajustaron cuatro modelos, con la función ‘Bayeshmmcts::bayes.PHMM’, para 2, 3, 4 y 5 estados, después, se estimó la log - verosimilitud marginal, utilizando muestreo por puente como alternativa a la 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 de 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 Jeffereys(1961) y modificada ligeramente por Lee y Wagenmakers en 2013:

B01	Decisión
>100	Evidencia extrema para H0
30 - 100	Evidencia muy fuerte para H0
10 - 30	Evidencia fuerte para H0
3 - 10	Evidencia moderada para H0
1 - 3	Evidencia apenas mencionable para H0
1	No hay evidencia
1/3 - 1	Evidencia apenas mencionable para H1
1/10 - 1/3	Evidencia moderada para H1
1/30 - 1/3	Evidencia fuerte para H1
1/100 - 1/30	Evidencia muy fuerte para H1
< 1/100	Evidencia extrema para H1

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.

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

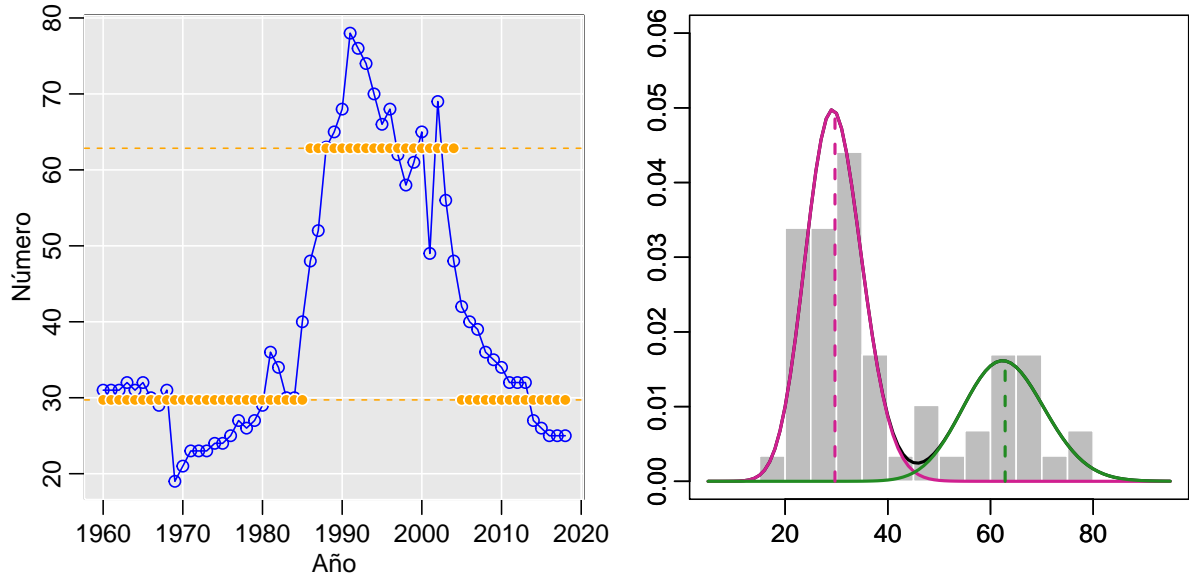


Figura 9: Algoritmo Viterbi aplicado a un PHMM de dos estados.

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
Γ_{11}	0.953	0.001	0.032	0.873	0.935	0.961	0.977	0.994	2491.677	1.000
Γ_{12}	0.047	0.001	0.032	0.006	0.023	0.039	0.065	0.127	2491.677	1.000
Γ_{21}	0.099	0.001	0.065	0.014	0.051	0.084	0.133	0.257	2417.781	1.000
Γ_{22}	0.901	0.001	0.065	0.743	0.867	0.916	0.949	0.986	2417.781	1.000
λ_1	29.715	0.018	0.871	28.097	29.111	29.684	30.299	31.460	2456.451	1.001
λ_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.

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	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	1

Tabla 11: Resultados de la decodificación global bayesiana con el algoritmo Viterbi.

Diagnosticos de la cadena

En esta sección se verificara el diagnostico 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 purebas 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

	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

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

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 este caso parece no 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 están 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 comportana bastante bien, pues hay un mínimo de muestras divergentes.

Los intervalos de credibilidad, para los parámetros calculados a partir de las muestras posteriores con todas las cadenas fusionadas. Índican valores consistentes en las estimaciones de los parámetros.

La prueba de convergencia utiliza la estadística de Cramer-von-Mises para probar la hipótesis nula de que los valores muestreados 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	P..Medio.ancho	Media	Valor.p
Γ_{11}	páso	0.396	páso	0.953	0.001
Γ_{21}	páso	0.978	páso	0.099	0.002
Γ_{12}	páso	0.396	páso	0.047	0.001
Γ_{22}	páso	0.978	páso	0.901	0.002
λ_1	páso	0.569	páso	29.701	0.034
λ_2	páso	0.862	páso	62.742	0.079
lp	páso	0.440	páso	-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.

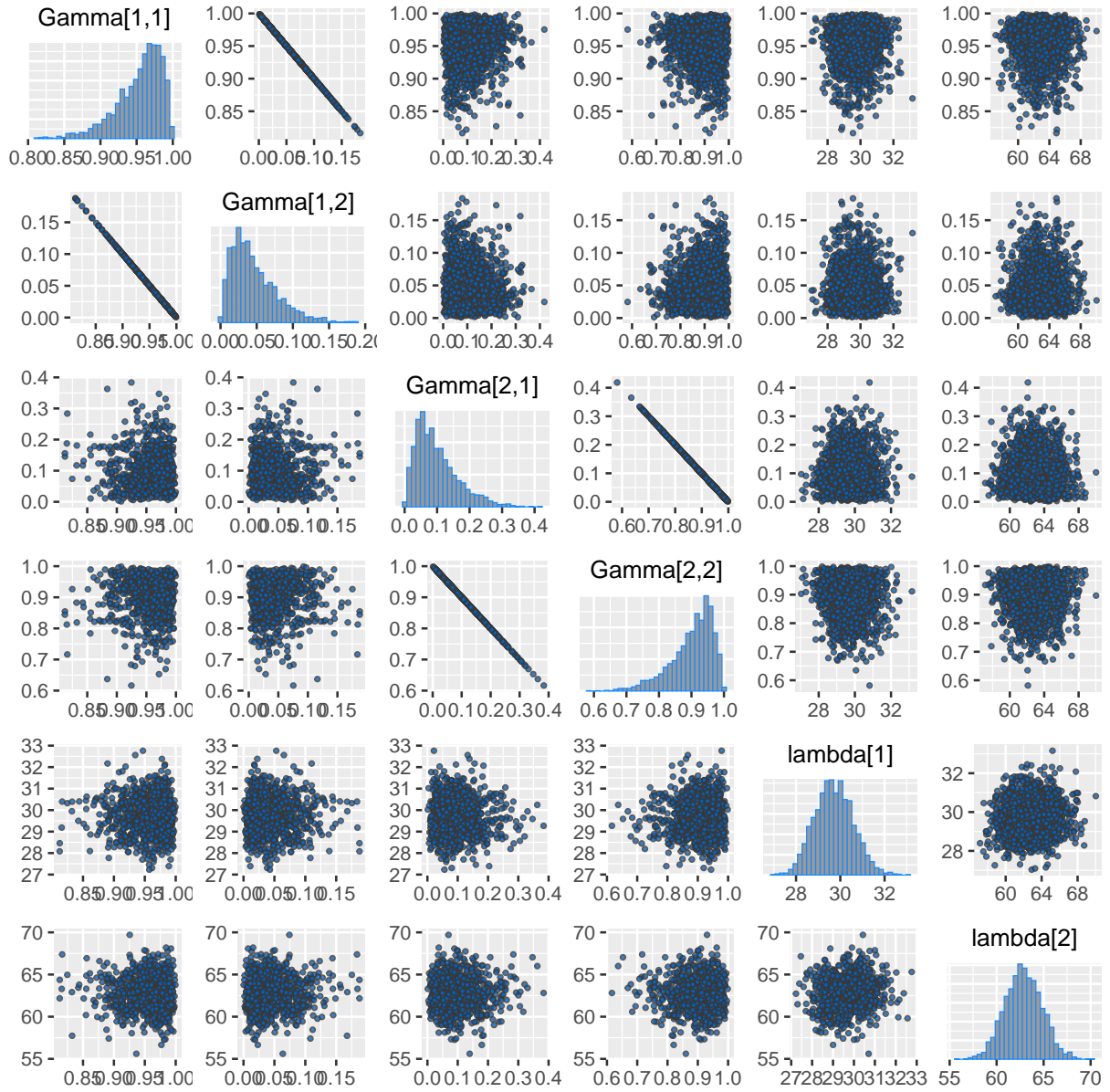


Figura 10: Gráfico de dispersión para las muestras MCMC.

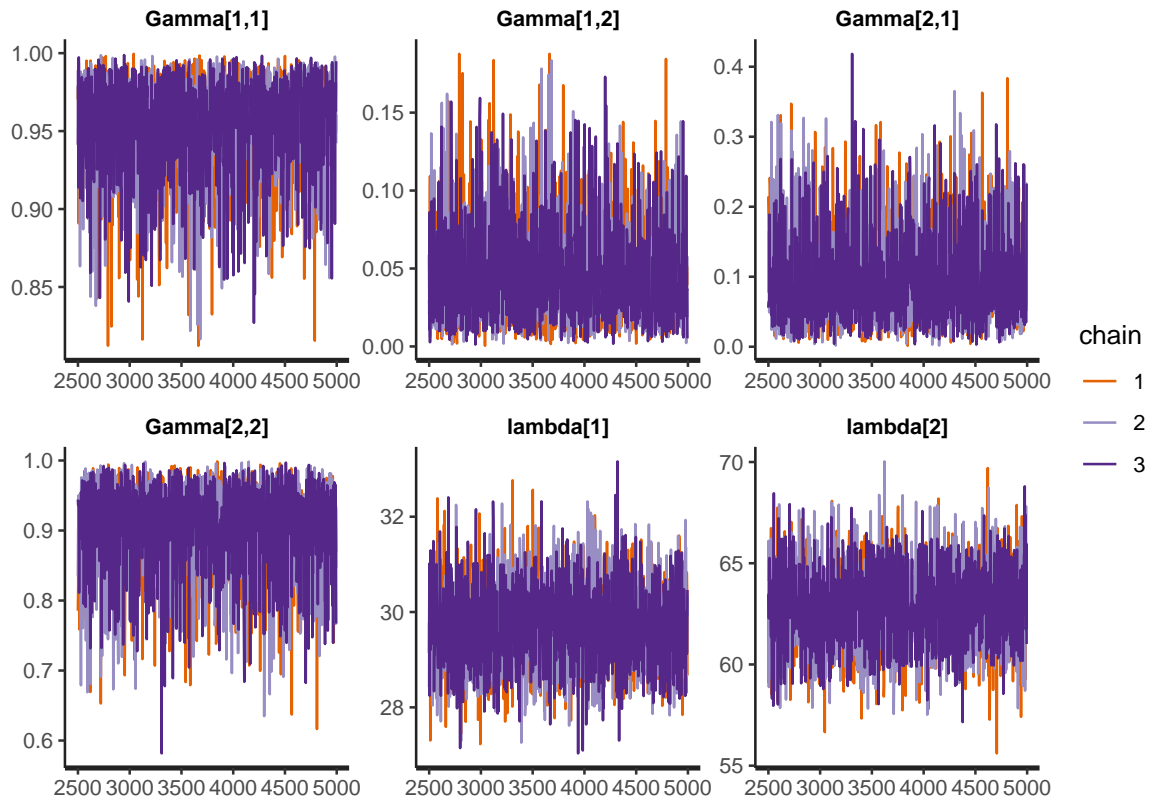


Figura 11: Gráfico de trazas de las cadenas, para cada iteración y por cadena.

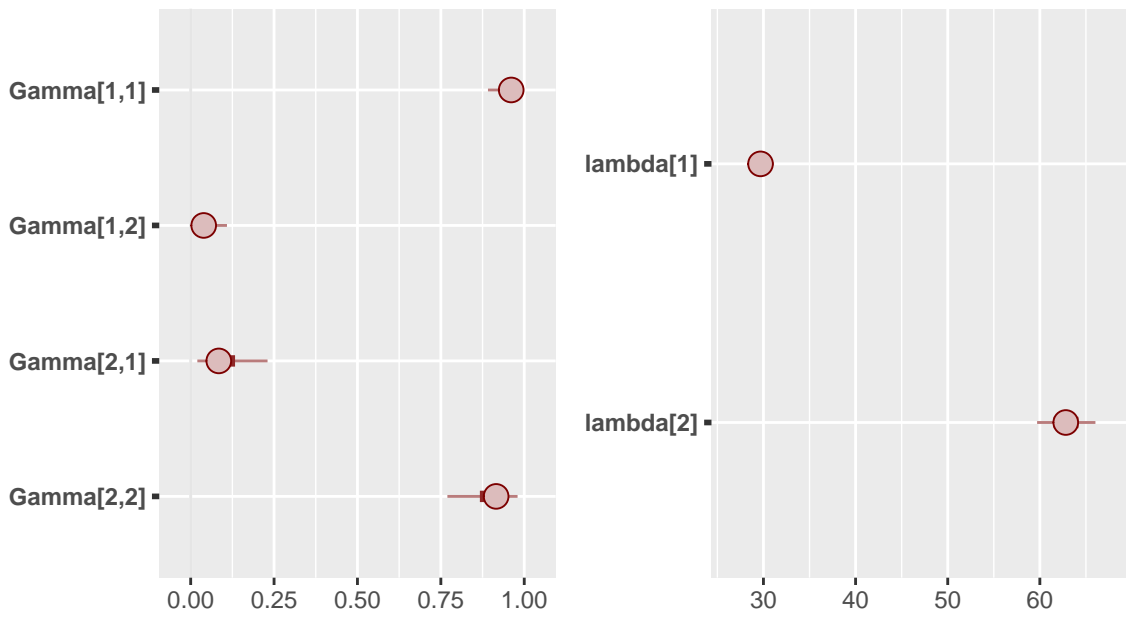


Figura 12: Intervalos de credibilidad.

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 a posteriori 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	EMV	2.5	97.5	Ancho
gamma11	0.953	0.873	0.994	0.120	0.980	0.844	1.000	0.156
gamma21	0.099	0.014	0.257	0.244	0.064	0.015	1.000	0.985
gamma12	0.047	0.006	0.127	0.120	0.020	0.000	0.156	0.156
gamma22	0.901	0.743	0.986	0.244	0.936	0.000	0.985	0.985
lambda1	29.715	28.097	31.460	3.363	29.716	27.689	31.648	3.959
lambda2	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.

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 ± 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.

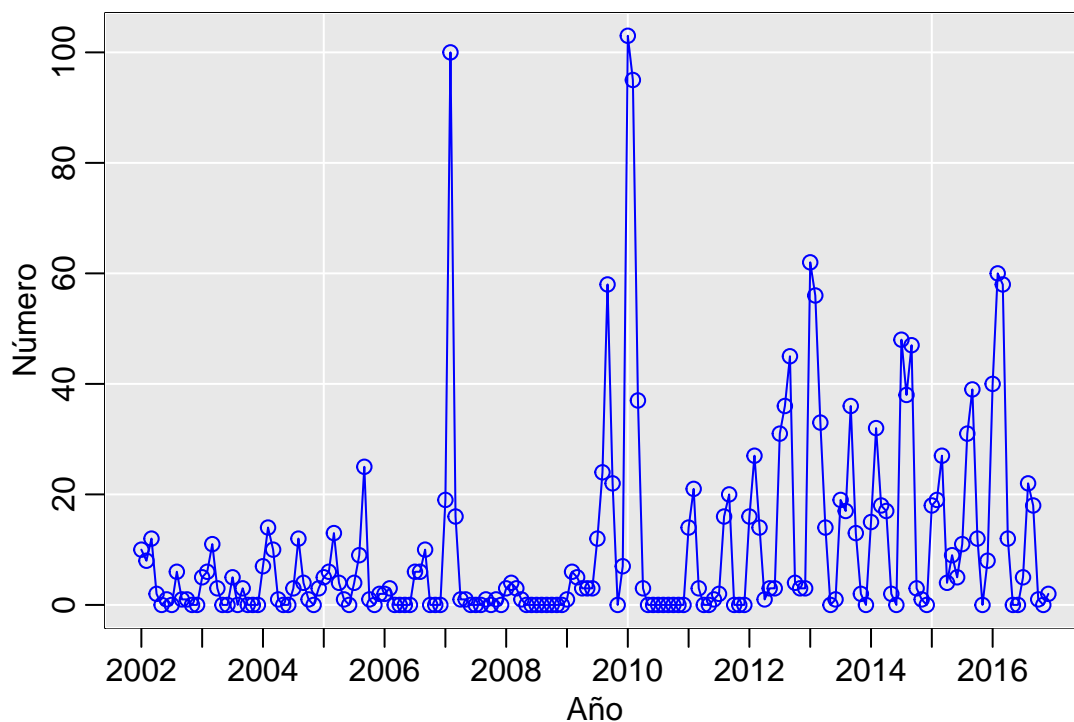


Figura 13: Serie de tiempo Grandes Incendios Forestales en Colombia desde el año 2002 hasta el año 2016.

En la figura 13, 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 periodos 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 sse 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.

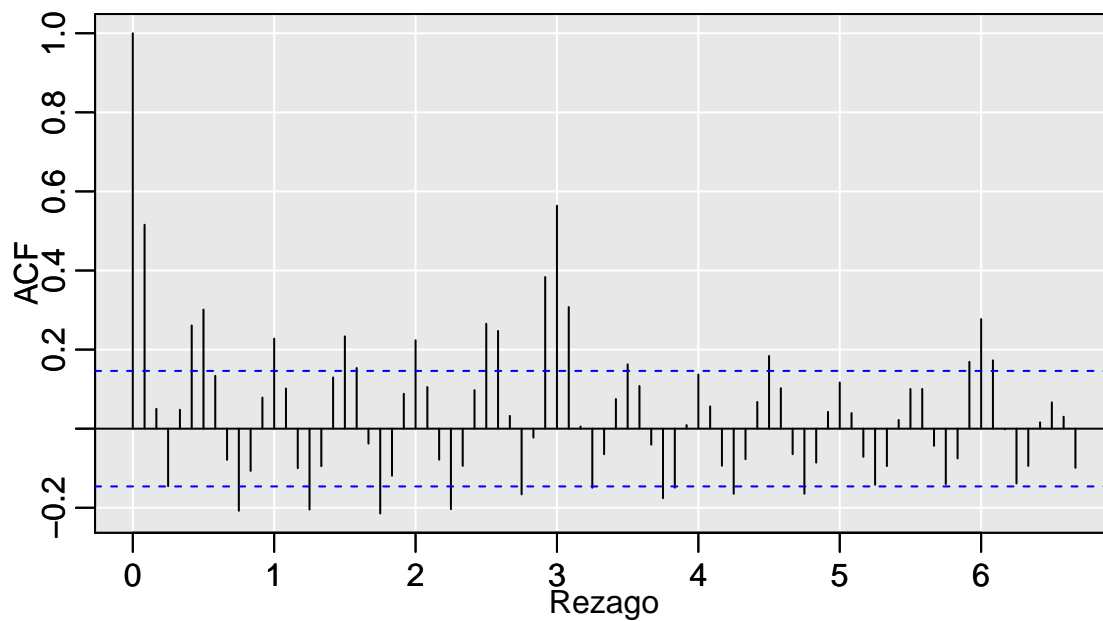


Figura 14: Función de autocorrelación muestral, para la serie GIF.

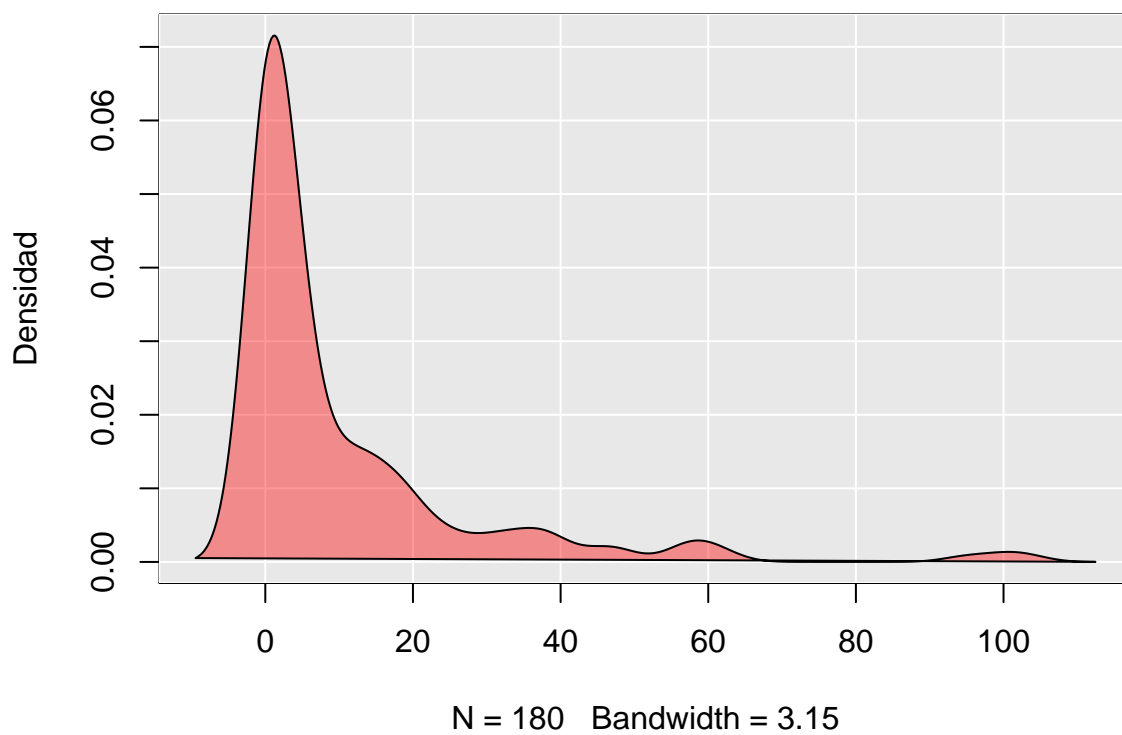


Figura 15: Kernel Densidad serie Grandes Incendios Forestales en Colombia (2002-2016).

1.2.1. Ajuste del ZIP-HMM

	Modelo	p	logL	AIC	BIC
1	ZIPHMM - Estados 2	6.00	764.61	1541.23	1560.39
2	ZIPHMM - Estados 3	12.00	592.81	1209.62	1247.94
3	ZIPHMM - Estados 4	20.00	536.07	1112.15	1176.01
4	ZIPHMM - Estados 5	30.00	521.87	1103.74	1199.53
5	ZIPHMM - Estados 6	42.00	508.60	1101.20	1235.31
6	ZIPHMM - Estados 7	56.00	510.79	1133.58	1312.39

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

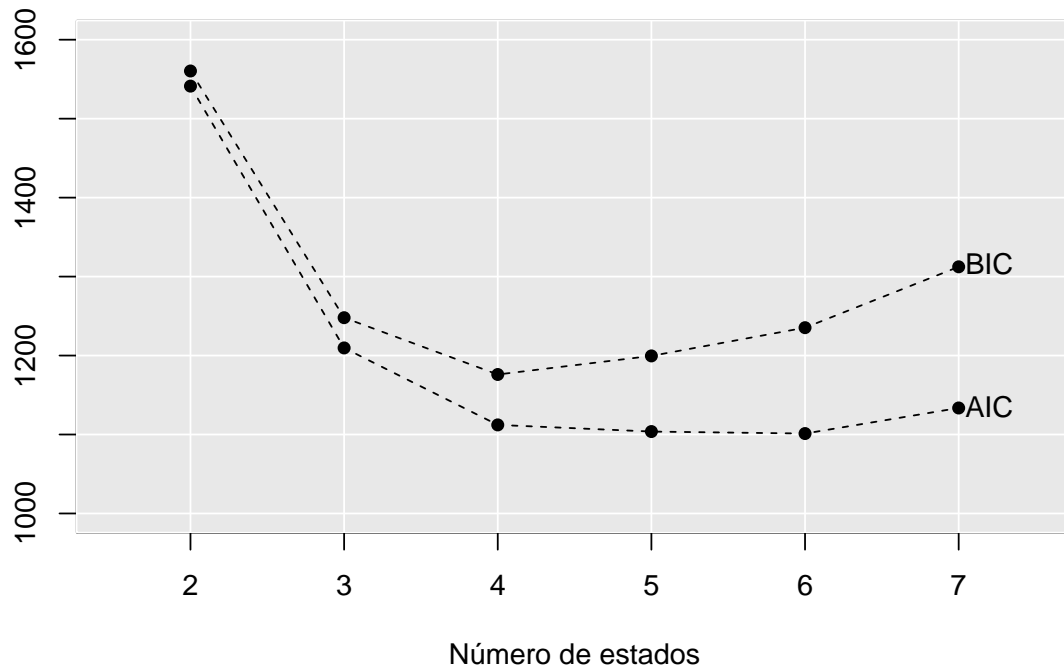


Figura 16: Serie incendios: selección de modelos AIC y BIC.

```
## [1] "Be careful of label switching when interpreting states!"
```

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.

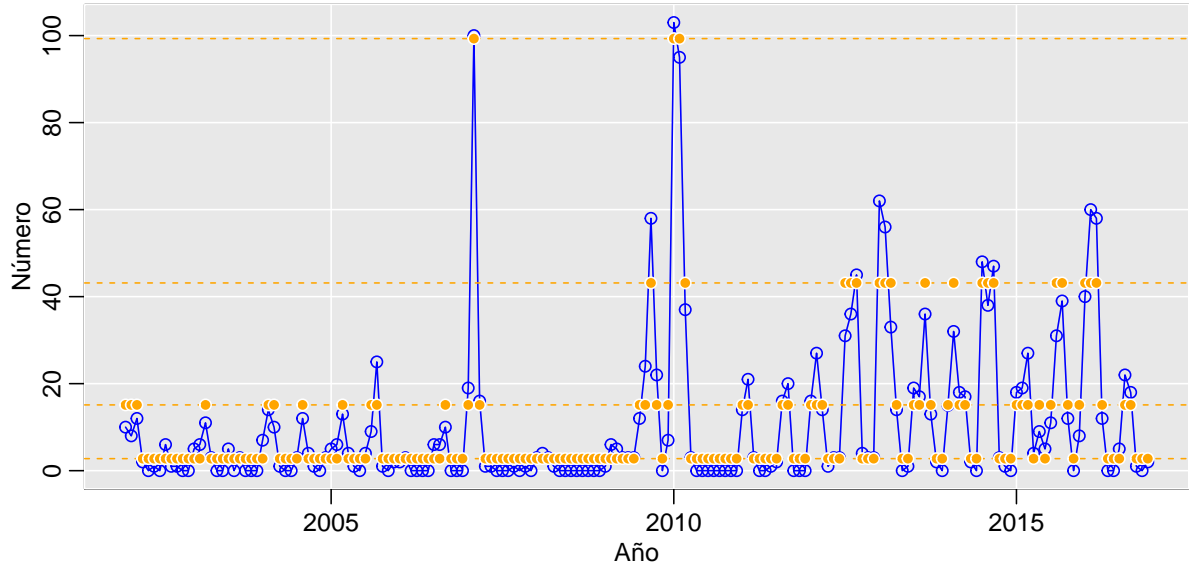


Figura 17: Algoritmo Viterbi aplicado al ZIPHMM de tres estados.

	Media	Err.Sta	Desv	2.5 %	25 %	50 %	75 %	97.5 %	n.eff	Rhat
θ	0.449	0.001	0.050	0.351	0.414	0.449	0.483	0.545	5219.355	0.999
λ_1	2.868	0.004	0.266	2.371	2.684	2.866	3.047	3.408	4810.999	0.999
λ_2	15.369	0.012	0.775	13.859	14.830	15.361	15.874	16.877	4492.668	0.999
λ_3	43.153	0.040	1.807	39.600	42.040	43.143	44.300	46.658	2036.872	1.001
λ_4	99.235	0.145	6.120	87.428	95.445	99.219	103.236	110.912	1790.696	1.001
Γ_{11}	0.795	0.001	0.037	0.719	0.772	0.797	0.821	0.861	4587.670	0.999
Γ_{12}	0.158	0.000	0.034	0.098	0.134	0.156	0.180	0.231	4676.747	0.999
Γ_{13}	0.034	0.000	0.017	0.009	0.022	0.032	0.044	0.076	4870.506	1.000
Γ_{14}	0.013	0.000	0.011	0.000	0.005	0.010	0.018	0.040	3815.711	0.999
Γ_{21}	0.470	0.001	0.079	0.318	0.415	0.469	0.523	0.630	5238.938	0.999
Γ_{22}	0.328	0.001	0.077	0.192	0.272	0.325	0.380	0.482	5096.108	0.999
Γ_{23}	0.146	0.001	0.058	0.053	0.104	0.140	0.183	0.274	4306.099	1.000
Γ_{24}	0.056	0.001	0.037	0.008	0.029	0.049	0.076	0.147	4611.951	1.000
Γ_{31}	0.188	0.001	0.082	0.056	0.127	0.179	0.238	0.376	5736.173	1.000
Γ_{32}	0.331	0.001	0.099	0.151	0.261	0.325	0.396	0.537	5985.893	0.999
Γ_{33}	0.436	0.001	0.100	0.249	0.367	0.434	0.504	0.634	5179.787	1.000
Γ_{34}	0.045	0.001	0.043	0.001	0.013	0.032	0.063	0.161	3481.429	1.000
Γ_{41}	0.137	0.002	0.119	0.004	0.045	0.102	0.200	0.435	4984.285	0.999
Γ_{42}	0.309	0.002	0.163	0.055	0.182	0.287	0.421	0.658	5002.104	0.999
Γ_{43}	0.280	0.002	0.162	0.041	0.155	0.253	0.388	0.639	4855.509	1.000
Γ_{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 16: Estimación bayesiana de los parámetros para un ZIPH-MM.

```
##### Packages ###  
library(Bayeshmmcts)  
library(broom)  
  
##### Data #####  
data("homicidios")  
data("incendios")  
  
##### Poisson - Hidden Markov Model #  
  
##### Zero Inflated Poisson - Hidden Markov Model #
```