Resultados

Rafael Eduardo Diaz

18 de junio de 2019

1. Aplicación e ilustración

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 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.
[1] ".preformat.ts"

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2002	0	1	4	0	0	0	0	0	0	0	0	0
2003	0	2	1	0	0	0	0	0	0	0	0	0
2004	1	6	3	0	0	0	0	1	0	0	0	0
2005	0	0	1	0	0	0	0	1	2	0	0	0
2006	0	0	0	0	0	0	2	0	0	0	0	0
2007	3	23	3	0	0	0	0	0	0	0	0	0
2008	0	1	1	0	0	0	0	0	0	0	0	0
2009	0	0	0	0	0	0	1	2	2	1	0	1
2010	20	16	6	0	0	0	0	0	0	0	0	0
2011	1	10	0	0	0	0	0	0	0	0	0	0
2012	1	3	3	0	0	0	2	1	8	0	0	0
2013	4	6	2	0	0	0	0	1	3	0	0	0
2014	4	4	11	4	0	0	5	4	3	0	0	0
2015	1	2	3	1	0	0	0	6	6	0	0	0
2016	1	10	18	0	0	0	0	2	1	0	0	0

Tabla 2: Número de incendios 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)

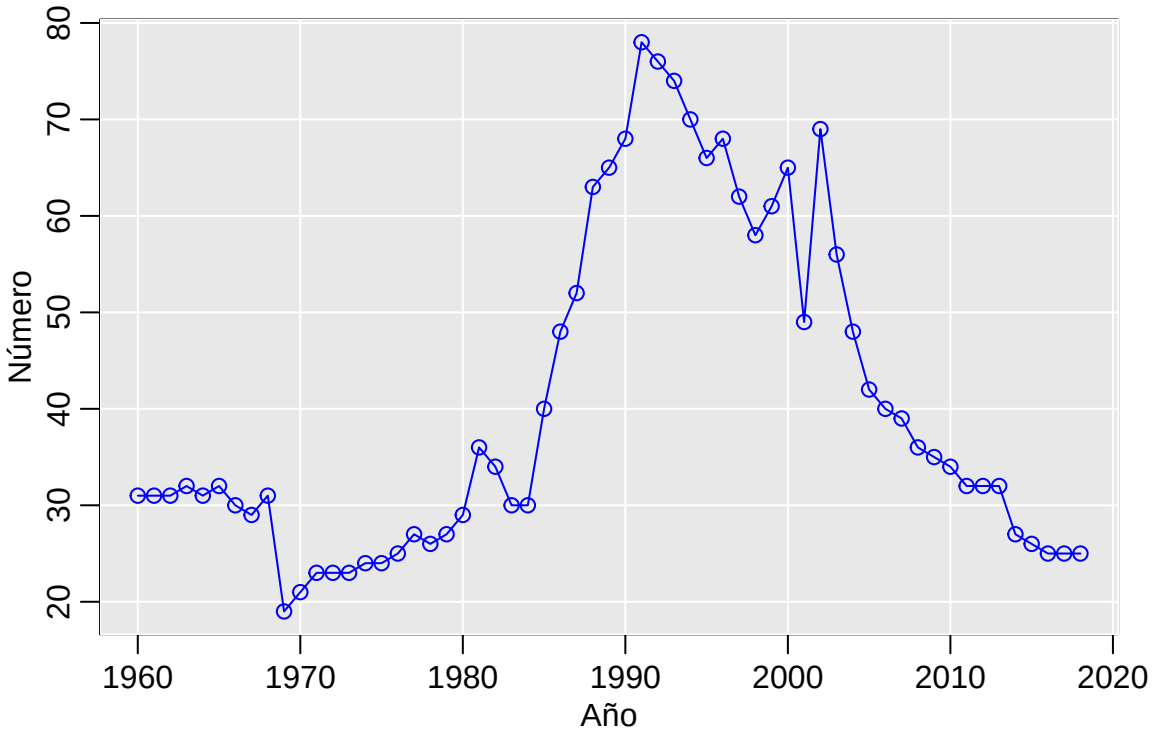


Figura 1: Serie de tiempo homicidios en Colombia desde el año 1960 hasta el año 2018.

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 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 homicidio40/100.000 en estas últimas seis décadas, con cifras comparables a las de países con guerra civil declarada. (Franco, 1980).

En la figura 2, se encuentra gráficada 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 seria 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.

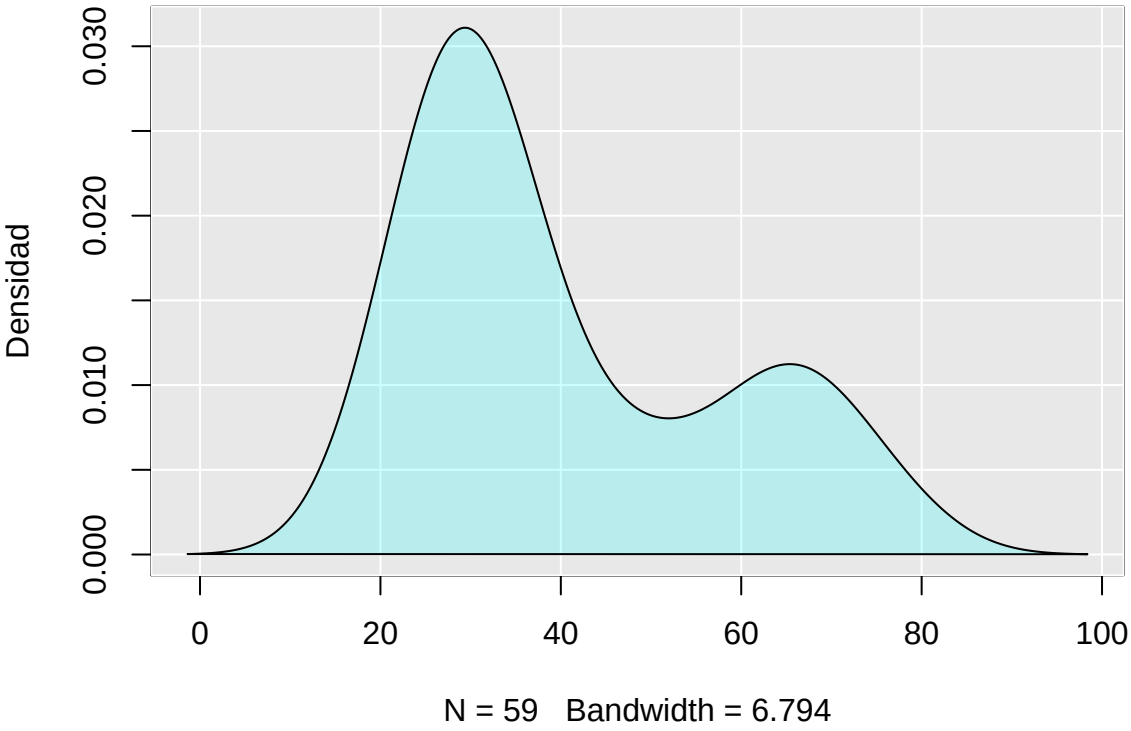


Figura 2: Kernel Densidad de homicidios en Colombia (1960-2018).

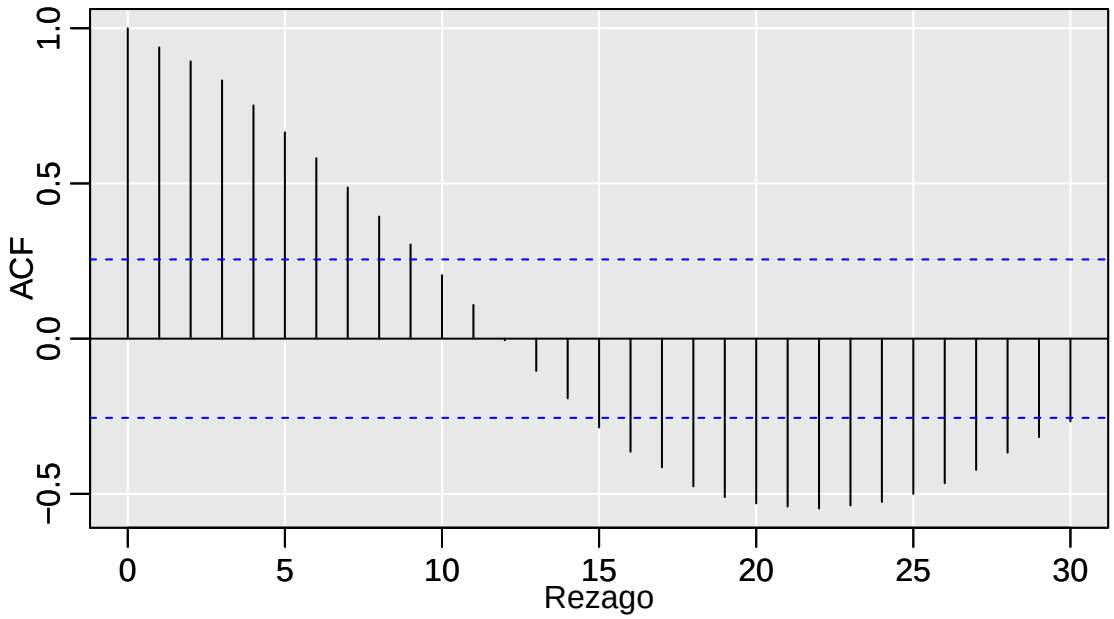


Figura 3: Función de autocorrelación muetral, para la seerie de homicidios.

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.

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 modelo son las siguientes la media de los estados dependientes $\lambda = (29.72, 62.8)$ y los valores de la distribución estacionaria son $\pi = (0.764, 0.236)$. Ahora veamos la estimación de la matriz de transición para el modelo con 2 estados:

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

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 son los siguientes:

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.

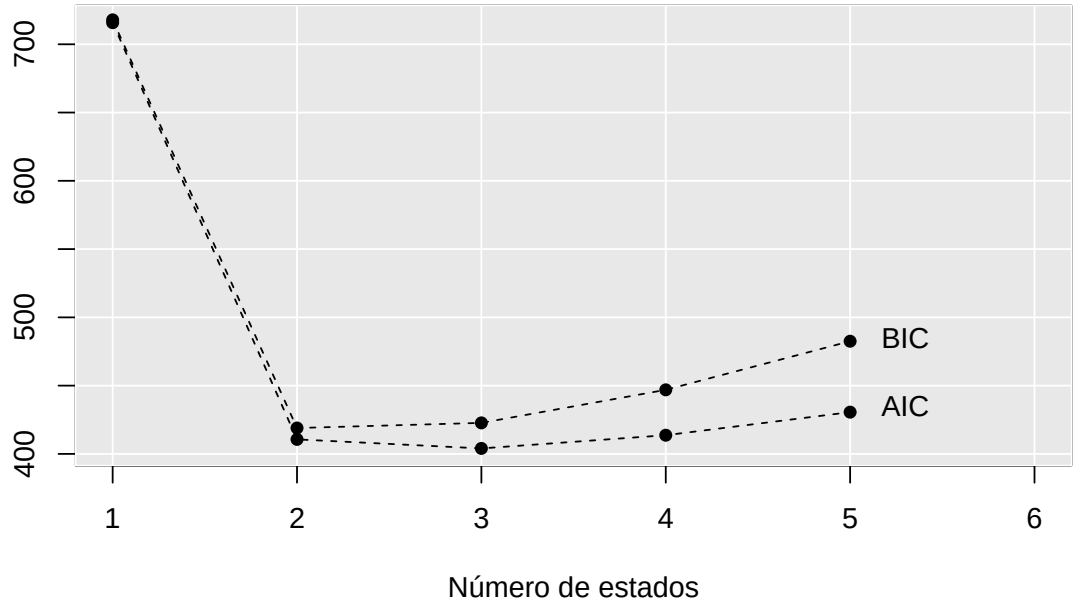


Figura 4: Serie homicidios: selección de modelos AIC y BIC.

	1	2	3	4	5	6	7	8	9	10	11	12
observaciones	0.94	0.89	0.83	0.75	0.67	0.58	0.49	0.39	0.30	0.20	0.11	-0.01
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	0.49	0.47	0.44

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

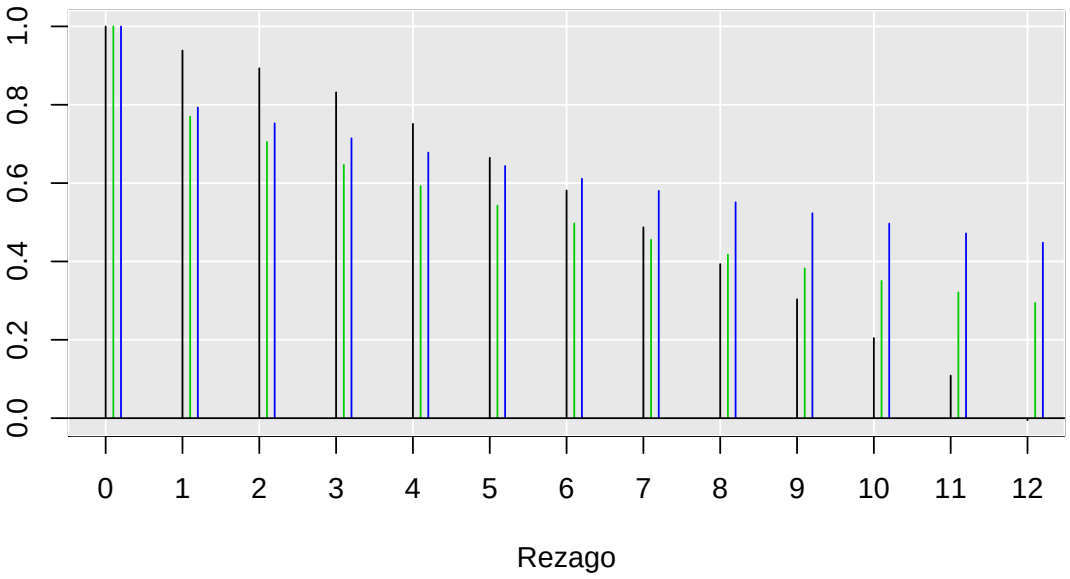


Figura 5: Datos homicidios: ACF y ACF de los PHMM con dos y tres estados.

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.

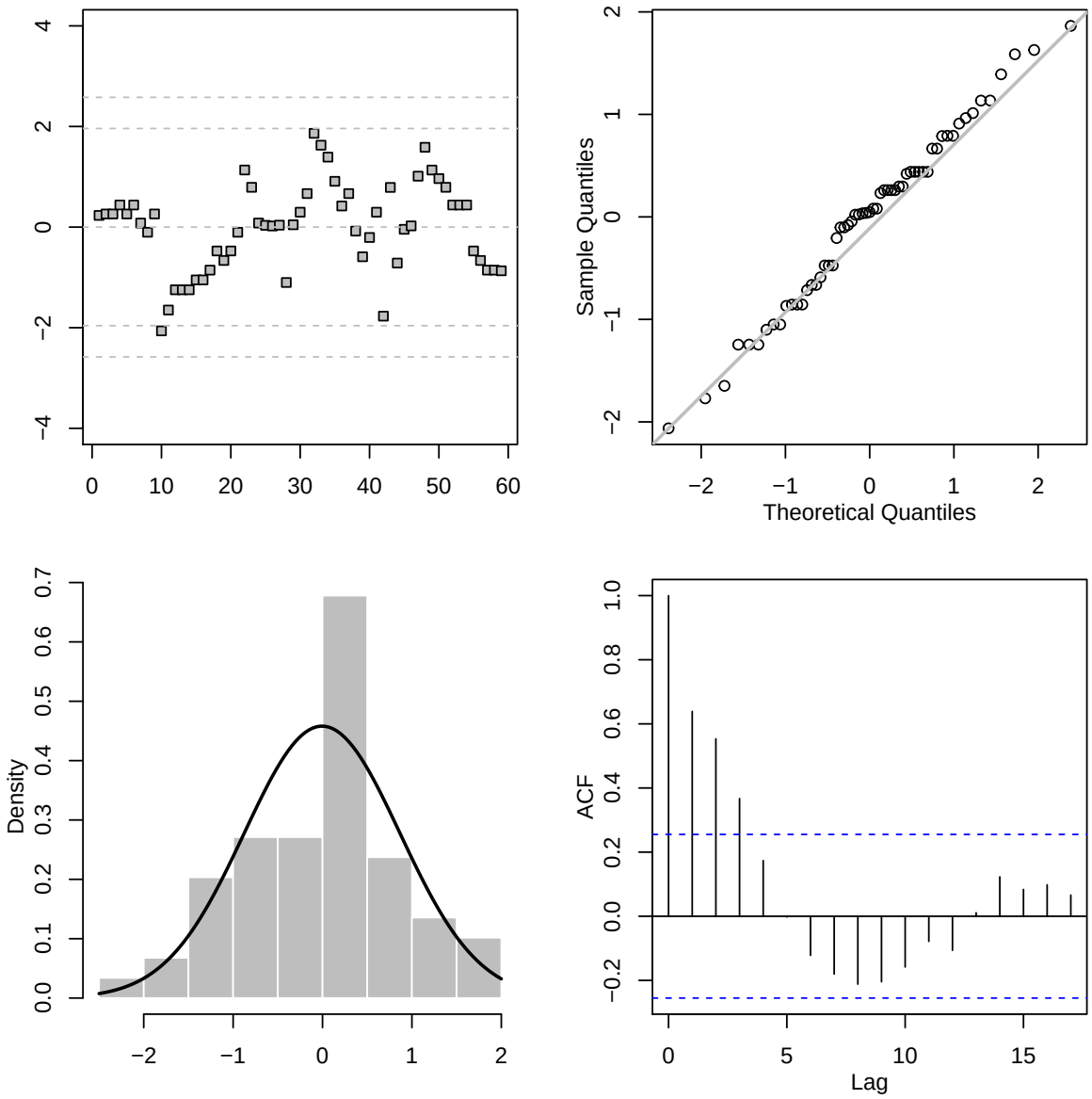


Figura 6: Grafico pseudo-residuales ordinarios para el PHMM de 2 estados.

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-

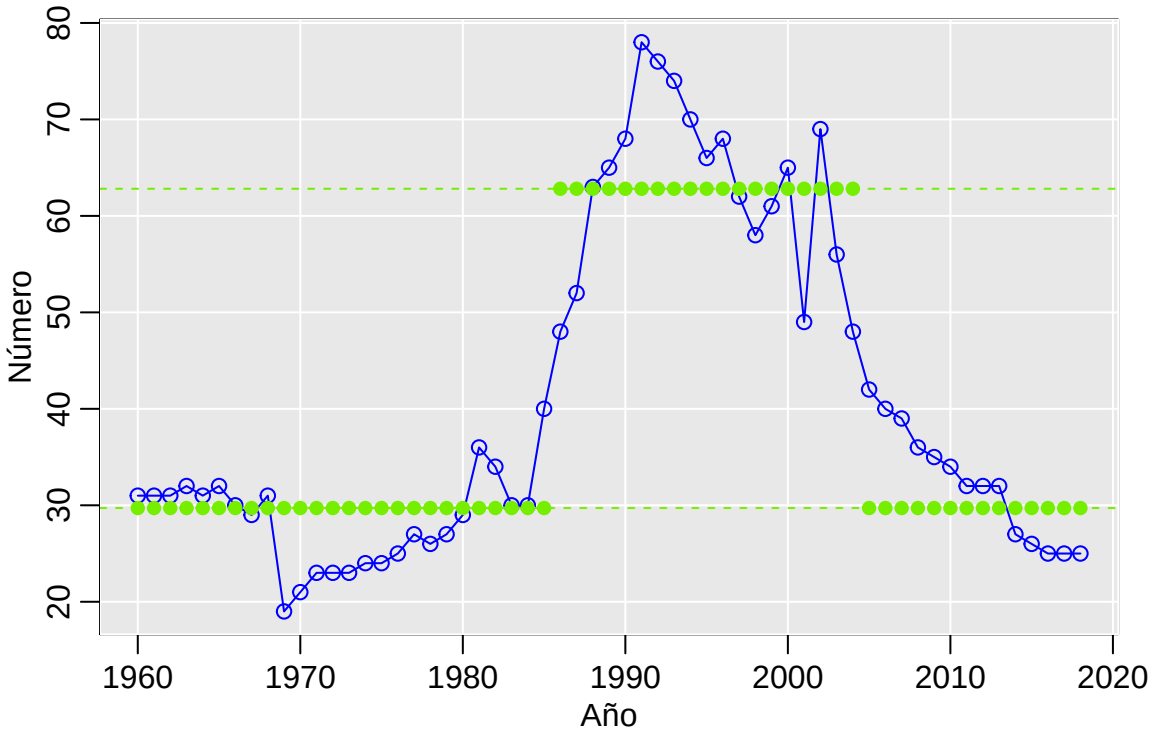


Figura 7: Algoritmo Viterbi aplicado a un PHMM de dos estados.

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

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	

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

El algoritmo Viterbi, permite realizar la decodificación global de los estados clasificando a cada una de las observaciones en su correspondiente estado, dando como resultado la anterior tabla, para un total de 40 observaciones en el estado 1 y 19 en el estado 2. En la grafica 5 se visualiza el algoritmo viterbi.

Ahora realizaremos la predicción de los estados para un rezago $h = 12$, dado y de la distribución para una año en especifico.

	Estado 1	Estado 2
2019	0.9802	0.0198
2020	0.9621	0.0379
2021	0.9456	0.0544
2022	0.9304	0.0696
2023	0.9164	0.0836
2024	0.9037	0.0963
2025	0.8920	0.1080
2026	0.8813	0.1187
2027	0.8714	0.1286
2028	0.8624	0.1376
2029	0.8542	0.1458
2030	0.8467	0.1533

Tabla 7: Predicción de las probabilidades para un rezago h = 12.

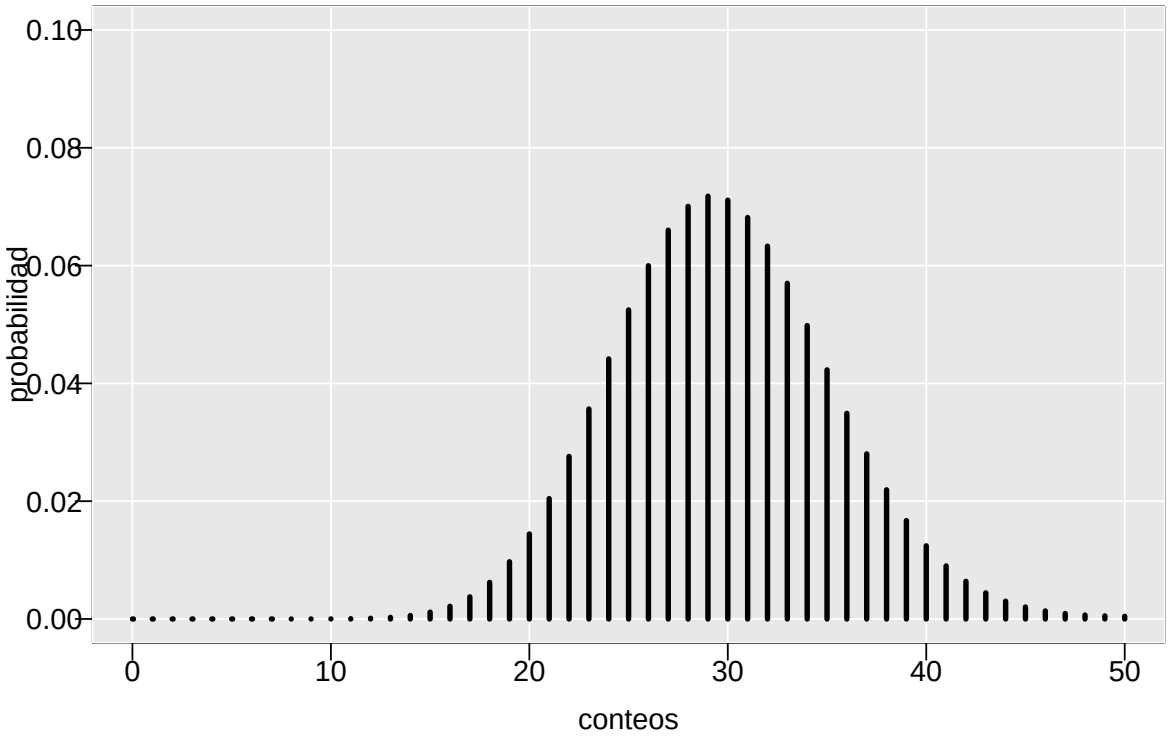


Figura 8: Serie homicidios: distribución de pronóstico para el año 2019.

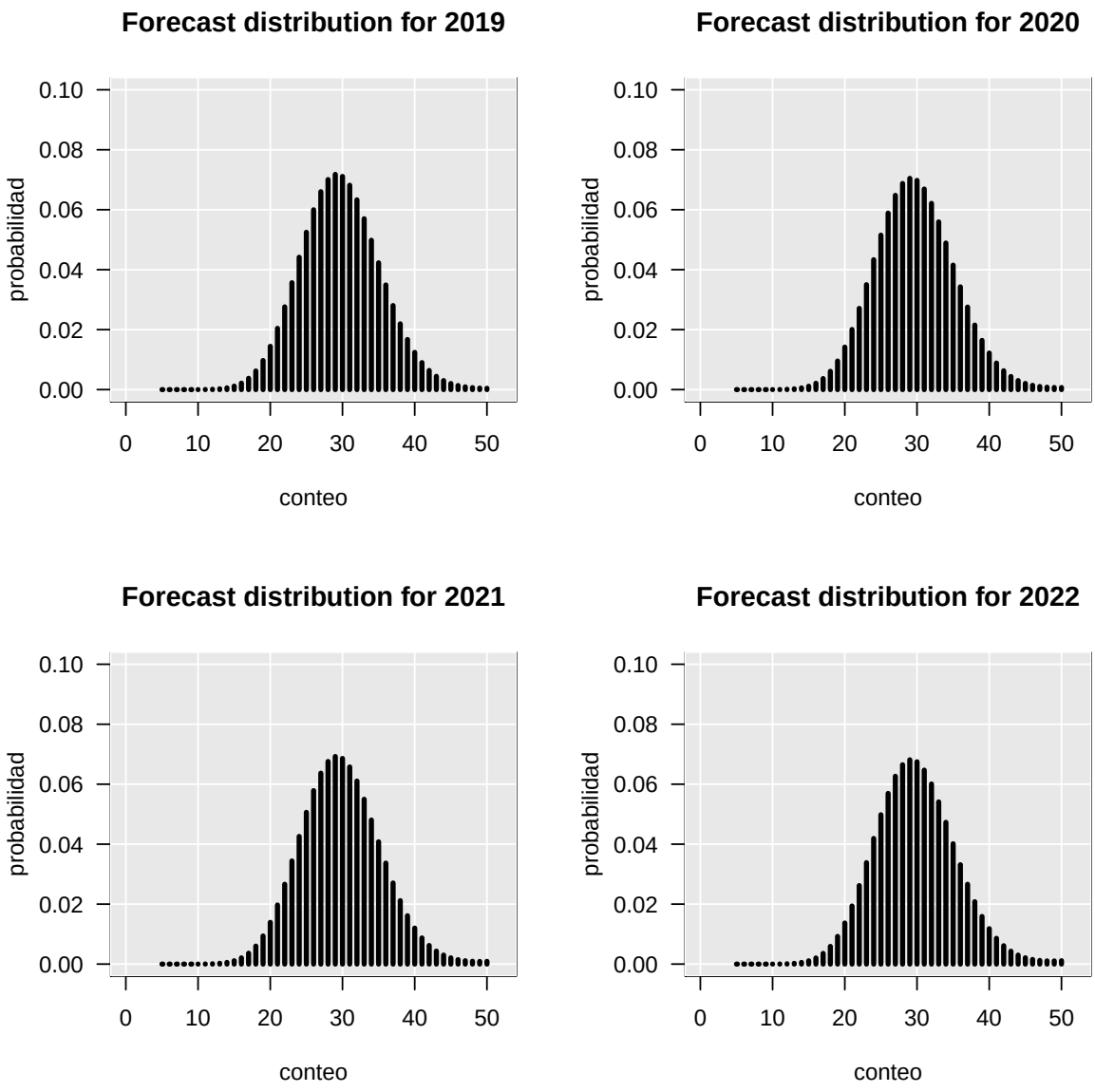


Figura 9: Pronostico de la distribución para los años 2019 a 2022.

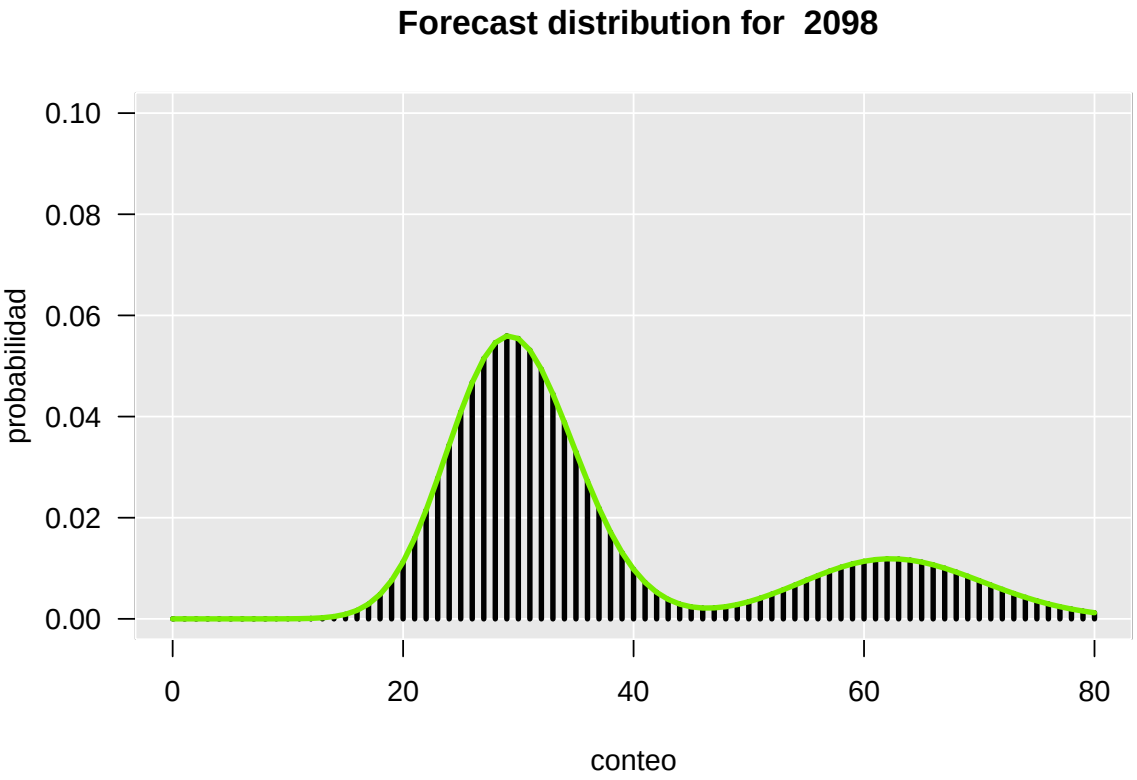


Figura 10: La distribución prónostica

Estimación Bayesiana del PHMM

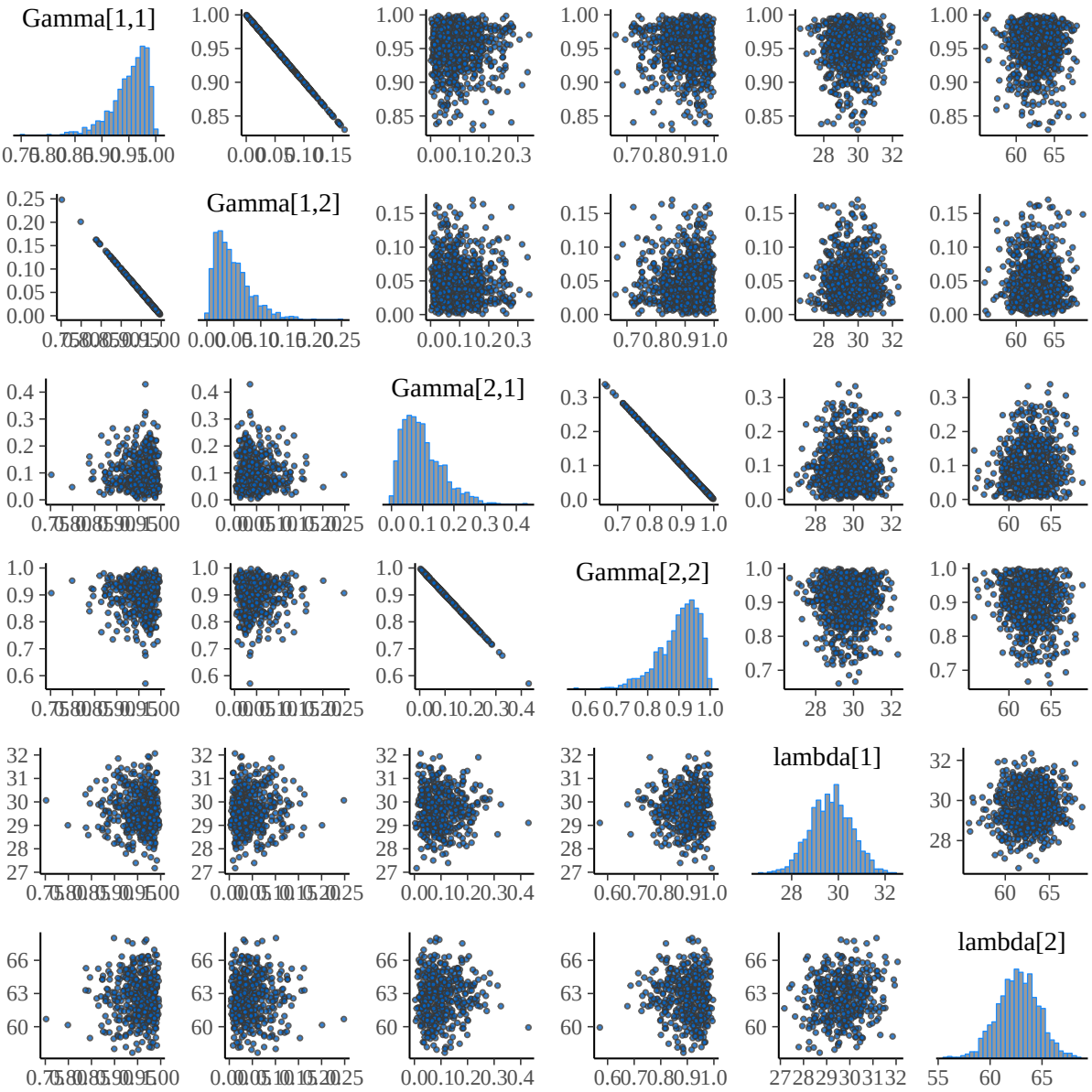
Primero se ajustaron cuatro modelos, con la función 'Bayeshmmcts::bayes.PHMM', para 2, 3, 4 y 5 estados, después, se estimo la log-verosimilitud marginal, utilizando muestreo por puente como alternativa a las propuesta hecha por Newton y Raftery(1994) que sugiere utilizar la verosimilitud integrada, 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 consitente tiene un gran problema varianza infinita. Mientras que el estimador de muestreo por puente, no presenta ese problema además de su facil implementación, pues esta metodología se puede ejecutar con la función 'bridge$_s ampler()$'$delpaquete$'$bridgesampling$', $delautorGronau.Inclusivecalcularelerrorparalaverosimilitudmarginal, q$ $estadosdeaparejas, yseleccionarelmásadecuado, lasiguientetablailustraelcontrastedehipótesis$:

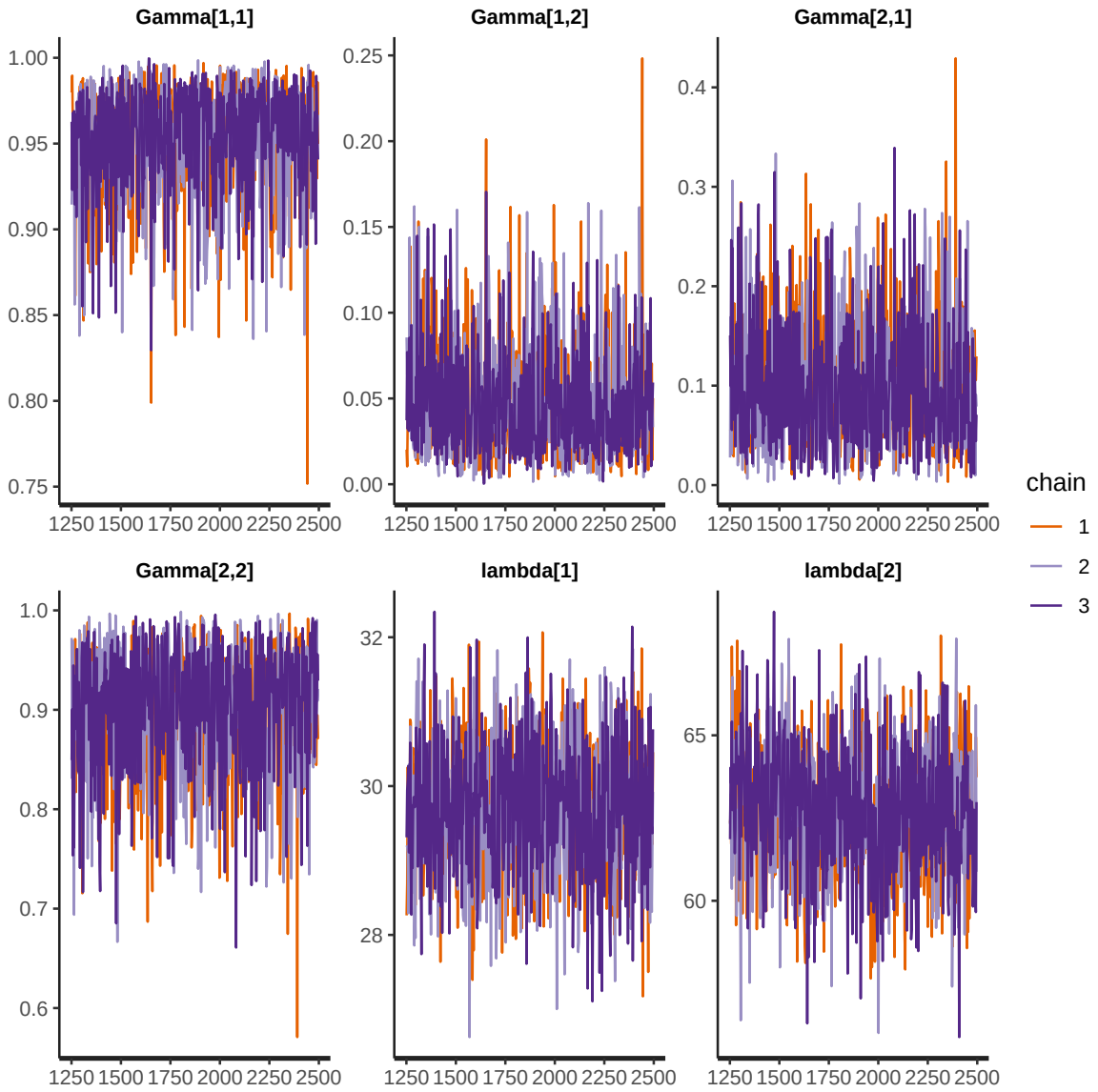
	mod 2 Estados	mod 3 Estados	mod 4 Estados	mod 5 Estados
mod 2 Estados		3.11	3047.46	390147608.00
mod 3 Estados			980.61	125542040.00
mod 4 Estados				128023.00

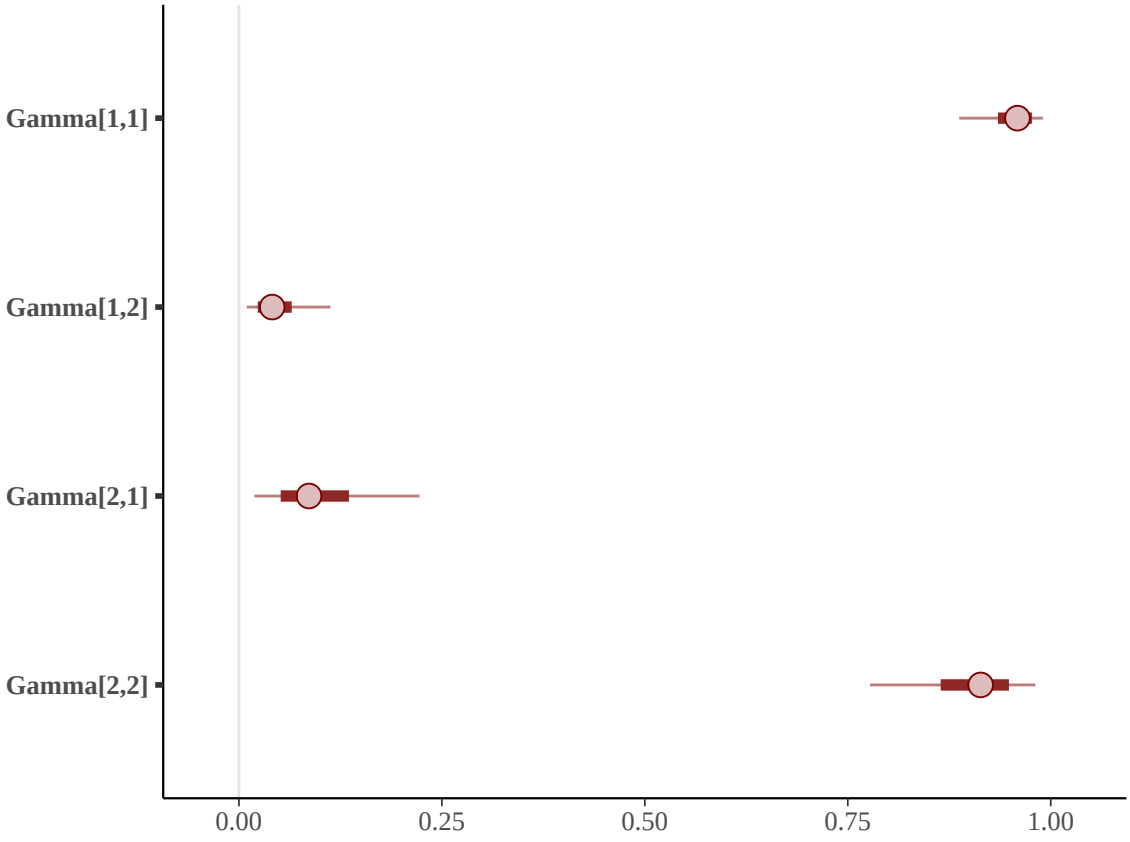
Tabla 8: Comparación resultados Factor de Bayes.

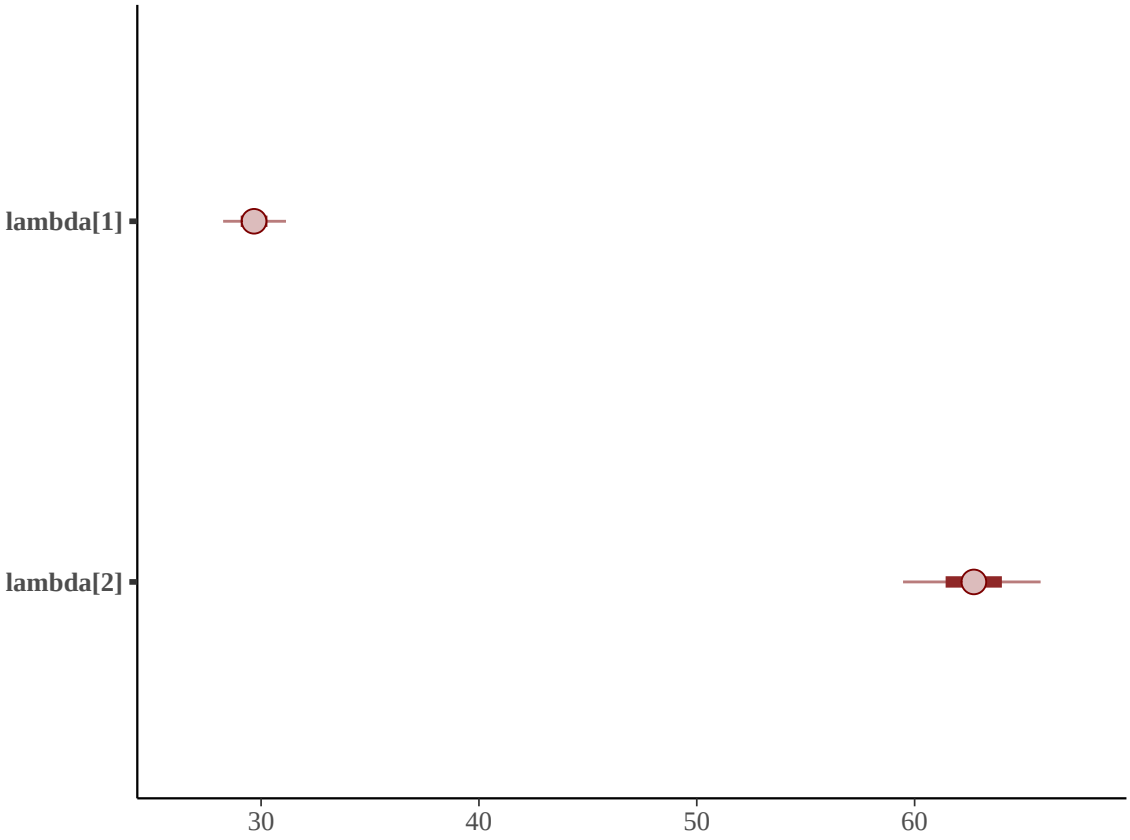
Por lo tanto se elige un PHMM de orden 2 lo cual coincide con la teoría clásica. A continuación mostramos las estimaciones bayesianas de la matriz de transición, y la media de los estados dependiesntes:

	mean	se_mean	sd	2.5 %	25 %	50 %	75 %	97.5 %	n_eff	Rhat
Gamma[1,1]	0.95	0.00	0.03	0.87	0.93	0.96	0.98	0.99	1180.11	1.00
Gamma[1,2]	0.05	0.00	0.03	0.01	0.02	0.04	0.07	0.13	1180.11	1.00
Gamma[2,1]	0.10	0.00	0.06	0.01	0.05	0.09	0.14	0.25	1232.99	1.00
Gamma[2,2]	0.90	0.00	0.06	0.75	0.86	0.91	0.95	0.99	1232.99	1.00
lambda[1]	29.68	0.03	0.88	28.00	29.06	29.68	30.29	31.41	1193.62	1.00
lambda[2]	62.69	0.06	1.95	58.98	61.42	62.72	64.01	66.46	1089.96	1.00
lp__	-210.57	0.05	1.50	-214.33	-211.20	-210.24	-209.52	-208.75	1068.37	1.00


Gamma[1,1]
Gamma[1,2]
Gamma[2,1]
Gamma[2,2]
lambda[1]
lambda[2]


Gamma[1,1]
Gamma[1,2]
Gamma[2,1]
Gamma[2,2]
lambda[1]
lambda[2]
1.00
0.95
0.90
0.85
0.80
0.75
0.25
0.20
0.15
0.10
0.05
0.00
0.4
0.3
0.2
0.1
0.0
1.0
0.9
0.8
0.7
0.6
32
30
28
65
60
1250 1500 1750 2000 2250 2500
chain
1
2
3


Gamma[1,1]
Gamma[1,2]
Gamma[2,1]
Gamma[2,2]
0.00
0.25
0.50
0.75
1.00


lambda[1]
lambda[2]
30
40
50
60

```
##
##            Stationarity start     p-value
##            test         iteration
## Gamma[1,1] passed       1         0.0765
## Gamma[2,1] passed       1         0.1884
## Gamma[1,2] passed       1         0.0765
## Gamma[2,2] passed       1         0.1884
## lambda[1]  passed       1         0.8587
## lambda[2]  passed       1         0.5854
## lp__       passed       1         0.7470
##
##            Halfwidth Mean      Halfwidth
##            test
## Gamma[1,1] passed       0.9518 0.00183
## Gamma[2,1] passed       0.0990 0.00350
## Gamma[1,2] passed       0.0482 0.00183
## Gamma[2,2] passed       0.9010 0.00350
## lambda[1]  passed      29.6772 0.04900
## lambda[2]  passed      62.6925 0.11668
## lp__       passed    -210.5730 0.09063
```

1.2. Modelo Poisson Cero inflado - Oculto de Markov

	Fecha	GIF
1	2002-01	0
2	2002-02	1
3	2002-03	4
4	2002-04	0
5	2002-05	0
6	2002-06	0

Tabla 9: Grandes Incendios forestales, en Colombia desde enero del 2002 a diceiembre del 2016.

```
##
##   0   1   2   3   4   5   6   8  10  11  16  18  20  23
## 124  19   9   8   6   1   5   1   2   1   1   1   1   1
##
##     0     1     2     3     4     5     6     8    10    11    16    18
## 0.689 0.106 0.050 0.044 0.033 0.006 0.028 0.006 0.011 0.006 0.006 0.006
##    20    23
## 0.006 0.006
```

2. Anexo Códigos

```
################
### Packages ###
################
library(Bayeshmmcts)
library(broom)
```

```
################
##### Data #####
################
data("homicidios")
data("incendios")

#####################################################
########## Poisson - Hidden Markov Model ##########
#####################################################


rm(list = ls())

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

incendios <- readRDS("incendios.rds")
GIF <- ts(data = incendios$GIF, start = c(2002,1), frequency = 12)
GIF
```