Universidad Nacional Autónoma de México

Facultad de Ciencias

Estadística Bayesiana.

Proyecto final.

Integrantes:

Aburto Hernández N. Carolina, N.L 1
Miranda Peñafiel Melissa Sofía, N.L 38
Pacheco Martínez Mariana, N.L 45
Pérez García Diego Eduardo, N.L 47
Tufiño Villegas Julio Ernesto, N.L 62
Zeferino Alvarado Rodrigo Emmanuel, N.L 66

30 Enero 2021

Introducción

En términos económicos, como mercado, el mercado de viviendas suele determinar sus precios por medio de la oferta y la demanda. Sin embargo, a pesar de poder ser alcanzado el equilibrio en el mercado, esto puede conllevar algunos problemas sociales como la inaccesibilidad a una vivienda.

En el caso particular de Londres, Inglaterra, se ha visto que en las últimas décadas los precios se han elevado de manera significativa: tan sólo a finales del año pasado, se rebasó por primera vez en la historia, el margen de £500,000 como costo promedio por propiedad. Esto convierte a Londres en la región más costosa para vivir de todo Reino Unido.

Refiriendonos nuevamente al problema de la inaccesibilidad a una vivienda, cabe ser señalado que el ritmo de crecimiento de los precios de las viviendas ha llegado a exceder el incremento en los ingresos individuales. Incluso, en 2014 se llegó a tener viviendas con costos 10 veces superiores al salario promedio, a diferencia de 1997, cuando estos eran sólo 4 veces superiores.

Derivado de la pandemia ocasionada por el virus COVID-19, se ha reportado que durante este periodo se ha visto una concentración de la demanda de viviendas en Inglaterra. Esto tiene diversos orígenes: En el caso de aquellos que son compradores primerizos, al no tener las complicaciones una persona que decide mudarse, como pued ser: asegurar tener una vivienda destino, concertar los arreglos y pagos y contemplar arreglos y remodelaciones; se ha visto incentivada la compra de propiedas, que normalmente están en el extremo inferior de precios de las viviendas. Esto a su vez, ha generado un aumento en la demanda y por tanto, elevado los precios.

Por otro lado, otro factor que ha tenido relevancia en el aumento de precios, son los cambios de preferencias que han tenido las personas respecto a sus viviendas, derivado de las estrictas medidas de confinamiento que el gobierno decretó a finales de Marzo de 2020

Base y análisis descriptivo

Se utilizó una base de datos que contiene información sobre las viviendas en Londres. Las variables que contiene son fecha, área, precio promedio de las viviendas (registrado en £GBP), código del área, número de viviendas vendidas y un indicador sobre si el área es un *borough* de Londres o no.

Los datos son actualizados cada mes desde enero de 1995 hasta enero de 2020 y se consideran 45 regiones de Londres, por lo tanto, cada región cuenta con 301 datos por variable.

Para los efectos de este proyecto, analizaremos la serie de tiempo del precio promedio de las viviendas para hacer una predicción. En particular trabajaremos con la región de Westminster. Entonces, para la variable *average_price* para la región escogida tenemos:

	Mínimo	Mediana	Media	Máximo
Precio promedio	131,468	502,387	543,866	1,117,408

Como es de esperarse, el mínimo corresponde a la primera observación (i.e. enero de 1995) y el máximo se alcanzó en febrero de 2018.

Al ser una serie de tiempo, nos van a importar más los datos más recientes. Para los datos a partir del 2018 tenemos que la media es de 990,359 entonces esperaríamos que nuestras predicciones ronden este valor.

Descripción del proyecto

Series de tiempo desde el enfoque Clásico:

En el enfoque clásico de series de tiempo, a grandes rasgos, se tiene una serie de valores correlacionados entre sí, debido a la dependencia que tienen respecto del tiempo. A través de las distintas técnicas, se busca analizar la información que se tiene, con el fin de identificar un patrón que permita describir esa información a lo largo del tiempo. Posteriormente se busca extender ese patrón a un periodo de tiempo determinado para llevar a cabo un pronóstico.

Series de tiempo desde el enfoque Bayesiano:

En el caso de los modelos Bayesianos de series de tiempo, encontramos que estos tienen una estructura matemáticamente igual a los del enfoque clásico, como el modelo de Box-Jenkins, pero difieren en que en la estimación de los parámetros, estos son considerados variables aleatorias y como tales tienen asociado un espacio de probabilidad.

Comparación para los modelos de series de tiempo Clásico vs. Bayesiano:

Para el modelo **Clásico** se buscará:

- En primer lugar hacer una prueba de Breusch-Pagan para verificar la homocedasticidad del modelo de las observaciones.
- En base al resultado anterior, de existir heterocedasticidad, se le aplicará una transformación Box-Cox a los datos para corregir la heterocedasticidad y nuevamente se hará una prueba de Breusch-Pagan para comprobar nuevamente esta condición.
- Posteriormente, de existir homocedasticidad, se utilizará una diferenciación para conferir estacionariedad a la serie, pues esta es necesaria para hacer ARMA o ARIMA. Luego, se realizarán los tests de Dickey-Fuller y KPSS para comprobar que exista estacionariedad en la serie.
- Finalmente, en base a los resultados anteriores, se propondrá un ajuste ARIMA o SARIMA y se hace una predicción para 5 valores futuros.

Para el modelo **Bayesiano** se buscará:

- En primer lugar, cabe señalar que se trabajará con la serie ya diferenciada que empleamos con el modelo clásico.
- Entonces, se definirán los valores iniciales en los parámetros para las distribuciones a priori; esto tendrá como objetivo estimar los promedios móviles.
- Se realizará el quemado de datos y se establecerá el número de cadenas de Markov para la estimación.
- Se comprobará la convergencia de los parámetros de manera gráfica y mediante una prueba de Gelman.
- Finalmente, se hace una predicción para 5 valores futuros.

Para terminar el procedimiento anterior, se comparan el modelo Clásico y Bayesiano en función de los criterios de bondad de ajuste, y se determina cuál ofrece un mejor modelo.

Finalmente, se propondrá un modelo Bayesiano nuevo que busque encontrar una mejor estimación de la varianza, realizando varias pruebas de diversos modelos y comparándolos en función de sus p-values y considerando que no existe correlación entre sus parámetros.

Serie de tiempo

Una serie de tiempo es la sucesión de observaciones generadas por un proceso estocástico, cuyo índice se toma en relación con el tiempo.
En las series de tiempo se supone que existe una estructura de correlación entre dos observaciones, no son independientes.

Realizamos una serie de tiempo para la columna de precios de la región Westminster.

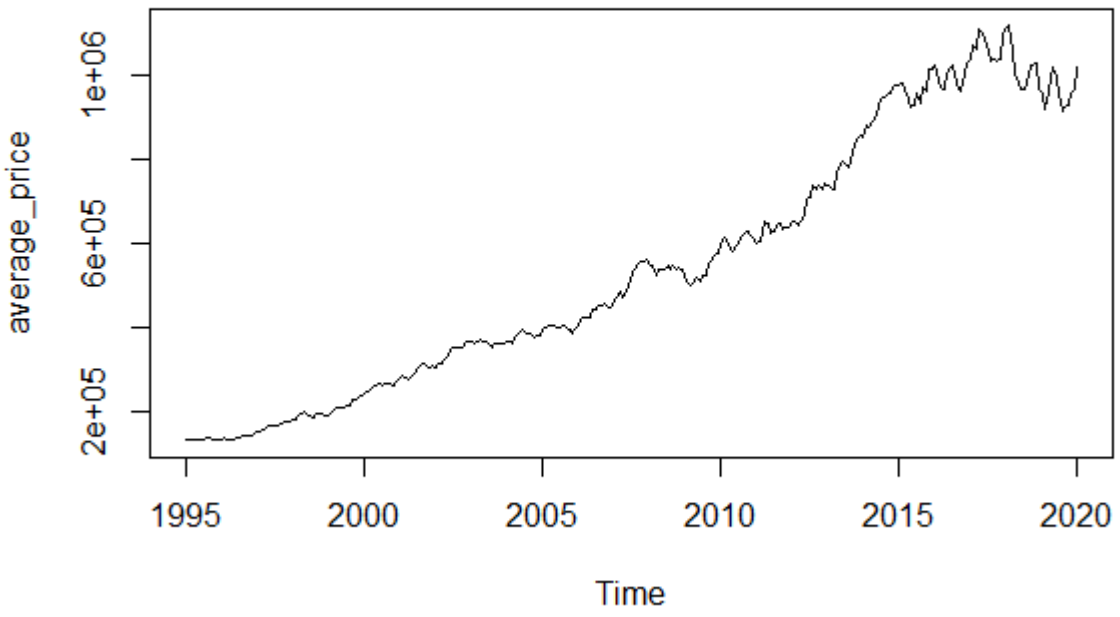

Figura 1: Serie de tiempo de los precios de la región de Westminster.

Para las series de tiempo necesitamos que tengan varianza constante, así realizamos una **prueba homosedástica**, bajo el *Test de Breusch-Pagan*, suponiendo:

H_0 = Los datos son homocedásticos (tienen varianza constante).
H_1 = Los datos son heterocedásticos (la varianza no es constante).

Obtuvimos un p-value < 0.05 así que los datos no son homocedásticos.

Así con la *transformaciónn de BoxCox* volvimos homocedástica (de varianza constante) nuestra serie de tiempo. El comando BoxCox aplica una transformación a los datos acorde a un parámetro lambda que nosotros le damos, para encontrar un valor de lambda que genere una transformación adecuada a los datos, usamos el comando *BoxCox.lambda*, entonces encontramos lambda con *BoxCox.lambda* y transformamos los datos con BoxCox usando el parámetro lambda.

Todo esto bajo el *método Guerrero*, que es simplemente la forma en que se va a calcular lambda. Podíamos haber usado el método *loglike* cambiando "guerrero"por "loglik", sin embargo esto cambiaría el valor de lambda y consecuentemente la transformación, como nos funcionó bien "guerrero", nos quedamos con este método.

Finalmente volvimos a realizar el test y con un p-value $= 0.2557 \Rightarrow$ aceptamos H_0, *i.e* nuestros datos ya son homocedásticos.

Luego realizamos las **pruebas de estacionariedad**, usando dos test: *Test de Dickey-Fuller* y *Test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS)*. Como no pasó el de D-F realizamos una differenciación y así pasó la prueba de estacionariedad, de igual manera pasó el test de KPSS.

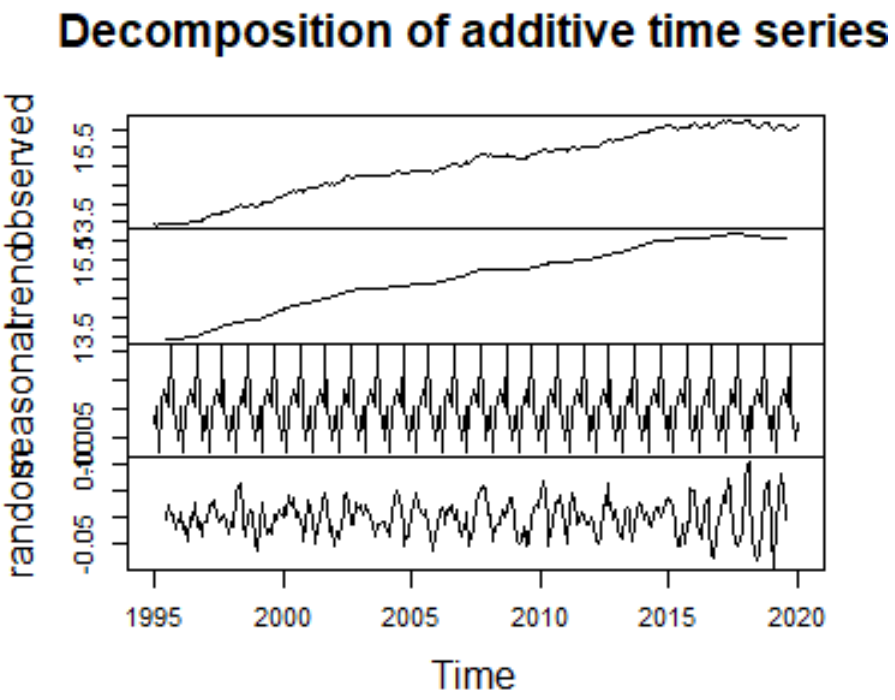

Figura 2: Descomposición de la serie de tiempo.

Usamos los ACF y PACF muestrales para darnos una idea del numero de retrasos que podriamos proponer para el ajuste con un ARIMA o SARIMA. Teóricamente un AR(p) Tendra los primeros p lags del

PACF por fuera de las bandas de confianza y después los lags tenderan rápidamente a cero, análogamente un MA(q) tendra los primeros p lags del PACF por fuera de las bandas de confianza y después los lags tenderan rápidamente a cero, de ahi se concluye que un ARMA(p,q) cumplirá ambas condiciones.

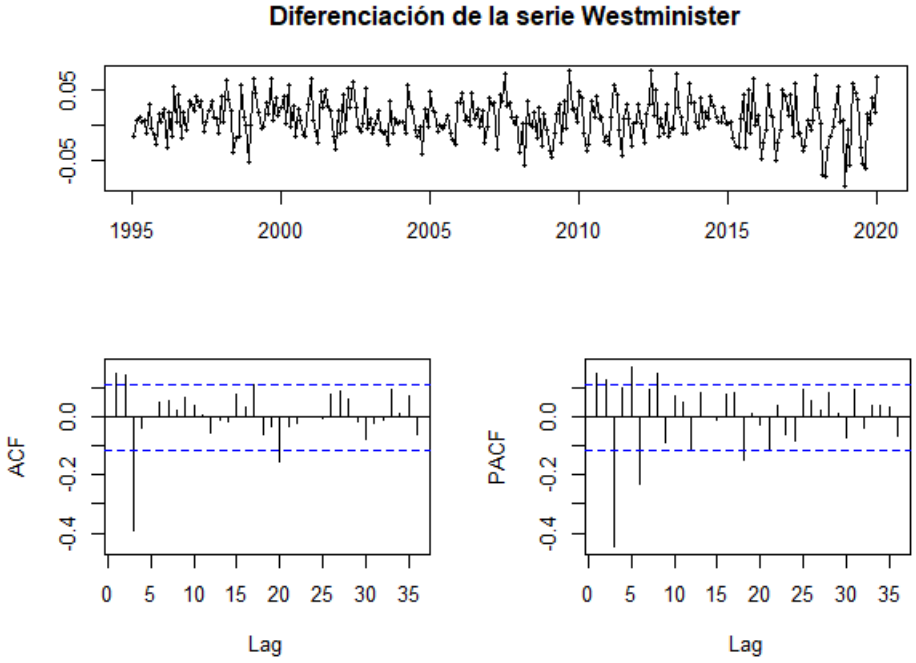

Figura 3: Diferenciación de la serie de tiempo.

Por medio de *auto.arima* obtuvimos el modelo: ARIMA(2,0,3) WITH NON-ZERO MEAN con $AIC = -1402.07$ $AICc = -1401.68$ y $BIC = -1376.14$.

Finalmente usamos *forecast* para obtener la gráfica de la predicción.

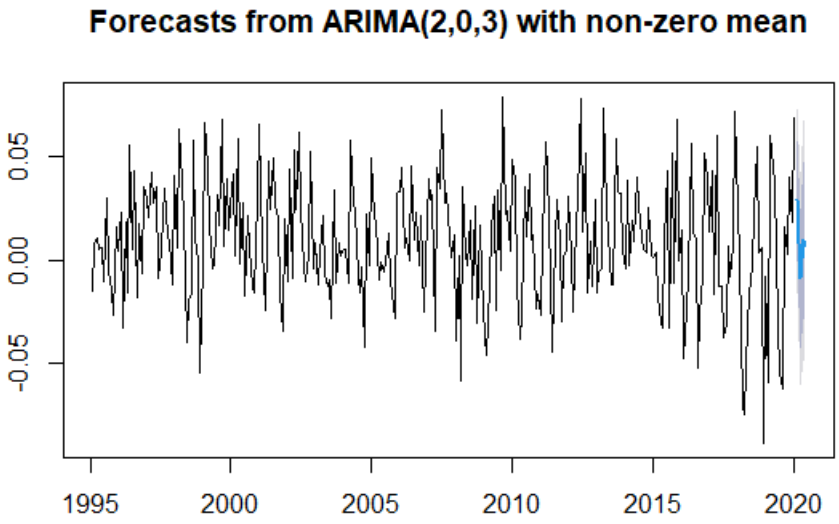

Figura 4: Proyección de valores futuros con bandas de confianza mostrando la serie completa.

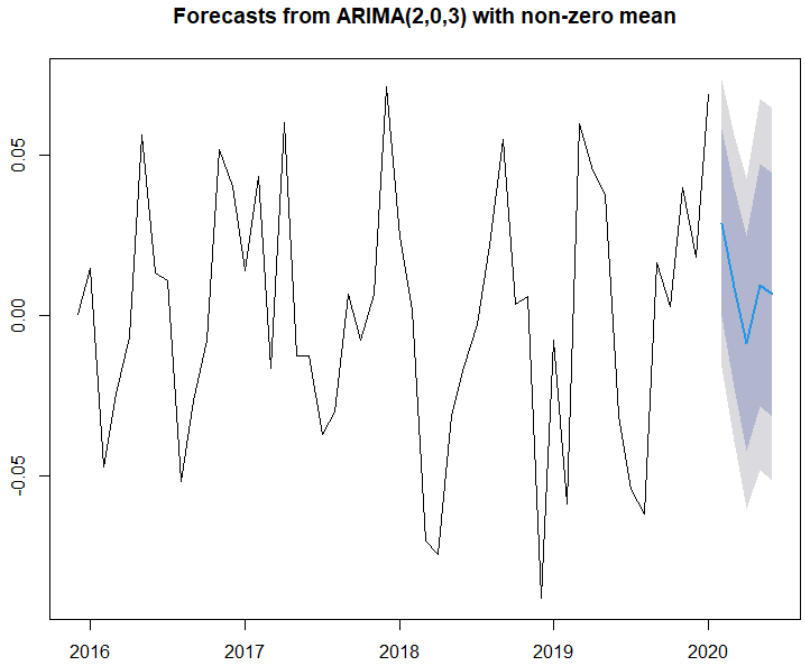

Figura 5: Proyección de valores futuros con bandas de confianza mostrando la serie a partir del 2016.

Serie de tiempo con enfoque bayesiano

Se ajustó el modelo ARIMA(2,0,3), propuesto en el enfoque clásico, usando JAGS y la paquetería `bayesforecast`. Para esto tenemos que definir la ecuación a trabajar, la cual requiere dos variables autoregresivas (las llamaremos ρ_1 y ρ_2) y 3 variables latentes (θ_1, θ_2 y θ_3) para los promedios móviles, una vez que tenemos esto definimos una variable auxiliar z para arrastrar los promedios móviles.

Se obtuvieron los siguientes resultados

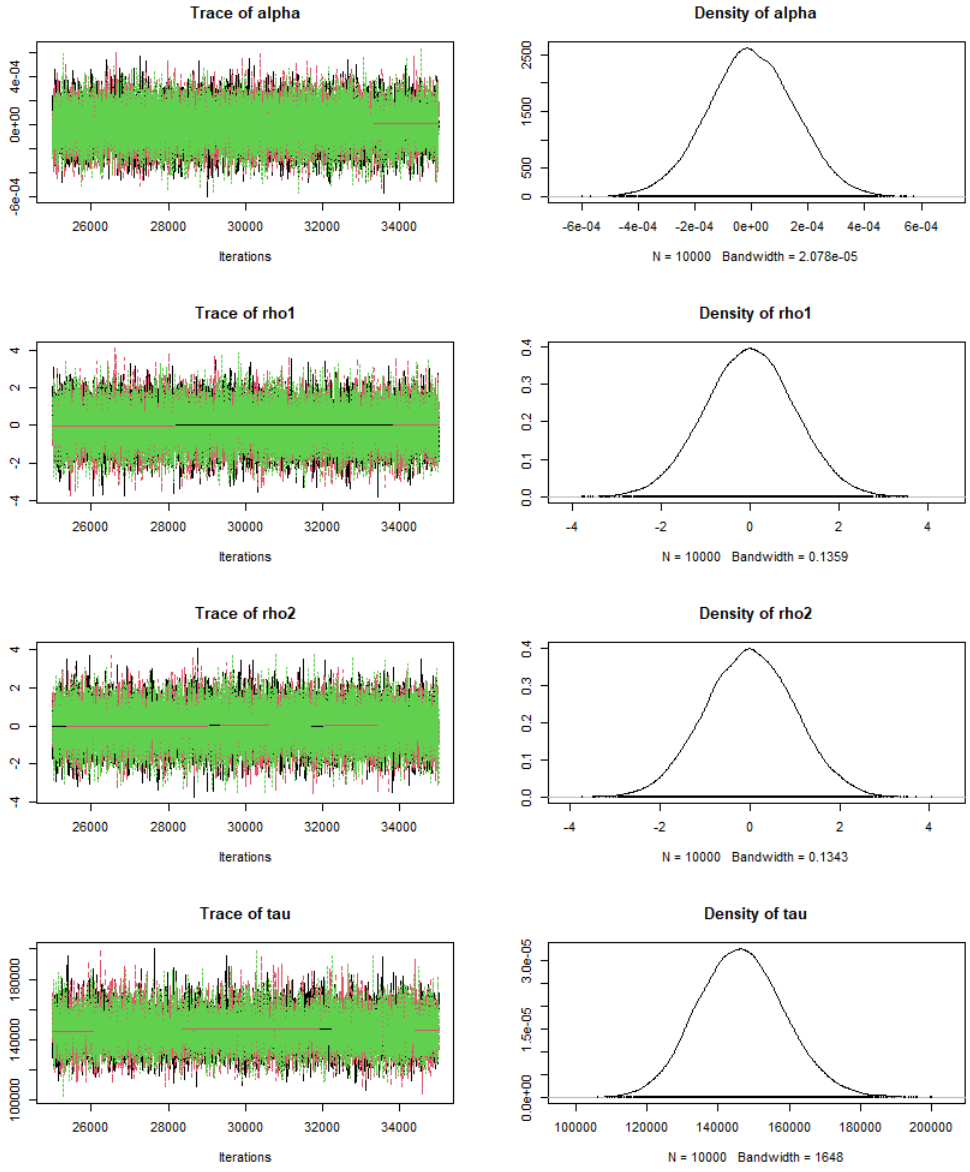

Figura 6: Trazas y densidad de los parámetros estimados utilizando 3 cadenas.

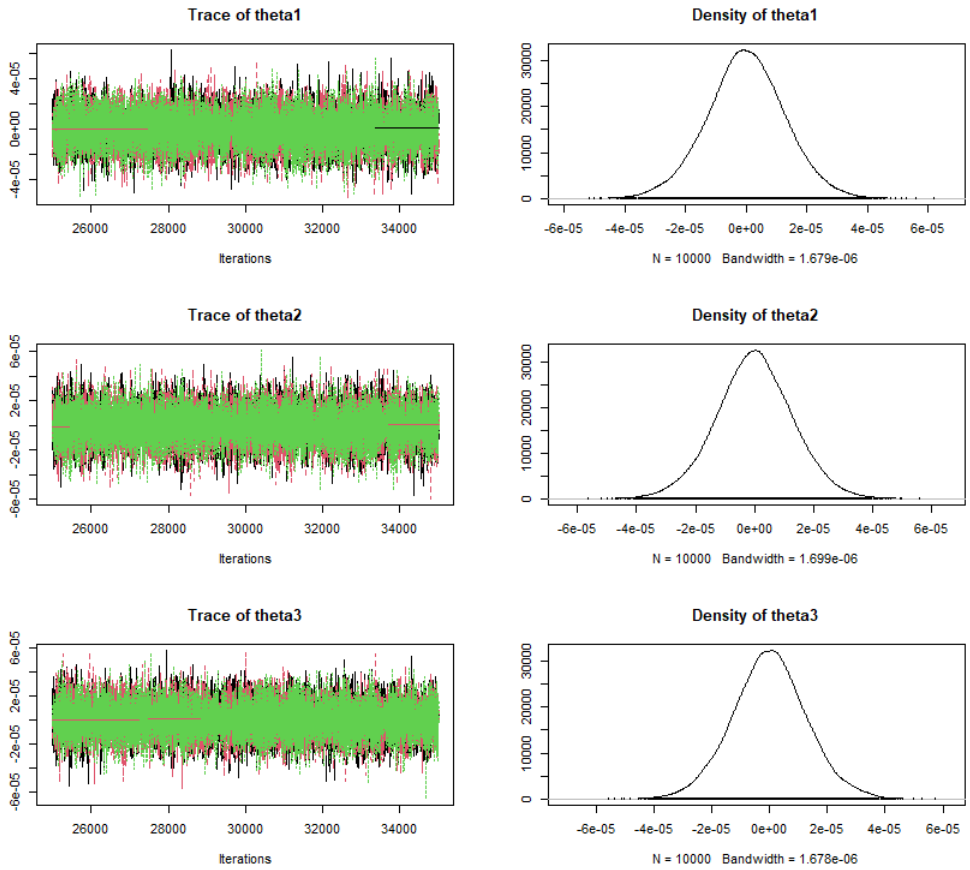

Figura 7: Trazas y densidad de los parámetros estimados utilizando 3 cadenas.

Podemos ver que las trazas convergen y las densidades también. Podemos ver que convergen muy cerca del cero, que es lo que también se pudo ver en el modelo clásico.

A su vez podemos ver el comportamiento de los residuales

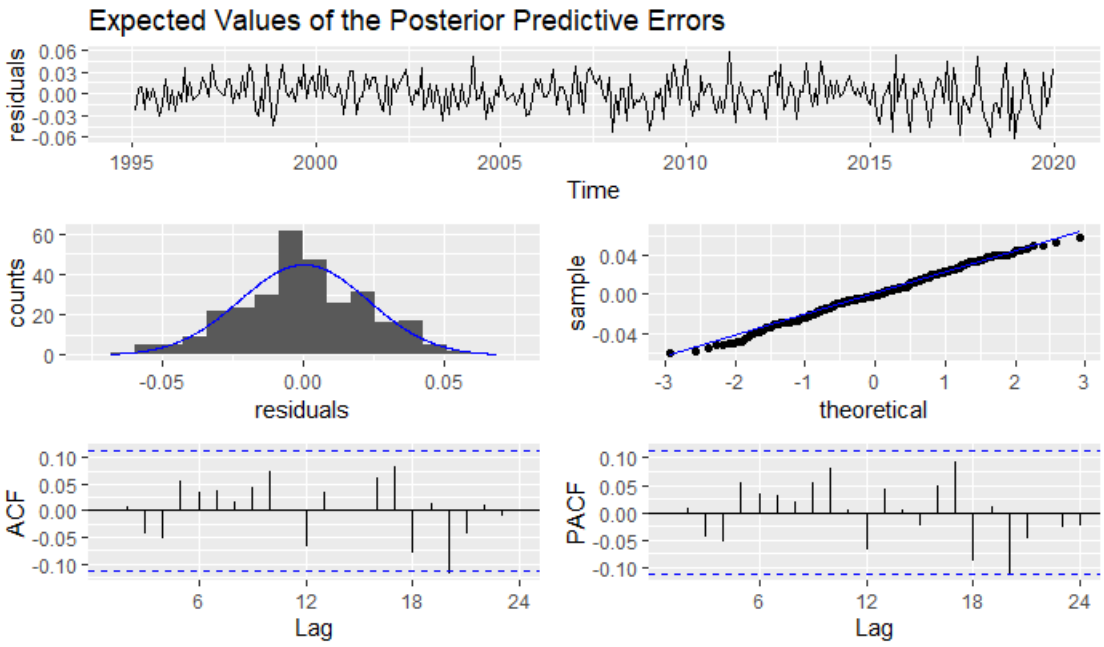

Figura 8: Residuales para el modelo ARIMA(2,0,3) usando el enfoque bayesiano.

Se puede ver que parecería que sí siguen un ruido blanco, las gráficas de ACF y PACF se mantienen dentro de las bandas. Se realizaron las pruebas de los supuestos y sí pasaron la prueba de independencia (Ljung-Box)

Finalmente vemos la predicción del modelo

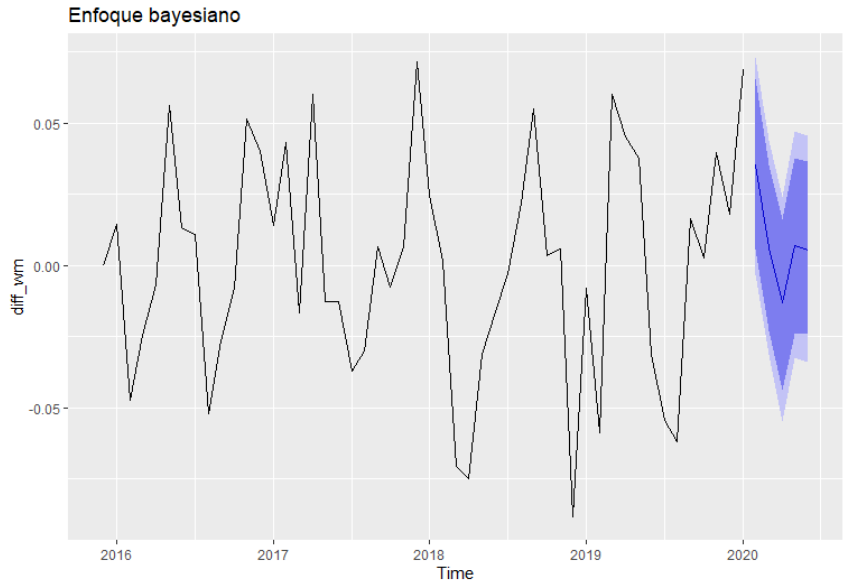

Figura 9: Predicción de 5 valores para el modelo ARIMA(2,0,3) bayesiano.

Es muy parecida a la predicción que se presentó en el modelo clásico y podríamos decir que ajusta bien,

mas si observamos la siguiente gráfica que muestra los valores ajustados contra los observados podemos notar que se está subestimando la varianza ya que los valores ajustados están por debajo de los observados.

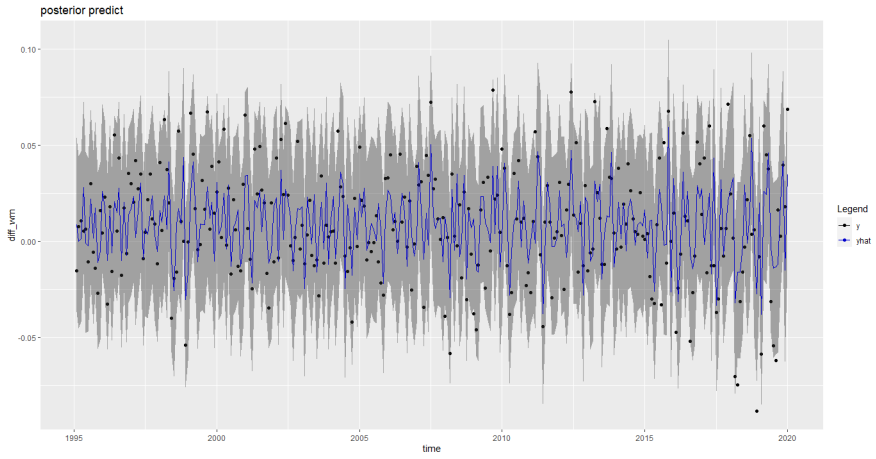

Figura 10: Valores observados contra ajustados. Los azules representan los ajustados mientras que los negros son los observados.

Como desde el principio se tuvieron problemas con que la varianza de la serie no era constante buscamos otro modelo para ajustar con el comando `auto.sarima`. Obtuvimos un ARIMA(1,0,2)

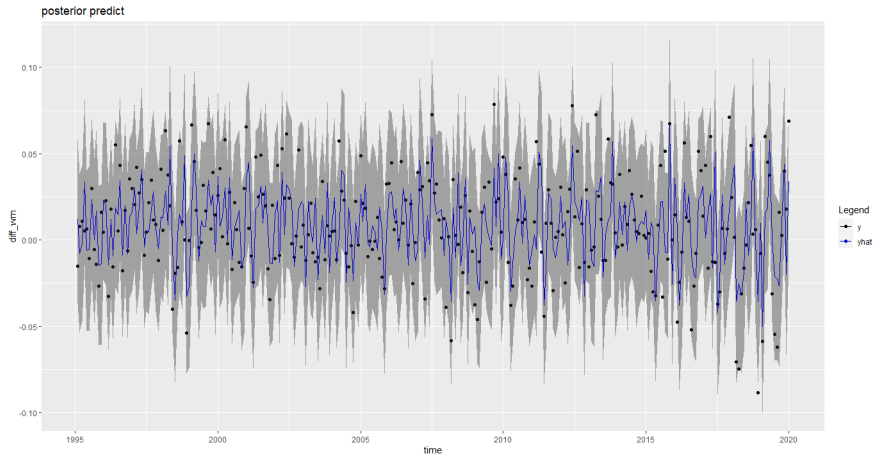

Figura 11: Valores observados contra ajustados del modelo ARIMA(1,0,2)

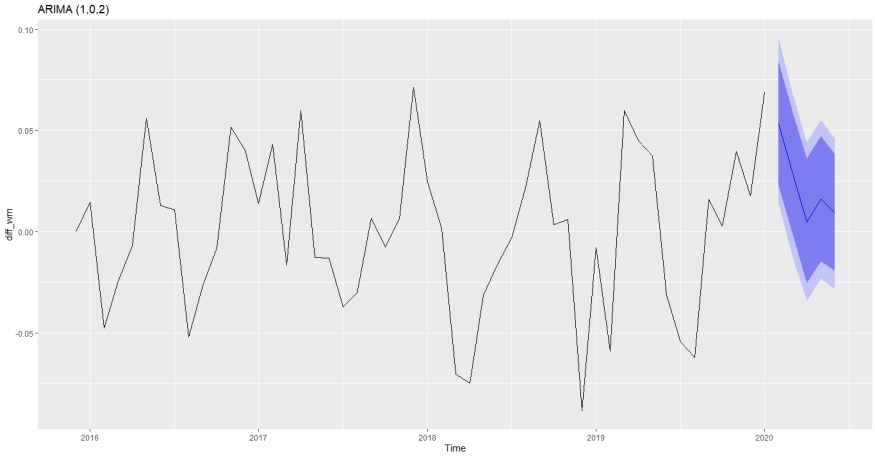

Figura 12: Predicción para el modelo

Tratamos de ajustar distintos modelos con la paquetería `bayesforecast`. Observamos que la mayoría no pasaba la prueba de la varianza constante por lo que decidimos tomar los que tuvieran mayor p-value para esta prueba y que no tuvieran correlación en los parámetros. La lista de modelos que conservamos fueron: ARIMA(2,0,3), ARIMA(3,0,3), ARIMA(4,0,2), ARIMA(1,0,2) y ARIMA(2,0,1).

Comparando estos modelos llegamos a que el que parecía que tenía un mejor ajuste era el ARIMA(1,0,2), que de hecho fue el sugerido por el comando `auto.sarima`.

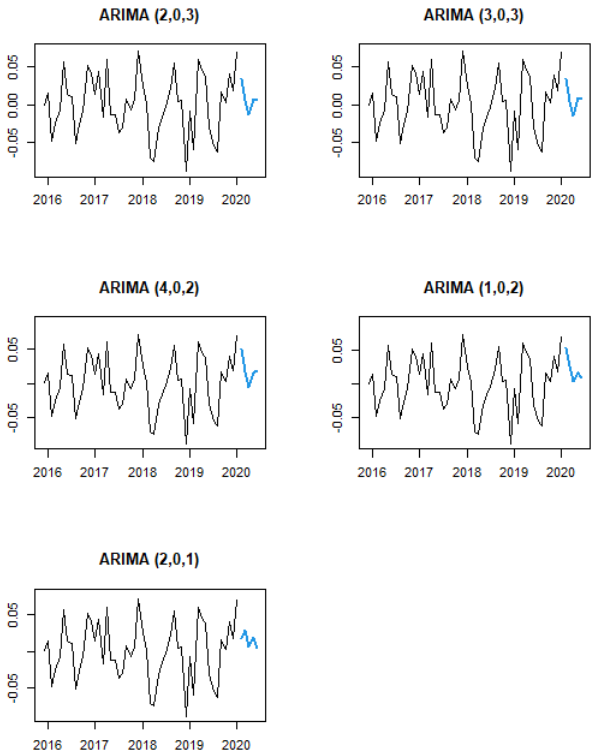

Figura 13: Predicciones para los 5 modelos

Entonces el mejor modelo que se obtuvo fue el ARIMA(1,0,2).

Conclusiones

Finalmente no hubo un único mejor modelo para el ajuste de la serie de tiempo. Se intentó trabajar con el mismo modelo obtenido en el enfoque clásico para la parte bayesiana pero para estos métodos no resultaba ser la mejor opción. Pudimos ver la importancia de tener distintas formas de abordar los modelados y lo difícil que puede ser llegar a un buen ajuste, sobre todo en la forma bayesiana porque requiere más trabajo computacional que tal vez no fue mucho con nuestros datos pero al tener bases con millones de datos se puede complicar el estar probando distintos modelos.

Referencias

- Base de datos obtenida de `https://www.kaggle.com/justinas/housing-in-london`
- Joel Marsden. (2015). House prices in London – an economic analysis of London's housing market. 30/01/2021, de GLA Economics Sitio web: `https://www.london.gov.uk/sites/default/files/house-prices-in-london.pdf`.
- Sanchez Mexicano, S., Valencia Ramirez, G. J. (1989). Modelos para series de tiempo: una perspectiva Bayesiana.

- The Week. (2021). London house prices: average hits record high of £514,000. 30/01/2021, de The Week Sitio web: https://www.theweek.co.uk/london-house-prices
- Natalie Jones. (2021). UK House Price Index: November 2020. 30/01/2021, de Office for National Statistics Sitio web: https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/housepricein
november2020