Metodología para la predicción con modelos SARIMA Bayesianos

Daniel Dala
Departamento de Estadística
Universidad Nacional Autónoma de Honduras
e-mail: daniel.dala@unah.hn

Índice

I. Introducción 2

II. Preliminares y Notación 2
II-A. Modelos SARIMA . . . 3
II-B. Metodología Box-Jenkins . . . 3
II-C. Inferencia Bayesiana . . . 4
II-D. Modelos SARIMA Bayesianos . . . 4

III. Adaptaciones a la Metodología Box-Jenkins 4
III-A. Estimación del Modelo . . . 4
III-B. Evaluación de la inferencia . . . 5
III-C. Diagnóstico del modelo . . . 5
III-D. Comparación de los modelos . . . 5

IV. Ilustraciones 6
IV-A. Temperatura promedio en Honduras . . . 6
IV-B. Precio de cierre en la acciones de Pfizer . . . 8

V. Conclusiones 9

Referencias 10

Índice de figuras

1. Metodología de Box-Jenkins (1970). El diagrama de flujo presenta el procedimiento a utilizar para un análisis de datos adecuado en un enfoque frequentista, para una mayor descripción de dicha metodología revisar [1]. . . . 3
2. Metodología para la predicción con modelos SARIMA Bayesianos. El diagrama de flujo presenta la adaptación del Método Box y Jenkins (Figura 1) adaptado a un enfoque Bayesiano basado en el Bayesian workflow propuesto por *Gelman, Vehtari et. al. (2020)* [2]. . . . 5
3. Registro de la temperatura promedio mensual en Celsius para Honduras desde el año 1980 hasta el 2013. La serie no presenta tendencia pero presenta fuertes variaciones y oscilaciones periódicas. Por lo tanto, un modelo SARIMA estacional es adecuado para el análisis de la serie. . . . 6
4. El gráfico superior muestra la serie con una diferencia estacional periodo 12 y luego diferenciados no estacionalmente. Los gráficos inferiores muestran las funciones ACF y PACF de los datos doblemente diferenciados. 6
5. Gráficos de densidades a posteriori y las cadenas simuladas para cada uno de los parámetros estimados. . . . 7
6. La serie de los residuos (parte superior). El histograma y gráfico de cuantiles (parte media). Gráficos de autocorrelación y autocorrelación parcial de los residuos del modelo (parte inferior). . . . 7
7. La parte superior compara las predicciones de los modelos con el conjunto de prueba. La parte inferior presenta predicción generada por el Modelo 1. . . . 7
8. Precio de cierre promedio en Dolares de enero 2010 hasta septiembre 2021. . . . 8
9. Los gráficos superiores muestran los datos diferenciados y doblemente diferenciados no estacionalmente. Los gráficos inferiores muestran las gráficas de los ACF y PACF de los datos doblemente diferenciados . . . 8
10. Gráficos de las densidades a posteriori y las cadenas simuladas para cada uno de los parámetros estimados. . . . 9
11. La serie de los residuos (parte superior). El histograma y gráfico de cuantiles (parte media). Gráficos de autocorrelación y autocorrelación parcial de los residuos del modelo (parte inferior). . . . 9
12. La parte superior compara las predicciones de los modelos con el conjunto de prueba. La parte inferior presenta predicción generada por el Modelo 1, desde el mes de octubre del 2021 hasta febrero del 2022. . . . 9

ÍNDICE DE CUADROS

I. Resumen de las distribuciones a posterioris de cada uno de los parámetros para el modelo 1. Los estadísticos presentados son la media a posteriori (media), Error estándar (SE), intervalos de credibilidad al 90 %, tamaño de muestra efectivo (ESS) y la reducción de escala potencial ($\hat{R}$.) 6
II. Resumen comparativo de la precisión en cada modelo, en donde el modelo 1 muestra los mejores resultados y el modelo 3 muestra los peores. 7
III. Resumen de las distribuciones a posteriori de cada uno de los parámetros para el modelo 1. Los estadísticos presentados son la media a posteriori (media), Error estándar (SE), intervalos de credibilidad al 90 %, tamaño de muestra efectivo (ESS) y la reducción de escala potencial ($\hat{R}$. 8
IV. Resumen comparativo de la precisión en cada modelo, en donde el modelo 1 muestra los mejores resultados y el modelo 3 muestra los peores. 9

Metodología para la predicción con modelos SARIMA Bayesianos

***Resumen*—El Método Box-Jenkins muestra un conjunto de pasos para realizar predicciones en series temporales con modelos ARIMA mediante un enfoque frecuentista. En este estudio exponemos una adaptación de dicho método a un enfoque Bayesiano. Para la estimación de los parámetros utilizamos inferencia Bayesiana aproximando la distribución a posteriori de los modelos mediante métodos de Markov Chain Monte Carlo. Para el diagnóstico de los modelos se analiza el comportamiento y distribución de los errores que siguen un supuesto de ruido blanco Gaussiano y para la comparación de modelos se miden las diferencias en la precisión de las predicciones usando validación cruzada y comparando las predicciones en un conjunto de prueba. Finalmente, mostramos el desempeño de la metodología propuesta con dos ejemplos, primero en la predicción de la temperatura promedio mensual en Honduras y segundo la predicción del precio de cierre mensual en las acciones de la empresa Pfizer.**

I. Introducción

Una de las aplicaciones más importantes en el análisis de serie temporales es la predicción, esto es, estimar valores futuros que generalmente son desconocidos, para esto existen diferentes metodologías como los modelos de espacio y estado [3], Prophet [4], redes neuronales [5], splines [6], procesos Gaussianos [7], entre otros. Una clase de modelos muy populares por su fácil interpretación y alta capacidad predictiva son los modelos SARIMA [8], [9], pero su implementación con datos reales es compleja debido a que seleccionar el orden del modelo es una tarea complicada. Box y Jenkins (1970) [1] propusieron una metodología para el uso adecuado de dichos modelos, la cual se basa en seis etapas iterativas: *visualización de los datos, selección del modelo, estimación de parámetros, diagnóstico, comparación de modelos y predicción*. Dicha metodología se ilustra en la Figura 1.

Existen muchos esquemas para el proceso de inferencia, y en los últimos años la inferencia Bayesiana se ha vuelto una alternativa muy utilizada para el análisis de datos con muchas aplicaciones en economía, física, química, psicología, entre otras. Su creciente popularidad se debe a su capacidad de incorporar información externa al modelo mediante una distribución a priori, y actualizar las creencias mediante el Teorema de Bayes. Este enfoque de inferencia en la práctica es muy complicado, por lo cual en los últimos años se han aproximado los resultados mediante los métodos de Markov Chain Monte Carlo [10]. Estos métodos consisten en generar una cadena de Markov cuya distribución estacionaria es la distribución a posteriori del modelo, existen muchos procedimientos para implementar estos métodos, uno de los más comunes es el Monte-Carlo Hamiltoneano que por su flexible implementación en el lenguaje Stan ha sido de utilidad en múltiples aplicaciones [11].

El mayor obstáculo al momento de realizar un análisis de datos adecuado en un enfoque Bayesiano, es que los procedimientos de estimación, diagnóstico, y selección utilizados en *Box y Jenkins (1970)* no son válidos en este nuevo enfoque. *Gelman, Vehtari et. al. (2020)* [2] proponen una extensa y robusta metodología denominada "*Bayesian workflow*", que presenta diferentes herramientas para un análisis de datos adecuado. Esta metodología se basa en la propuesta por *Box y Jenkins (1970)*, y se generaliza para cualquier tipo de modelamiento que involucre un enfoque de inferencia probabilístico.

Los dos principales problemas del método de *Gelman, Vehtari et. al. (2020)* al ser aplicados en el análisis de series temporales son su compleja estructura, y que algunas herramientas no son adecuadas para datos con supuestos de dependencia, por lo tanto, en este estudio presentamos una simplificación del *Bayesian Workflow* con ligeras variaciones en algunas de las herramientas para su adecuado uso en series temporales. Finalmente, aplicamos nuestra metodología propuesta mediante dos ejemplos. En el primer ejemplo predecimos la temperatura promedio mensual en Honduras con registros desde el año 1980 hasta el año 2013 obtenido de [12], y un segundo ejemplo analizando el precio de cierre promedio mensual de las acciones en la empresa farmacéutica Pfizer con un conjunto de datos registrados mensualmente desde el año 2010 hasta el año 2021 obtenidos de [13].

II. Preliminares y Notación

Para los objetivos de este estudio un proceso estocástico es una colección arbitraria de variables aleatorias $\{Y_1, Y_2, \ldots\}$, y una serie de tiempo o simplemente serie, es una realización o muestra finita $\{y_1, y_2, \ldots, y_n\}$ del proceso. Una propiedad importante a considerar es la estacionaridad, diremos que un proceso $\{y_i\}_{i\in\mathbb{Z}}$ es *estacionario fuerte* si para cualquier colección finita del proceso su distribución conjunta se mantiene constante en el tiempo. Esto es

$$F_X(y_{t_1}, y_{t_2}, ..., y_{t_n}) = F_X(y_{t_{1+\tau}}, y_{t_{2+\tau}}, ..., y_{t_{n+\tau}}),$$

para $t \in \mathbb{Z}_+$ con $n \in \mathbb{N}$ y cualquier $\tau \in \mathbb{Z}_+$. Una propiedad menos restrictiva es la estacionaridad débil, diremos que el proceso $\{y_i\}_{i\in\mathbb{Z}}$ es *estacionario débil* si el proceso tiene una media y varianza constante a través del tiempo, y la autocorrelación es una función lineal de la diferencia de dos tiempos.

$$\mu(t) = \mu \ \ , \ \ \sigma^2(t) = \sigma^2, \ \ corr(t,k) = \tau|t-k|.$$

Para $t, k \in \mathbb{Z}$ y $\tau > 0$. Una serie $\{y_t\}$ presenta tendencia sobre la media del proceso, si la media puede representarse como una función en el tiempo $y_t = f(t) + \varepsilon_t$, donde $\{\varepsilon_t\}$ es un proceso con media cero y $f : \mathbb{Z} \to \mathbb{R}$ es una función

medible. Para transformar un proceso con tendencia en uno estacionario, aplicamos el operador diferencia

$$\nabla y_t = y_t - y_{t-1}. \tag{1}$$

Para un proceso y con tendencia lineal, el proceso ∇y obtenido en la ecuación 1 es estacionario. La ciclicidad en una serie, implica múltiples oscilaciones periódicas en la media del proceso, un caso particular es la estacionalidad, esta sucede cuando la serie presenta una oscilación periódica constante de periodo m en la media. Para transformar un proceso estacional en uno estacionario, aplicamos el operador diferencia estacional

$$\nabla_m y_t = y_t - y_{t-m}. \tag{2}$$

Donde m es un entero positivo que representa el periodo de la serie, y para una serie y con estacionalidad, el proceso $\nabla_m y$ obtenido en la ecuación 2 es estacionario. Es importante recalcar que las series con tendencia o estacionalidad son no estacionarias, por los efectos de modelado, es necesario trabajar con procesos estacionarios. Un ejemplo de procesos estacionarios son los ruidos blancos, una colección de variables independientes con distribución normal, media cero, y varianza constante positiva.

II-A. Modelos SARIMA

Sea $\{Y_i\}_{i=i}^n$ una serie de tiempo, decimos que la serie sigue un modelo *Autorregresivo Integrado de Medias móviles* $ARIMA(p,d,q)$ si para cualquier tiempo Y_t, se puede escribir de la forma:

$$\nabla^d y_t = \mu_0 + \sum_{i=1}^{p} \phi_i \nabla^d y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{3}$$

donde, μ_0 es la media inicial del proceso, $p \in \mathbb{Z}_+$ y $\{\phi_i\}_{i=1}^p$ son el orden y parámetros de la componente autorregresiva respectivamente, $q \in \mathbb{Z}_+$ y $\{\theta_i\}_{i=1}^q$ son el orden y parámetros de la componente de medias móviles respectivamente, $d \in \mathbb{Z}_+$ representa el número de diferencias no estacionales y $\varepsilon_t \sim N(0,\sigma_0)$ es ruido blanco Gaussiano centrado en cero y con varianza constante positiva.

El modelo propuesto en la ecuación 3 se puede adaptar para analizar series de tiempo con estacionalidad, esto se puede lograr agregando componentes autorregresivas y de medias móviles para modelar la estacionalidad de forma aditiva, y adaptando la diferencia estacional de forma multiplicativa. Sea $\{Y_i\}_{i=i}^n$ una serie de tiempo con estacionalidad y periodo $m \in \mathbb{Z}_+$, decimos que la serie sigue un modelo *ARIMA estacional Multiplicativo* $SARIMA(p,d,q)\times(P,D,Q)_m$ si para cualquier tiempo Y_t,

$$\begin{gathered} Z_t = \mu_0 + \sum_{i=1}^{p} \phi_i Z_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{P} \Phi_i Z_{t-im} + \sum_{i=1}^{Q} \Theta_i \varepsilon_{t-im} + \varepsilon_t, \\ Z_t = \nabla_m^D \nabla^d y_t, \end{gathered} \tag{4}$$

donde $\varepsilon_t \sim N(0,\sigma_0)$ es un ruido blanco Gaussiano con varianza constante positiva, los parámetros $\mu_0, p, \{\phi_i\}_{i=1}^p, q, \{\theta_i\}_{i=1}^q, d$ son los mismos definidos en la ecuación 3, $P \in \mathbb{Z}_+$ y $\{\Phi_i\}_{i=1}^P$ son el orden y parámetros de la componente autorregresiva estacional respectivamente, $Q \in \mathbb{Z}_+$ y $\{\Theta_i\}_{i=1}^Q$ son el orden y parámetros de la componente de medias móviles estacionales respectivamente, y $D \in \mathbb{Z}_+$ representa el número de diferencias estacionales. Note que Z_t es la transformación obtenida al aplicar diferencias y transformaciones estacionales de forma multiplicativa.

II-B. Metodología Box-Jenkins

El procedimiento para predicción de valores futuros en series de tiempo requiere de dos etapas fundamentales: análisis de los datos y selección del modelo de predicción que mejor se ajuste a los datos. A continuación se expondrá brevemente cada uno de los pasos, para un estudio más profundo del método leer [14].

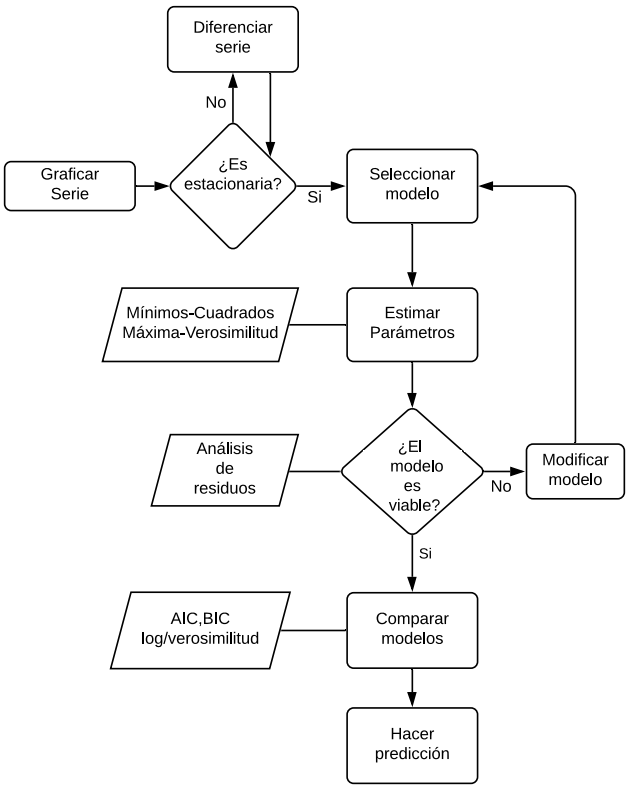

Figura 1. Metodología de Box-Jenkins (1970). El diagrama de flujo presenta el procedimiento a utilizar para un análisis de datos adecuado en un enfoque frequentista, para una mayor descripción de dicha metodología revisar [1].

1. *Visualización de los datos:* la visualización mediante gráficos, permite detectar patrones como tendencia, estacionalidad, ciclos u observaciones atípicas, que deben ser filtrados diferenciando la serie o aplicando diferencias estacionales. Los gráficos ACF (función de autocorrelación) [15] y PACF (función de autocorrelación parcial) [15] son de gran importancia para seleccionar los ordenes del modelo [3].
2. *Selección:* para definir un modelo inicial es necesario establecer los valores p,d,q,P,D,Q y m. Lo valores d,D son el número de diferencias necesarias para que la serie sea estacionaria o un ruido blanco, esto se logra graficando la serie original y la serie diferenciada. Los valores (p,P) y (q,Q) se identifican con los gráficos

PACF y ACF respectivamente, como el número de retardos (lags) diferentes de cero en la serie diferenciada, para más detalles ver [3].

3. *Estimación:* una vez se ha definido un modelo inicial, es necesario estimar los $n_p = p+P+q+Q+2$ parámetros, los métodos más usados son: *mínimos cuadrados* [16], *máxima verosimilitud* [16], y la *ecuación de Yule-Walker* [17], [18].
4. *Diagnóstico:* Los modelos $SARIMA$ siguen el supuesto que los errores siguen un ruido blanco Gaussiano, es decir, los errores son estacionarios con distribución normal. Para diagnóstico de estacionaridad, se utilizan las pruebas de Portmanteau [19] y Lijung-Box [20], o pruebas de raíz unitaria como la prueba Augmented Dickey-Fuller [21], Phillips-Perron [21], y KPSS [22]. Para medir normalidad pruebas como Epps [23], Lobato-Velasco [24] y proyecciones aleatorias [25] que miden normalidad en procesos estacionarios son las más adecuadas. Para más detalles ver [26].
5. *Comparación:* El criterio de selección de modelos más utilizado es el *Akaike's Information Criteria (AIC)* propuesto por Akaike en 1974 [27]. Sea $n_p = p + q + P + Q + 2$ el número de parámetros estimados en el modelo, luego se eligen los valores de p, q, P, Q que minimicen el AIC:
$$AIC = -2logL + 2n_p, \tag{5}$$
donde L denota la verosimilitud. Existen varias modificaciones del AIC que también son usadas como el BIC (Bayesian Information Criteria) y la log-verosimilitud.
6. *Predicción:* Una vez elegido el modelo que mejor se ajuste a cada una de las pruebas anteriores se procede a hacer la predicción de las observaciones futuras.

II-C. Inferencia Bayesiana

En un enfoque de inferencia Bayesiano, se analiza la probabilidad de un parámetro $\theta \in \Theta$ dada la muestra obtenida y, donde Θ es un espacio de probabilidad para el conjunto de parámetros. Para eso, debemos iniciar por establecer un modelo que proviene de una distribución de probabilidad conjunta para θ y y. La función de probabilidad conjunta puede ser escrita como el producto de dos densidades:

$$p(\theta, y) = p(\theta)p(y|\theta)$$

donde a $p(\theta)$ se le llama la distribución a *priori* y a $p(y|\theta)$ la distribución de la muestra o *verosimilitud*. Al condicionar el valor conocido de los datos y y usando el Teorema de Bayes obtenemos la distribución a *posteriori*:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

Una forma equivalente de la ecuación anterior omite el factor $p(y)$ el cual, al no depender de θ se considera una constante, lo que resulta en una distribución posteriori no normalizada, en otras palabras la posteriori es proporcional a la verosimilitud y priori, mediante la siguiente ecuación:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{6}$$

Note que $p(y|\theta) : \Theta \to \mathbb{R}$ es una función de θ cuando la muestra y es fija, y se le conoce como *función de verosimilitud.* Estas ecuaciones bastan para realizar inferencia Bayesiana en donde primero se establece un modelo $p(\theta, y)$ y luego se desarrollan los cálculos computacionales para obtener $p(\theta|y)$ el cual funciona como la información actualizada de los datos.

Por otro lado, al inferir observaciones desconocidas o predicciones, seguimos un procedimiento similar. Después de inferir nuestros parámetros θ a partir de los datos observados y, podemos predecir una observación desconocida $\tilde{y}$. La distribución de $\tilde{y}$ se conoce como la *distribución predictiva a posteriori* y se obtiene con la siguiente ecuación:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

II-D. Modelos SARIMA Bayesianos

En base a la definiciones previas, se define un *modelo SARIMA Bayesiano* como un modelo ARIMA estacional y una selección de prioris independientes entre si para cada uno de los parámetros desconocidos, siguiendo las siguientes ecuaciones:

$$\begin{gathered}
y \sim SARIMA(p, d, q)_{\times}(P, D, Q)_m \\
\phi_i \sim p(\phi_i), \quad i = 1, ..., p \\
\theta_j \sim p(\theta_j), \quad j = 1, ..., q \\
\Phi_k \sim p(\Phi_k), \quad k = 1, ..., P \\
\Theta_w \sim p(\Theta w), \quad w = 1, ..., Q \\
\mu_0 \sim p(\mu_0) \\
\sigma_0 \sim p(\sigma_0)
\end{gathered}$$

Donde los datos y son una realización de un proceso estocástico que siguen la ecuación (4), y $\theta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \Phi_1, \ldots, \Phi_P, \Theta_1, \ldots, \Theta_Q, \mu_0, \sigma_0) \in \mathbb{R}^{n_p}$ es el vector de parámetros desconocidos.

III. Adaptaciones a la Metodología Box-Jenkins

Al proponer un modelo SARIMA Bayesiano para hacer la predicción debemos hacer modificaciones y adaptaciones al Método Box-Jenkins, esto dado que las estimaciones y diagnósticos de parámetros se basan en inferencia Bayesiana. Inicialmente los pasos de *visualización de datos* y *selección del modelo* se mantienen igual, por lo que en las siguientes secciones se expondrán los pasos de *selección, diagnóstico y comparación de los modelos* que se verán modificados siguiendo la linea del *Bayesian Workflow* adaptado al análisis de series temporales.

III-A. Estimación del Modelo

Una vez establecido el modelo inicial. y las distribuciones a priori de cada uno de los parámetros, se aproxima la distribución a posteriori mediante Monte Carlo Hamiltoneano el cual simula una cadena de markov estacionaria que converge a la distribución de cada uno de los parámetros. Para una mejor compresión de este proceso de inferencia se recomienda leer [28].

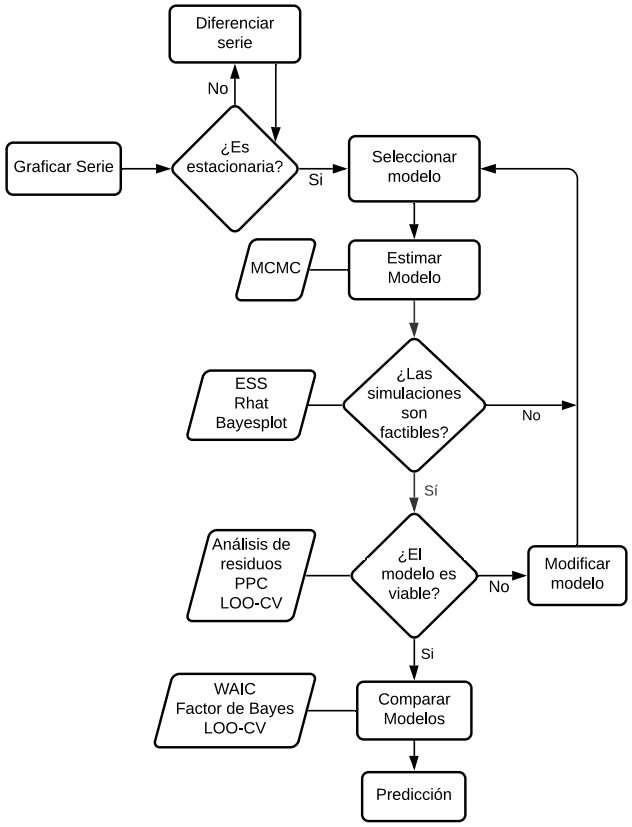

Figura 2. Metodología para la predicción con modelos SARIMA Bayesianos. El diagrama de flujo presenta la adaptación del Método Box y Jenkins (Figura 1) adaptado a un enfoque Bayesiano basado en el Bayesian workflow propuesto por *Gelman, Vehtari et. al. (2020)* [2].

III-B. Evaluación de la inferencia

Luego de hacer la inferencia es necesario evaluar la convergencia de las simulaciones midiendo la estacionaridad y combinación de las cadenas simuladas para cada uno de los parámetros. De esta manera, para decidir la viabilidad de los resultados se utilizarán dos estadísticos, el Tamaño de muestra efectiva *(Effective Sample Size ESS)* y la Reducción de escala potencial (*potencial scale reduction* $\hat{R}$), el primero indica el tamaño suficiente de las simulaciones para aproximar correctamente los parámetros y el segundo es un indicador de convergencia de las cadenas que para este estudio se tomará como estimador factible si $\hat{R} < 1.1$. Las propiedades generales de ambos estimadores se encuentran en [28].

Es recomendable además, el análisis gráfico del ajuste a cada parámetro, esto es, observar los histogramas y gráficos de las cadenas generadas en busca de indicios de multimodalidad en la distribución de los parámetros y verificar que las cadenas muestren convergencia. Este análisis gráfico se puede realizar con el paquete *bayesplot* [29] y se pueden observar ejemplos en el artículo [30].

III-C. Diagnóstico del modelo

Al igual que en la Metodología Box-Jenkins es recomendable hacer un diagnóstico de los modelos, comprobando que los errores se comporten como un ruido blanco Gaussiano. Por otro lado en la inferencia Bayesiana existen métodos para constatar que el modelo representa efectivamente los datos observados, el principal método es la *Verificación predictiva a posteriori (PPC) (Box, 1980, Rubin, 1984, Gelman, Meng, y Stern, 1996)*. Si el modelo se ajusta bien, este debería generar datos con el mismo comportamiento de las observaciones. No obstante PPC no es factible en el análisis de series temporales debido a que los supuestos de intercambiabilidad no se cumplen en los datos al ser realizaciones de un proceso estocástico. Para evadir estos problemas se recomienda hacer PPC en los residuos del modelo ($\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$) los cuales son estacionarios e intercambiables, por lo que muestran condiciones óptimas para la creación de histogramas y obtención de resultados concluyentes.

Generalmente la PPC es suficiente para encontrar errores en el ajuste del modelo, sin embargo, dado que usamos las observaciones para ajustar el modelo y hacer las evaluaciones es posible que en algunos casos se dejen pasar comportamientos anormales en los datos. Un camino alternativo es hacer el diagnóstico con validación cruzada *Leave-one-out cross-validation (LOO-CV)* en donde una parte de los datos es utilizada para ajustar el modelo y el resto se utiliza para medir la precisión de predicción. En [2] se aconsejan tres maneras de abordar la evaluación usando validación cruzada: 1.Verificaciones de calibración utilizando la distribución predictiva de validación cruzada 2. Identificar qué observaciones o grupos de observaciones son más difíciles de predecir 3. Identificar qué tan influyentes son las observaciones particulares, esto es, cuánta información proporcionan además de otras observaciones. Para una mayor comprensión de LOO-CV se recomienda leer [31].

III-D. Comparación de los modelos

En la comparación de modelos Bayesianos frecuentemente se utilizan LOO-CV y el Criterio de Información Watanabe-Akaike, WAIC (Watanabe, 2010). Ambos métodos estiman la precisión puntual en la predicción en un modelo usando una muestra de la log-verosimilitud. No obstante, el método de estimación de precisión para modelos en series de tiempo que mejores estimaciones propone es la validación cruzada de series de tiempo [32], pero debido a la dificultad de implementación en este estudio se trabajará con LOO-CV y WAIC solamente. La implementación de estos métodos se puede realizar con el paquete *loo* [33]. Para una mayor comprensión de los mismos se recomienda leer [31], [34].

Finalmente, luego de elegir los modelos que presenten mejores resultados en las pruebas anteriores se proceden a hacer la predicciones con los modelos seleccionados. Las herramientas propuestas ofrecen la precisión suficiente para una buena estimación y selección de modelos, sin embargo, hace falta un análisis cuidadoso a la hora de hacer predicciones de valores futuros. En las siguientes secciones se mostrarán ejemplos con datos reales, aplicando nuestra metodología propuesta para seleccionar el modelo SARIMA que mejor se ajuste a los datos en un enfoque Bayesiano.

IV. Ilustraciones

Aplicaremos la nueva metodología estudiando dos conjuntos de datos y realizando sus respectivas predicciones. Para la inferencia y análisis de los modelos utilizaremos el paquete *bayesforecast* que implementa dichos modelos usando un Monte-Carlo Hamiltoneano, generando 4 cadenas de 2,000 iteraciones y un warm-up de 1,000 iteraciones. Como diagnóstico de convergencia usaremos el estadístico $\hat{R}$ [28] y la comparación de los modelos será con validación cruzada usando el paquete *loo* [31].

IV-A. Temperatura promedio en Honduras

El primer conjunto de datos muestra la temperatura promedio mensual en Honduras desde el año 1980 hasta el 2013, con un total de 405 observaciones en donde las primeras 385 servirán como conjunto de entrenamiento del modelo y el resto será el conjunto de prueba. Estos datos fueron obtenidos de *Berkeley Earth* [12]. La Figura 3 muestra que los datos

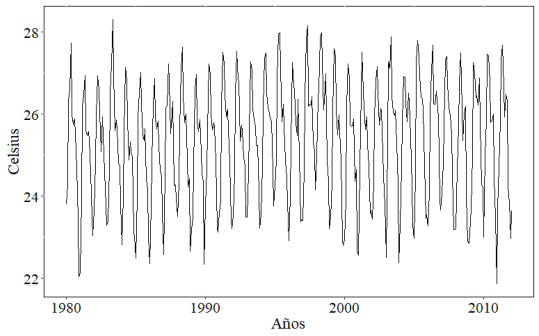

Figura 3. Registro de la temperatura promedio mensual en Celsius para Honduras desde el año 1980 hasta el 2013. La serie no presenta tendencia pero presenta fuertes variaciones y oscilaciones periódicas. Por lo tanto, un modelo SARIMA estacional es adecuado para el análisis de la serie.

no siguen un comportamiento estacionario, y dado que son observaciones climatológicas se espera que existan patrones estacionales anuales, por lo tanto se aplicará una diferencia estacional con periodo 12 y una diferencia no estacional.

La Figura 4 muestra que al aplicar una diferencia estacional la serie de la parte superior no luce estacionaria dado que la media no es constante, sin embargo la estacionalidad se reduce, lo que indica que basta hacer una segunda diferencia no estacional. El gráfico intermedio muestra la serie doblemente diferenciada y esta parece tener una media constante en cero y varianza estable. Al observar las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la parte inferior de la Figura 4, estas muestran una leve correlación de a lo más dos retardos en los datos, el orden del modelo SARIMA inicial a considerar entonces es $p = 2$, $d = 1$, $q = 2$. Por otra parte, ambos gráficos muestran ciertos patrones periódicos y los retardos indican una posible componente autorregresiva estacional, por lo tanto, el orden de la componente estacional es $P = 2$, $D = 1$, $Q = 0$. Finalmente, consideramos las distribuciones a priori de cada uno de los parámetros, en este caso seleccionamos prioris poco informativas. De esta manera el modelo completo es:

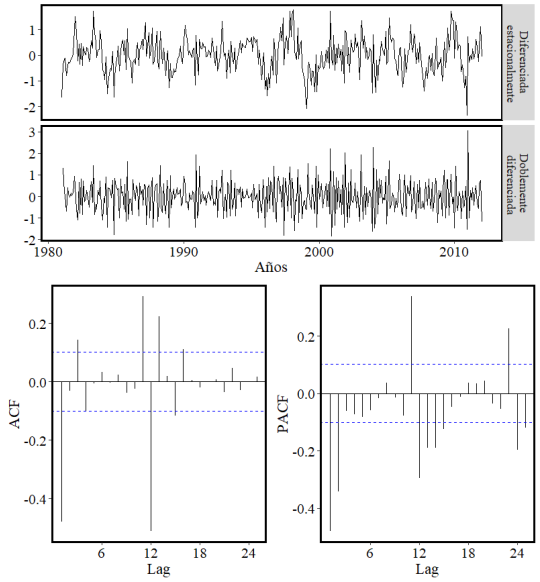

Figura 4. El gráfico superior muestra la serie con una diferencia estacional periodo 12 y luego diferenciados no estacionalmente. Los gráficos inferiores muestran las funciones ACF y PACF de los datos doblemente diferenciados.

$$\text{Modelo } 1 \sim SARIMA(2,1,2)_{\times}(2,1,0)_{12}$$
$$\mu_0 \sim t(0, 2.5, 6)$$
$$\sigma_0 \sim t(7)$$
$$ar_i, ma_i \sim N(0, 0.5) \quad i = 1, 2$$
$$sar_i \sim N(0, 0.5) \quad i = 1, 2$$

En la Cuadro I se muestra un resumen de las distribuciones a posteriori de cada uno de los parámetros. El estadístico $\hat{R}$ de cada uno de ellos indica que las cadenas convergen, y los tamaños de muestra efectivo (ESS) son valores mayores al número total de iteraciones, indicando un tamaño de muestra factible para la representación efectiva de los parámetros, por lo tanto, aceptamos la aproximación de las posterioris obtenidas.

	media	SE	5 %	95 %	ESS	$\hat{R}$
μ_0	0.01	0.00	-0.01	0.03	3904.71	0.9998
σ_0	0.51	0.00	0.48	0.54	3489.52	0.9999
ar.1	-0.00	0.00	-0.13	0.13	4122.60	0.9998
ar.2	-0.02	0.00	-0.10	0.06	4087.97	1.0001
ma.1	-0.65	0.00	-0.81	-0.48	3967.56	0.9998
ma.2	0.03	0.00	-0.09	0.16	3888.24	0.9999
sar.1	-0.68	0.00	-0.77	-0.60	4093.16	0.9999
sar.2	-0.36	0.00	-0.44	-0.27	3934.67	0.9998
loglik	-277.24	0.03	-281.00	-274.51	4084.98	0.9998

Cuadro I
Resumen de las distribuciones a posterioris de cada uno de los parámetros para el modelo 1. Los estadísticos presentados son la media a posteriori (media), Error estándar (SE), intervalos de credibilidad al 90 %, tamaño de muestra efectivo (ESS) y la reducción de escala potencial ($\hat{R}$.)

En la Figura 5 se observa que las cadenas parecen ser estacionarias indicando convergencia, además no se observa multimodalidad en la distribución a posteriori de los parámetros

por lo tanto podemos aceptar las estimaciones del modelo y continuar con el diagnóstico del ajuste en los datos.

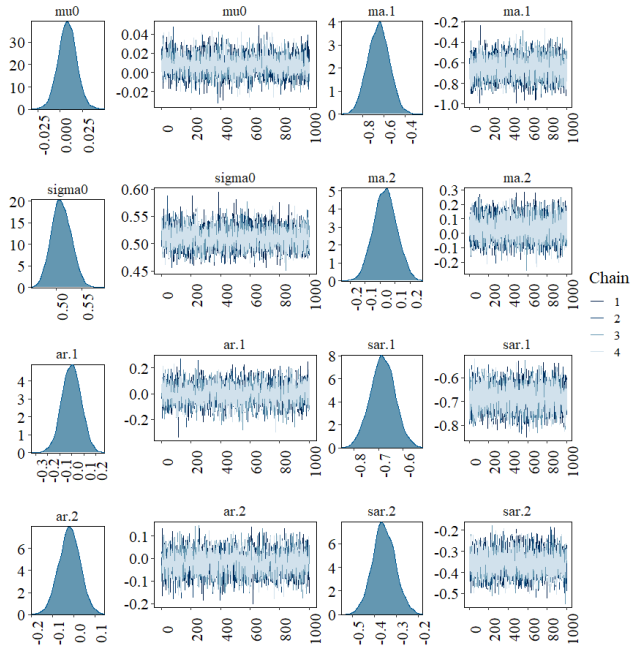

Figura 5. Gráficos de densidades a posteriori y las cadenas simuladas para cada uno de los parámetros estimados.

La Figura 6 muestra un resumen de los residuos del modelo 1. El gráfico superior muestra la serie de los residuos, que presenta una leve ciclicidad despreciable y no presenta volatilidad. Los gráficos ACF y PACF en la parte inferior de la Figura 6 muestran una baja correlación manteniéndose en los intervalos de confianza por lo tanto, los residuos parecen estacionarios. El histograma y el gráfico de cuantiles (intermedio) muestra que la media a posteriori de los residuos tienen distribución simétrica y sin colas pesadas indicando normalidad. Por lo tanto, concluimos que los residuos siguen un ruido blanco Gaussiano, satisfaciendo los supuestos para el Modelo 1.

	$elpd_{diff}$	SE_{diff}
modelo 1	0.00	0.00
modelo 2	-2.99	8.89
modelo 3	-29.68	12.28

Cuadro II
RESUMEN COMPARATIVO DE LA PRECISIÓN EN CADA MODELO, EN DONDE EL MODELO 1 MUESTRA LOS MEJORES RESULTADOS Y EL MODELO 3 MUESTRA LOS PEORES.

Comparamos el modelo obtenido con dos modelos alternativos de ordenes diferentes a los propuestos para el modelo inicial, que se definen en las siguientes dos ecuaciones:

$$\text{Modelo 2} \sim SARIMA(1,1,0) \times (1,1,1)_{12}$$

$$\text{Modelo 3} \sim SARIMA(1,0,0) \times (1,1,0)_{12}$$

El Modelo 2 tiene distribuciones a priori ligeramente diferente a los demás modelos, para los parámetros auto-regresivos y de medias móviles (ar,ma, sma) se seleccionó una distribución $Beta(2,2)$ transformada de tal forma que los valores obtenidos estén en el rango $[-1,1]$, garantizando la estacionaridad

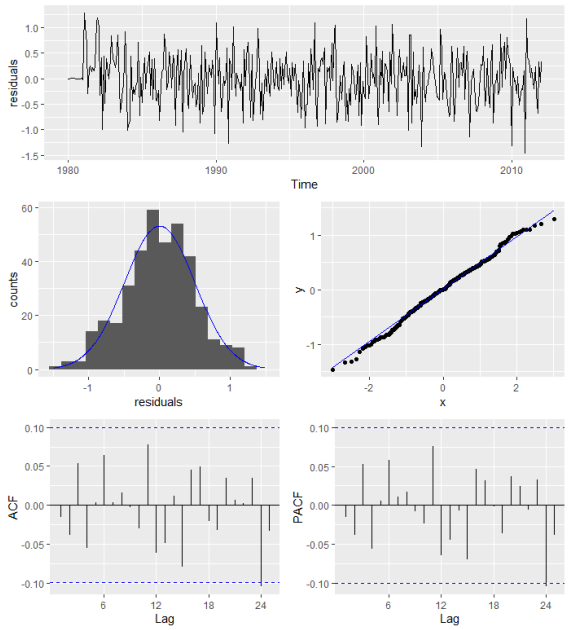

Figura 6. La serie de los residuos (parte superior). El histograma y gráfico de cuantiles (parte media). Gráficos de autocorrelación y autocorrelación parcial de los residuos del modelo (parte inferior).

del proceso. El resto de los parámetros se mantendrán con las mismas distribuciones que el Modelo 1. Es importante destacar que se utilizó el mismo procedimiento para estimar y diagnosticar los Modelos 2 y 3, validando cada una de las etapas del proceso.

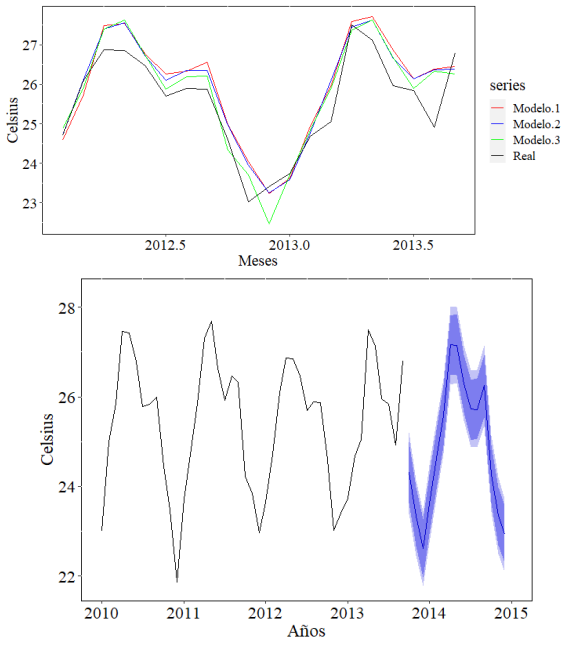

Figura 7. La parte superior compara las predicciones de los modelos con el conjunto de prueba. La parte inferior presenta predicción generada por el Modelo 1.

Procedemos a comparar los modelos usando validación cruzada, que estima la medida $elpd$ (Expected log predictive den-

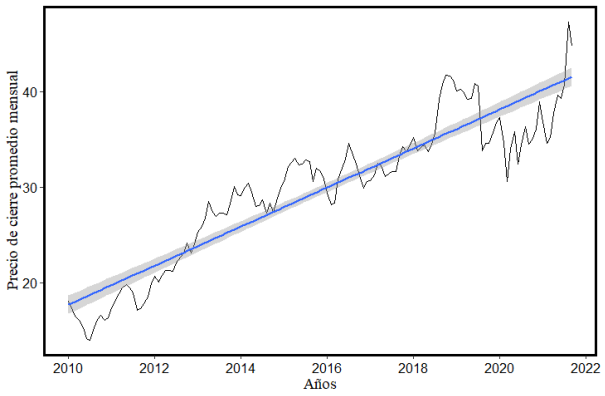

Figura 8. Precio de cierre promedio en Dolares de enero 2010 hasta septiembre 2021.

sity) para establecer la capacidad predictiva de cada modelo. El Cuadro II muestra la diferencia entre $elpds$, donde el Modelo 1 presenta una mayor capacidad predictiva que los otros dos. Adicionalmente, la parte superior de la Figura 7 presenta las predicciones realizadas por cada modelo comparando los resultados con los datos del conjunto de prueba, se observa que cada uno de los modelos generan predicciones muy confiables y similares entre si. Finalmente, seleccionamos el Modelo 1 para predecir los siguientes 15 meses, las predicciones se realizan a partir del mes de Septiembre del año 2013 y los resultados se visualizan en la parte inferior de la Figura 7.

IV-B. Precio de cierre en la acciones de Pfizer

Para el segundo ejemplo estudiamos el precio de cierre de las acciones en la empresa farmacéutica Pfizer, en este caso el conjunto de datos es obtenido de [13], la data muestra los atributos principales de las acciones desde junio de 1972 hasta septiembre de 2021 con 4 mediciones mensuales. Por conveniencia analizaremos el promedio mensual del precio de cierre desde enero del 2010. La serie final con que trabajaremos consta de 141 observaciones, en donde las primeras 134 será el conjunto de entrenamiento y el resto el conjunto de prueba. La Figura 8 muestra que los datos poseen tendencia creciente que se modela con un modelo SARIMA con tendencia determinista [35]. Por otra parte, en la Figura 9 observamos que los datos se estabilizan desde la primera diferencia, además es importante notar la alta volatilidad que se presenta en el periodo entre 2019 y 2021. En base a estos hechos y tomando como referencia los gráficos ACF y PACF que indican un posible modelo autorregresivo definiremos el modelo a utilizar:

$$\text{Modelo 1: } y_t = \beta_1 t + \eta_t$$

$$\eta_t \sim SARIMA(1,1,0)_{\times}(1,0,0)_{12}$$

$$\mu_0 \sim t(0, 2.5, 6) \ , \ \beta_1 \sim t(0, 2.5, 6)$$

$$\sigma_0 \sim t(7)$$

$$ar, sar \sim N(0, 0.5)$$

El modelo y_t es una modelo de regresión con sus error η_t, de esta forma inferimos en cada uno de los parámetros del modelo

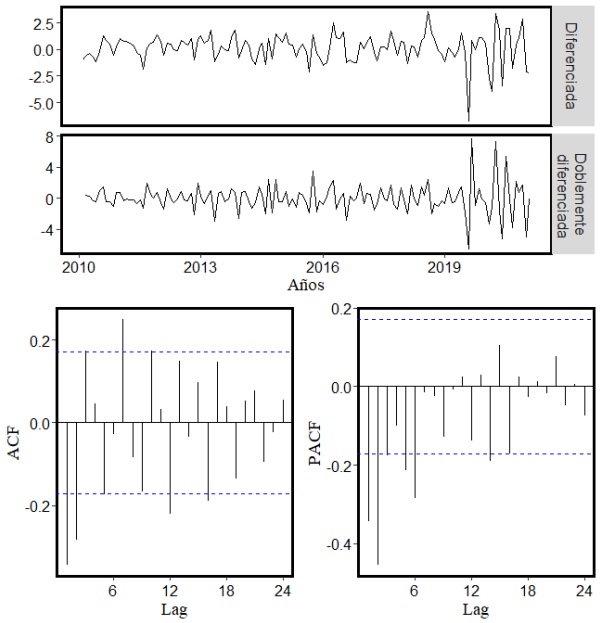

Figura 9. Los gráficos superiores muestran los datos diferenciados y doblemente diferenciados no estacionalmente. Los gráficos inferiores muestran las gráficas de los ACF y PACF de los datos doblemente diferenciados

SARIMA y el parámetro de regresión. En el Cuadro III se muestra un resumen de las distribuciones a posteriori de los parámetros. El estadístico $\hat{R}$ indica convergencia para cada uno de los parámetros, lo cual se confirma visualmente en la Figura 10, los gráficos de las distribuciones a posteriori no muestran multi-modalidad y las cadenas se parecen estacionarias, las cuales en este contexto son condiciones suficientes para suponer que se obtuvo una inferencia factible. Luego, al

	mean	SE	5 %	95 %	ESS	$\hat{R}$
μ_0	0.12	0.03	-2.92	3.17	3548.40	1.0012
σ_0	1.32	0.00	1.20	1.47	3884.63	1.0025
ar	0.06	0.00	-0.08	0.20	3738.73	1.0004
sar	-0.25	0.00	-0.41	-0.08	4392.88	1.0003
β_1	0.02	0.03	-3.03	3.07	3567.41	1.0012
loglik	-225.95	0.02	-228.70	-224.33	3227.89	1.0016

Cuadro III
RESUMEN DE LAS DISTRIBUCIONES A POSTERIORI DE CADA UNO DE LOS PARÁMETROS PARA EL MODELO 1. LOS ESTADÍSTICOS PRESENTADOS SON LA MEDIA A POSTERIORI (MEDIA), ERROR ESTÁNDAR (SE), INTERVALOS DE CREDIBILIDAD AL 90 %, TAMAÑO DE MUESTRA EFECTIVO (ESS) Y LA REDUCCIÓN DE ESCALA POTENCIAL ($\hat{R}$.

observar la Figura 11 se muestra en la serie de los residuos que el modelo no explica correctamente el periodo entre 2019 y 2021 esto debido a la alta volatilidad en esos años, lo cual también se representa en la gráfica de densidad y cuantiles que muestran presencia de colas pesadas, sin embargo, dada la baja autocorrelación mostradas en los gráficos ACF, PACF y el resto de la serie de los residuos este es un modelo factible para el ajuste de los datos. Por otra parte, los gráficos de autocorrelación muestran que un aumento en el orden del modelo autorregresivo podría mejorar los resultados. Ahora, en la comparación de modelos se proponen otras dos regresiones con errores ARIMA los cuales son:

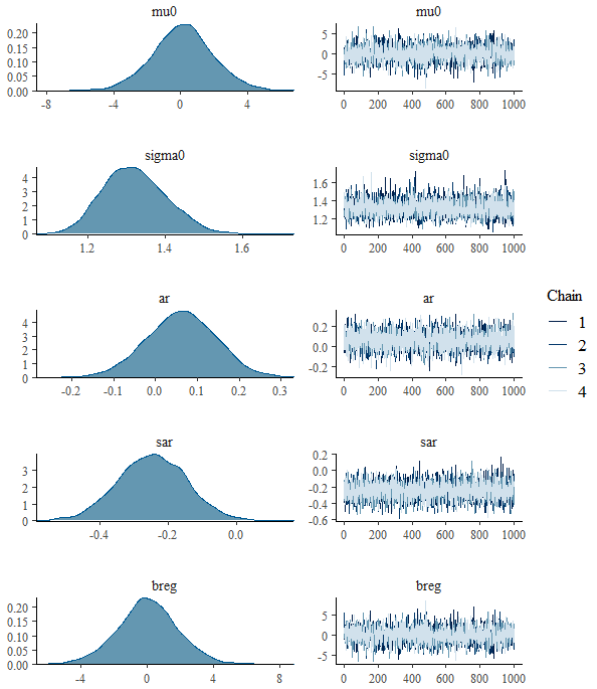

Figura 10. Gráficos de las densidades a posteriori y las cadenas simuladas para cada uno de los parámetros estimados.

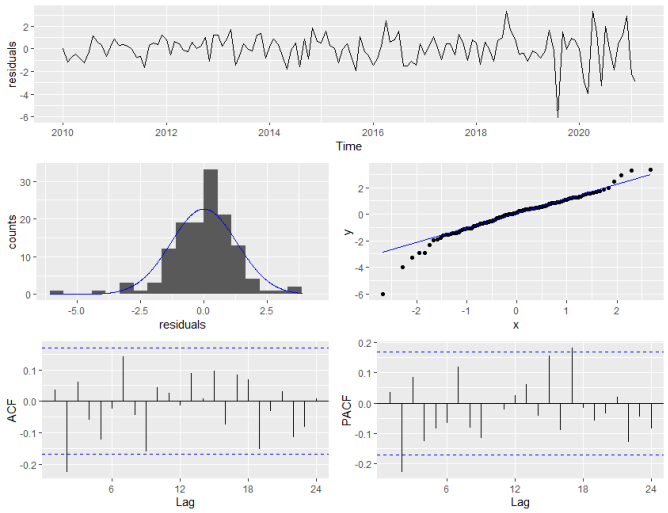

Figura 11. La serie de los residuos (parte superior). El histograma y gráfico de cuantiles (parte media). Gráficos de autocorrelación y autocorrelación parcial de los residuos del modelo (parte inferior).

$$\text{Modelo 2: } y_t = \beta_1 t + \eta_t$$

$$\eta_t \sim ARIMA(1,1,0)$$

$$\text{Modelo 3: } y_t = \beta_1 t + \eta_t$$

$$\eta_t \sim ARIMA(1,0,0)$$

Para el modelo 2, se prueba analizar los datos sin una componente estacional y para el modelo 3 se propone un modelo $AR(1,0,0)$ sin hacer una diferencia lo que indica que los errores se basan únicamente en el ajuste autorregresivo, por lo que la comparación se basa específicamente en el orden de los errores ya que las distribuciones a priori de los tres modelos son las mismas. La Figura 12 muestra que el Modelo 3 presenta los peores resultados, sin embargo, los Modelos 1 y 2 muestran predicciones similares.

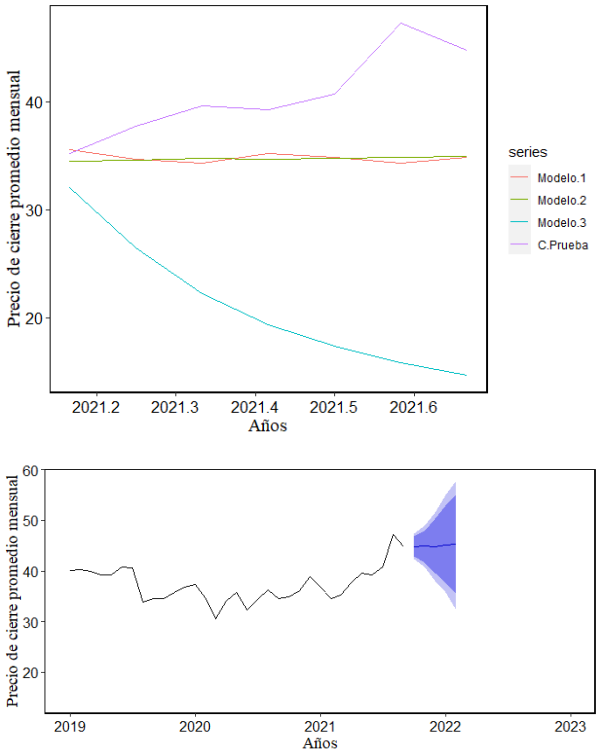

Figura 12. La parte superior compara las predicciones de los modelos con el conjunto de prueba. La parte inferior presenta predicción generada por el Modelo 1, desde el mes de octubre del 2021 hasta febrero del 2022.

Por otro lado, al aplicar CV-LOO y calcular las diferencias de precisión en los modelos se obtiene que el Modelo 1 presenta una mayor precisión que el Modelo 2, por lo tanto se procede a realizar la predicción final mediante el primer modelo. Al observar la Figura 12 se muestra una predicción con comportamiento lineal, sin embargo en el contexto de las observaciones y la tendencia creciente es posible que los valores reales del futuro sean mayores a la predicción obtenida, no obstante, es una buena primera predicción que se puede optimizar mediante la inclusión de variables exógenas.

	$elpd_{diff}$	SE_{diff}
modelo 1	0.00	0.00
modelo 2	-1.74	3.21
modelo 3	-47.19	35.44

Cuadro IV
RESUMEN COMPARATIVO DE LA PRECISIÓN EN CADA MODELO, EN DONDE EL MODELO 1 MUESTRA LOS MEJORES RESULTADOS Y EL MODELO 3 MUESTRA LOS PEORES.

V. CONCLUSIONES

En este estudio introducimos una nueva metodología para el análisis y predicción de series temporales con modelos

SARIMA en un enfoque Bayesiano. La metodología propuesta permite realizar un proceso adecuado de inferencia, diagnostico y selección de modelos, para el analisis de series temporales con modelos SARIMA. Hemos presentado el desempeño y aplicabilidad de la metodología mediante dos ejemplos en donde ambos presentaron resultados satisfactorios.

Disponibilidad de Datos

El conjunto de herramientas y procedimientos explícitos realizados en las Ilustraciones se muestra en [36] además de los conjuntos de datos utilizados.

Referencias

[1] R. J. Hyndman, *Box-Jenkins modelling*. Hans Daellenbach and Robert Flood, 2002, ch. Informed Student Guide to Management Science. [Online]. Available: https://robjhyndman.com/papers/BoxJenkins.pdf

[2] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák, "Bayesian workflow," 2020. [Online]. Available: https://arxiv.org/abs/2011.01808

[3] J. Durbin and S. Koopman, *Time Series analysis by state space methods*. Oxford University press, 2012, no. Second Edition.

[4] S. Taylor and B. Letham, "Forecasting at scale," *PeerJ Preprints 5:e3190v2*, 2017. [Online]. Available: https://doi.org/10.7287/peerj.preprints.3190v2

[5] O. Bogdan and C. Stefan, Cristian, "Time series forecasting using neural networks," *CoRR*, vol. abs/1401.1333, 2014. [Online]. Available: http://arxiv.org/abs/1401.1333

[6] U. Lotrič and A. Dobnikar, "Using smoothing splines in time series prediction with neural networks," in *Artificial Neural Nets and Genetic Algorithms*. Vienna: Springer Vienna, 1999, pp. 121–126.

[7] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Phil. Trans. R. Soc. A.371201105502011055O*, 2013. [Online]. Available: http://doi.org/10.1098/rsta.2011.0550

[8] R. J. Hyndman, *ARIMA processes*. Hans Daellenbach and Robert Flood, 2002, ch. Informed Student Guide to Management Science. [Online]. Available: https://robjhyndman.com/papers/ARIMA.pdf

[9] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 2nd ed. OTexts Melbourne, Australia, 2018, ch. 8.9. [Online]. Available: https://otexts.com/fpp2/seasonal-arima.html

[10] J. S. Speagle, "A conceptual introduction to markov chain monte carlo methods," 2020. [Online]. Available: https://arxiv.org/abs/1909.12313

[11] T. S. D. Team. Stan. [Online]. Available: https://mc-stan.org/

[12] R. Muller and E. Muller. Berkeley earth. [Online]. Available: http://berkeleyearth.org

[13] Kaggle. Pfizer stock price (all time). [Online]. Available: https://www.kaggle.com/kannan1314/pfizer-stock-price-all-time

[14] R. J. H. Spyros G. Makridakis, Steven C. Wheelwright, *Forecasting Methods and Applications*, 1998, ch. 7.

[15] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, 4th ed. Springer International Publishing, 2017, ch. 3.3.

[16] G. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970, ch. 7.

[17] P. Stoica, B. Friedlander, and T. Söderström, "A high-order yule-walker method for estimation of the ar parameters of an arma model," *Systems & Control Letters*, vol. 11, no. 2, pp. 99–105, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167691188900825

[18] C. Dimitriou-Fakalou, "Yule-walker estimation for the moving-average model," *International Journal of Stochastic Analysis*, vol. 2011, p. 151823, Aug 2011. [Online]. Available: https://doi.org/10.1155/2011/151823

[19] E. Mahdi, "Portmanteau test statistics for seasonal serial correlation in time series models," *SpringerPlus*, vol. 5, 2016. [Online]. Available: https://doi.org/10.1186/s40064-016-3167-4

[20] G. M. LJUNG and G. E. P. BOX, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 08 1978. [Online]. Available: https://doi.org/10.1093/biomet/65.2.297

[21] R. R. Chaired, "Unit root tests," University of Washington. [Online]. Available: https://faculty.washington.edu/ezivot/econ584/notes/unitroot.pdf

[22] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of Econometrics*, vol. 54, no. 1, pp. 159–178, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/030440769290104Y

[23] T. W. Epps and L. B. Pulley, "A test for normality based on the empirical characteristic function," *Biometrika*, vol. 70, no. 3, pp. 723–726, 1983. [Online]. Available: http://www.jstor.org/stable/2336512

[24] I. N. Lobato and C. Velasco, "A simple test of normality for time series," *Econometric Theory*, vol. 20, no. 4, pp. 671–689, 2004. [Online]. Available: http://www.jstor.org/stable/3533541

[25] M. E. Lopes, L. J. Jacob, and M. J. Wainwright, "A more powerful two-sample test in high dimensions using random projection," 2015.

[26] A. A. Matamoros, A. Nieto-Reyes, R. Hyndman, M. O'Hara-Wild, and T. A., "nortstest: Assessing normality of stationary process," 2021, r package version 1.0.3. [Online]. Available: https://CRAN.R-project.org/package=nortsTest

[27] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[28] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 2021, ch. 11.

[29] J. Gabry and T. Mahr, "bayesplot: Plotting for bayesian models," 2021, r package version 1.8.1. [Online]. Available: https://mc-stan.org/bayesplot/

[30] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman, "Visualization in bayesian workflow," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 182, no. 2, pp. 389–402, 2019. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12378

[31] A. Vehtari, A. Gelman, and J. Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and Computing*, vol. 27, no. 5, p. 1413–1432, Aug 2016. [Online]. Available: https://arxiv.org/pdf/1507.04544.pdf

[32] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 3rd ed., 2021, ch. 5.10. [Online]. Available: https://otexts.com/fpp3/tscv.html

[33] A. Vehtari, J. Gabry, M. Magnusson, Y. Yao, P.-C. Bürkner, T. Paananen, and A. Gelman, "loo: Efficient leave-one-out cross-validation and waic for bayesian models," 2020, r package version 2.4.1. [Online]. Available: https://mc-stan.org/loo/

[34] M. Magnusson, M. R. Andersen, J. Jonasson, and A. Vehtari, "Leave-one-out cross-validation for bayesian model comparison in large data," 2020. [Online]. Available: https://arxiv.org/abs/2001.00980

[35] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 3rd ed., 2021, ch. 10. [Online]. Available: https://otexts.com/fpp3/dynamic.html

[36] A. Dala. Seminario de investigación. [Online]. Available: https://github.com/Andres-Dala/Seminario_de_Investigacion_MM700