A Robust Workflow for Forecasting with Bayesian SARIMA Models

by Daniel Dala and Asael Matamoros

Abstract The Box-Jenkins method shows a set of steps to make predictions on time series with ARIMA models using a frequentist approach. In this study we present an adaptation of this method to a Bayesian approach. For the parameter estimation we use Bayesian inference approximating the posterior distribution of the models using Markov Chain Monte Carlo methods. For the model checking, is analysed the behaviour and distribution of the errors that follow an assumption of Gaussian white noise and for the model comparison the differences in the precision of the predictions are measured using cross-validation and comparing the predictions in a test set. Finally, we demonstrate the performance of the proposed methodology on two applications, one forecasting the average monthly temperature in Honduras and one forecasting the monthly closing price in the stock market of Pfizer Inc.

Introduction

One of the most important applications in time series analysis is prediction, that is, estimating future values that are generally unknown, for this there are different methodologies such as space and state models Durbin and Koopman (2012), Prophet Taylor and Letham (2017), neural networks Bogdan and Stefan (2014), splines Lotrič and Dobnikar (1999), Gaussian processes Roberts et al. (2013), among others. A very popular class of models for their easy interpretation and high predictive capacity are the SARIMA models Hyndman (2002a); Hyndman and Athanasopoulos (2018), but their implementation with real data is complex because selecting the order of the model is a complicated task. Box and Jenkins (1970) proposed a methodology for the proper use of these models, which is based on six iterative stages: *data visualization, model selection, parameter estimation, model checking, model comparison and forecasting*. This methodology is illustrated in Figure 1.

There are many schemes for the inference process, and in recent years Bayesian inference has become a widely used alternative for data analysis with many applications in economics, physics, chemistry, psychology, among others. Its growing popularity is due to its ability to incorporate external information into the model through a priori distribution, and update beliefs through Bayes' Theorem. This inference approach in practice is very complicated, which is why in recent years the results have been approximated using the Markov Chain Monte Carlo methods Speagle (2020). These methods consist in generating a Markov chain whose stationary distribution is the posterior distribution of the model, there are many procedures to implement these methods, one of the most common is the Monte-Carlo Hamiltoneano, which due to its flexible implementation in the Stan language has been useful in multiple applications Team.

The biggest obstacle when performing an adequate data analysis in a Bayesian approach is that the estimation, checking, and selection procedures used in Box and Jenkins (1970b) are not valid in this new approach. propose an extensive and robust methodology called *"Bayesian workflow"*, which presents different tools for an adequate data analysis. This methodology is based on the one proposed by Box and Jenkins (1970b), and it is generalized for any type of modeling that involves a probabilistic inference approach.

The two main problems of the Gelman et al. (2020) method when applied in the analysis of time series are its complex structure, and that some tools are not suitable for data with dependency assumptions, therefore, in this study we present a simplification of the *Bayesian Workflow* with slight variations in some of the tools for their adequate use in time series. Finally, we apply our proposed methodology through two examples. In the first example we predict the average monthly temperature in Honduras with records from 1980 to 2013 obtained from Muller and Muller, and a second example analyzing the average monthly closing price of shares in the pharmaceutical company Pfizer with a set of recorded data monthly from 2010 to 2021 obtained from Kaggle.

Preliminaries and notation

For the purposes of this study, a stochastic process is an arbitrary collection of random variables $\{Y_1, Y_2, \ldots\}$, and a time series or simply series, it is a realization or finite sample $\{y_1, y_2, \ldots, y_n\}$ of the process. An important property to consider is stationarity, we will say that a process $\{y_i\}_{i\in\mathbb{Z}}$ is *strong*

stationary if for any finite collection of the process its joint distribution holds constant in time. This is

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

for $t \in \mathbb{Z}_+$ with $n \in \mathbb{N}$ and any $\tau \in \mathbb{Z}_+$. A less restrictive property is weak stationarity, we will say that the process $\{y_i\}_{i\in\mathbb{Z}}$ is *weak stationary* if the process has a constant mean and variance through the time, and the auto-correlation is a linear function of the difference of two times.

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

for $t, k \in \mathbb{Z}$ and $\tau > 0$. A series $\{y_t\}$ shows a trend over the mean of the process, if the mean can be represented as a function in time $y_t = f(t) + \varepsilon_t$, where $\{\varepsilon_t\}$ is a process with zero mean and $f : \mathbb{Z} \to \mathbb{R}$ is a measurable function. To transform a process with a trend into a stationary one, we apply the difference operator

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

For a process y with a linear trend, the process ∇y obtained in the equation 1 is stationary. The cyclicality in a series implies multiple periodic oscillations in the mean of the process, a particular case is seasonality, this happens when the series presents a constant periodic oscillation of period m in the mean. To transform a seasonal process into a stationary one, we apply the seasonal difference operator

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

Where m is a positive integer that represents the period of the series, and for a series y with seasonality, the process $\nabla_m y$ obtained in the equation 2 is stationary. It is important to emphasize that the series with trend or seasonality are non-stationary, for modelling purposes, it is necessary to work with stationary processes. An example of stationary processes is white noise, a collection of independent variables with normal distribution, zero mean, and positive constant variance.

SARIMA models

Let $\{Y_i\}_{i=1}^n$ be a time series, we say that the series follows a model textit Autoregressive Integrated Moving Average $ARIMA(p,d,q)$ if for any time Y_t, can be written in the form:

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

where μ_0 is the initial mean of the process, $p \in \mathbb{Z}_+$ and $\{\phi_i\}_{i=1}^p$ are the order and parameters of the autoregressive component respectively, $q \in \mathbb{Z}_+$ and $\{\theta_i\}_{i=1}^q$ are the order and parameters of the moving average component respectively, $d \in \mathbb{Z}_+$ represents the number of non-seasonal differences and $\varepsilon_t \sim N(0, \sigma_0)$ is Gaussian white noise centered at zero and with constant positive variance.

The model proposed in the equation 3 can be adapted to analyse time series with seasonality, this can be achieved by adding autoregressive and moving averages components to model seasonality additively, and adapting the seasonal difference of multiplicative form. Let $\{Y_i\}_{i=i}^n$ be a time series with seasonality and period $m \in \mathbb{Z}_+$, we say that the series follows a *Multiplicative Seasonal ARIMA model* $SARIMA(p,d,q) \times (P,D,Q)_m$ if for any time Y_t,

$$Z_t = \mu_0 + \sum_{i=1}^{p} \phi_i Z_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{P} \Phi_i Z_{t-im} + \sum_{i=1}^{Q} \Theta_i \varepsilon_{t-im} + \varepsilon_t, \tag{4}$$

$$Z_t = \nabla_m^D \nabla^d y_t,$$

where $\varepsilon_t \sim \mathrm{N}(0, \sigma_0)$ is a Gaussian white noise with positive constant variance, the parameters $\mu_0, p, \{\phi_i\}_{i=1}^p, q, \{\theta_i\}_{i=1}^q, d$ are the same as defined in the equation 3, $P \in \mathbb{Z}_+$ and $\{\Phi_i\}_{i=1}^P$ are the order and parameters of the seasonal autoregressive component respectively, $Q \in \mathbb{Z}_+$ and $\{\Theta_i\}_{i=1}^Q$ are the order and parameters of the seasonal moving average component respectively, and $D \in \mathbb{Z}_+$ represents the number of seasonal differences. Note that Z_t is the transformation obtained by applying differences and seasonal transformations in a multiplicative way.

Box-Jenkins Method

The procedure for forecasting future values in time series requires two fundamental stages: analysis of the data and selection of the forecast model that best fits the data. Each of the steps will be briefly explained below, for a more in-depth study of the method read Spyros G. Makridakis (1998).

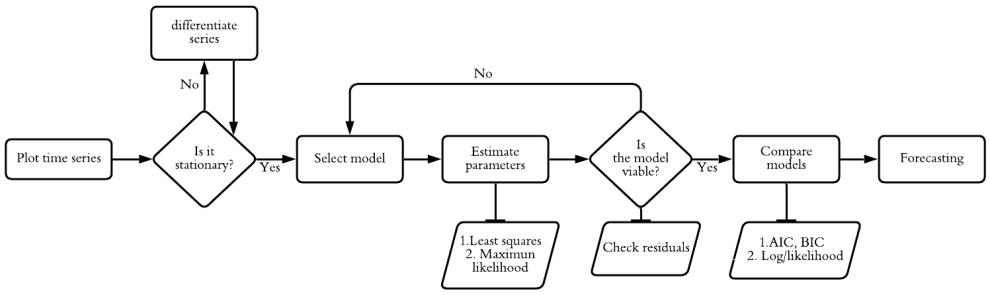

Figure 1: Box-Jenkins methodology (1970).The flow diagram presents the procedure to be used for an adequate data analysis in a frequentist approach, for a further description of said methodology, review Hyndman (2002b).

1. *Data visualization:* visualization through graphs allows the detection of patterns such as trends, seasonality, cycles or atypical observations, which must be filtered by differentiating the series or applying seasonal differences. The graphs ACF (autocorrelation function) Shumway and Stoffer (2017) and PACF (partial autocorrelation function) Shumway and Stoffer (2017) are of great importance to select the orders of the model Durbin and Koopman (2012).
2. *Model selection:* to define an initial model it is necessary to establish the values p, d, q, P, D, Q and m. The values d, D are the number of differences necessary for the series to be stationary or a white noise, this is achieved by graphing the original series and the differentiated series. The values (p, P) and (q, Q) are identified with the graphs PACF and ACF respectively, as the number of lags different from zero in the differentiated series, for more details see Durbin and Koopman (2012).
3. *Parameter estimation:* once an initial model has been defined, it is necessary to estimate the $n_p = p + P + q + Q + 2$ parameters, the most used methods are: *least squares* Box and Jenkins (1970a), *maximum likelihood* Box and Jenkins (1970a), and the *Yule-Walker equation* Stoica et al. (1988); Dimitriou-Fakalou (2011).
4. *Model checking:* the *SARIMA* models follow the assumption that the errors follow a Gaussian white noise, that is, the errors are stationary with normal distribution. For diagnosis of stationarity, the Portmanteau Mahdi (2016) and Lijung-Box tests LJUNG and BOX (1978) are used, or unit root tests such as the Augmented Dickey-Fuller Chaired, Phillips-Perron Chaired, and KPSS tests Kwiatkowski et al. (1992). To measure normality tests such as Epps Epps and Pulley (1983), Lobato-Velasco Lobato and Velasco (2004) and random projections Lopes et al. (2015) that measure normality in stationary processes are the most appropriate. For more details see Matamoros et al. (2021).
5. *Model comparison:* the most widely used model selection criterion is the *Akaike's Information Criteria (AIC)* proposed by Akaike in 1974 Akaike (1974). Let $n_p = p + q + P + Q + 2$ be the number of parameters estimated in the model, then the values of p, q, P, Q that minimize the AIC are chosen:
$$AIC = -2logL + 2n_p, \tag{5}$$
where L denotes the likelihood. There are several modifications of the AIC that are also used as the BIC (Bayesian Information Criteria) and the log-likelihood.
6. *Forecasting:* once the model that best fits each of the previous tests has been chosen, the prediction of future observations is made.

Bayesian inference

In a Bayesian inference approach, the probability of a parameter $\theta \in \Theta$ is analysed given the sample obtained y, where Θ is a probability space for the set of parameters. For that, we must start by establishing a model that comes from a joint probability distribution for θ and y. The joint probability function can be written as the product of two densities:

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

where $p(\theta)$ is called the *prior* distribution and $p(y|\theta)$ the sample distribution or *likelihood*. By conditioning the known value of the data y and using Bayes Theorem we obtain the *posterior* distribution:

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

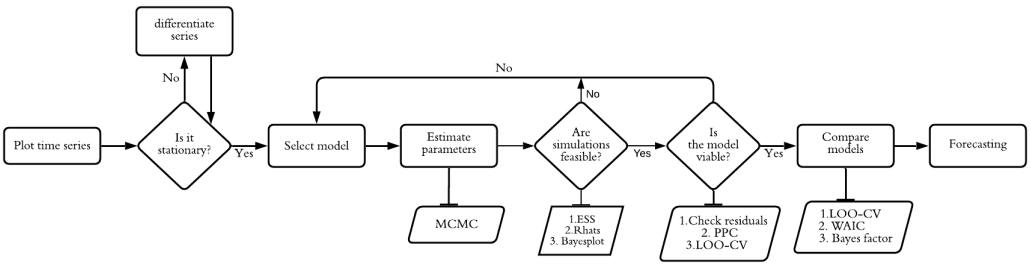

Figure 2: Methodology for prediction with Bayesian SARIMA models. The flow diagram presents the adaptation of the Box-Jenkins method adapted to a Bayesian approach based on the Bayesian workflow proposed by Gelman et al. (2020).

An equivalent form of the previous equation omits the factor $p(y)$ which, as it does not depend on θ is considered a constant, resulting in a non-normalized posterior distribution, in other words the posterior is proportional to the likelihood and prior, using the following equation:

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

Note that $p(y|\theta) : \Theta \to \mathbb{R}$ is a function of θ when the sample y is fixed, and is known as a *likelihood function.* These equations are enough to perform Bayesian inference where first a model $p(\theta, y)$ is established and then the computational calculations are developed to obtain $p(\theta|y)$ which works as the updated information of the data.

On the other hand, when inferring unknown observations or predictions, we follow a similar procedure. After inferring our parameters θ from the observed data y, we can predict an unknown observation $\tilde{y}$. The distribution of $\tilde{y}$ is known as the *posterior predictive distribution* and is obtained with the following equation:

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

Bayesian SARIMA models

Based on the previous definitions, a *Bayesian SARIMA model* is defined as a seasonal ARIMA model and a selection of independent priors for each of the unknown parameters, following the equations:

$$y \sim SARIMA(p, d, q)_{\times}(P, D, Q)_m$$

$$\begin{aligned} \phi_i &\sim p(\phi_i), \quad i = 1, ..., p \\ \theta_j &\sim p(\theta_j), \quad j = 1, ..., q \\ \Phi_k &\sim p(\Phi_k), \quad k = 1, ..., P \\ \Theta_w &\sim p(\Theta w), \quad w = 1, ..., Q \end{aligned}$$

$$\begin{aligned} \mu_0 &\sim p(\mu_0) \\ \sigma_0 &\sim p(\sigma_0) \end{aligned}$$

Where the data *and* are a realization of a stochastic process that follow the equation (4), y $\theta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \Phi_1, \ldots, \Phi_P, \Theta_1, \ldots, \Theta_Q, \mu_0, \sigma_0) \in \mathbb{R}^{n_p}$ is the vector of unknown parameters.

Adaptations to the Box-Jenkins method

When proposing a Bayesian SARIMA model to make the forecasting, we must make modifications and adaptations to the Box-Jenkins method, this given that the parameter estimation and checking are based on Bayesian inference. Initially the *data visualization* and *model selection* steps remain the same, so the following sections will show the *selection, checking and comparison of the models* steps that will be seen modified following the line of the *Bayesian Workflow* adapted to the analysis of time series.

Model estimation

Once the initial model is established and the prior distributions of each parameters, the posterior distribution is approximated by Hamiltonian Monte Carlo which simulates a stationary markov chain that converges to the distribution of each parameter. For a better understanding of this inference process it is recommended to read Gelman et al. (2021).

Inference evaluation

After making the inference, it is necessary to evaluate the convergence of the simulations by measuring the stationarity and combination of the simulated chains for each of the parameters. In this way, to decide the viability of the results, two statistics will be used, the *(Effective Sample Size ESS)* and the (*Potential scale reduction* $\hat{R}$), the first indicates the sufficient size of the simulations to correctly approximate the parameters and the second is an indicator of convergence of the chains that for this study will be taken as a feasible estimator if $\hat{R} < 1.1$. The general properties of both estimators are found in Gelman et al. (2021).

It is also recommended to do a graphic analysis of the fit to each parameter, that is, to observe the histograms and graphs of the generated chains in search of signs of multimodality in the distribution of the parameters and to verify that the chains show convergence. This graphical analysis can be performed with the *bayesplot* Gabry and Mahr (2021) package and examples can be seen in the article Gabry et al. (2019).

Model checking

As in the Box-Jenkins methodology, it is advisable to make a check of the models, checking that the errors behave like a Gaussian white noise. On the other hand, in Bayesian inference, there are methods to verify that the model effectively represents the observed data, the main method is *Posterior predictive check (PPC) (Box, 1980, Rubin, 1984, Gelman, Meng, and Stern , nineteen ninety six)*. If the model fits well, it should generate data with the same behaviour as the observations. However, PPC is not feasible in the analysis of time series because the assumptions of interchangeability are not fulfilled in the data as they are realizations of a stochastic process. To avoid these problems, it is recommended to do PPC on the model residuals ($\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$) which are stationary and interchangeable, so they show optimal conditions for the creation of histograms and obtaining conclusive results.

Generally, the PPC is sufficient to find errors in the model fit, however, since we use the observations to fit the model and make the evaluations, it is possible that in some cases atypical behaviours are allowed to pass through the data. An alternative way is to make the diagnosis with *Leave-one-out cross-validation (LOO-CV)* where a part of the data is used to fit the model and the rest is used to measure the precision of prediction. In Gelman et al. (2020) three ways of approaching evaluation using cross-validation are advised: 1. Calibration checks using the cross-validation predictive distribution 2. Identify which observations or groups of observations are most difficult to predict 3. Identify how influential are the particular observations, that is, how much information they provide in addition to other observations. For a better understanding of LOO-CV it is recommended to read Vehtari et al. (2016).

Model comparison

In the comparison of Bayesian models, LOO-CV and the Watanabe-Akaike Information Criterion, WAIC (Watanabe, 2010) are frequently used. Both methods estimate the point precision in the prediction in a model using a log-likelihood sample. However, the method for estimating precision for time series models that proposes the best estimates is the cross-validation of time series Hyndman and Athanasopoulos (2021a), but due to the difficulty of implementation in this study we will work with LOO-CV and WAIC only. The implementation of these methods can be done with the *loo* Vehtari et al. (2020) package. For a better understanding of them, it is recommended to read Magnusson et al. (2020); Vehtari et al. (2016).

Finally, after choosing the models that present better results in the previous tests, the predictions are made with the selected models. The proposed tools offer sufficient precision for a good estimation and selection of models, however, careful analysis is required when making predictions of future values. The following sections will show examples with real data, applying our proposed methodology to select the SARIMA model that best fits the data in a Bayesian approach.

Illustrations

We will apply the new methodology by studying two data sets and making their respective predictions. For the inference and analysis of the models we will use the *bayesforecast* package that implements these models using an Hamiltonian Monte Carlo, generating 4 chains of 2,000 iterations and a warm-up of 1,000 iterations. As a convergence diagnosis we will use the statistic $\hat{R}$ Gelman et al. (2021) and the comparison of the models will be with cross-validated using the *loo* Vehtari et al. (2016) package.

Average temperature in Honduras

The first data set shows the monthly average temperature in Honduras from 1980 to 2013, with a total of 405 observations where the first 385 will serve as the training set of the model and the rest will be the test set. These data were obtained from *Berkeley Earth* Muller and Muller.

```
> library(bayesforecast)
> library(ggplot2)

> autoplot(object = train, ylab="Celsius", xlab = "Years",size=1) +
+     theme(panel.background = element_rect(fill = "white", colour = "black",
+                                           size = 1, color = "black"),
+           panel.grid.major = element_blank(),
+           panel.grid.minor = element_blank(),
+           axis.text=element_text(size=12, color = "black"),
+           axis.title=element_text(size=14,family="serif"))
```

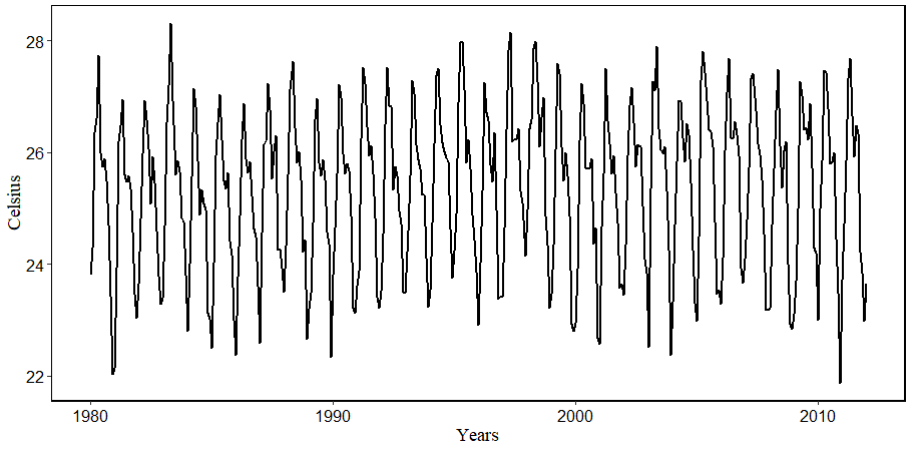

Figure 3: Record of the monthly average temperature in Celsius for Honduras from 1980 to 2013. The series does not present a trend but it does present strong variations and periodic oscillations. Therefore, a seasonal SARIMA model is suitable for the analysis of the series.

```
> head(serie)
Jan    Feb    Mar    Apr    May    Jun
1980 23.798 24.296 26.346 26.595 27.735 25.984
```

Figure 3 shows that the data do not follow a stationary behaviour, and since they are climatological observations it is expected that there will be annual seasonal patterns, therefore a seasonal difference with period 12 and a non-seasonal difference will be applied.

```
library(cowplot)

g1 = cbind("Seasonally differenced" = diff(train, 12),
"Doubly differenced" = diff(diff(train, 12))) %>%
autoplot(facets = TRUE) + xlab("Years") + ylab("")

g2 = ggacf(y = diff(train))
```

```
g3 = ggpacf(y = diff(train))

g4 = plot_grid(g2,g3, labels = NULL)
plot_grid(g1, g4, labels = NULL, ncol = 1, rel_heights = c(1.3,1))
```

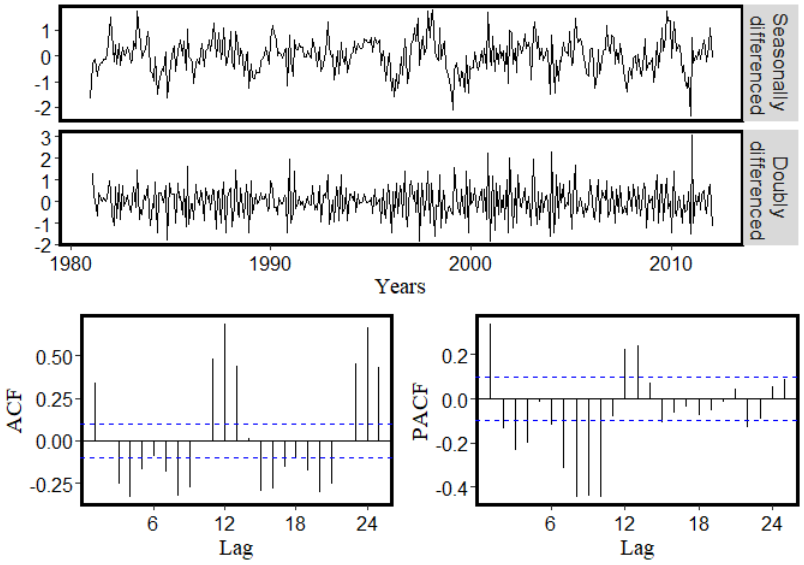

Figure 4: The upper graph shows the series with a seasonal difference with period 12 and then non-seasonally differentiated. The lower graphs show the ACF and PACF functions of the doubly differentiated data.

Figure 4 shows that when applying a seasonal difference the series in the upper part does not look stationary since the mean is not constant, however the seasonality is reduced, which indicates that it is enough to make a second difference not seasonal. The middle graph shows the doubly differentiated series and this appears to have a constant mean at zero and a stable variance. When observing the autocorrelation (ACF) and partial autocorrelation (PACF) functions of the lower part of Figure 4, they show a slight correlation of at most two lags in the data, the order of the initial SARIMA model to consider then is $p = 2$, $d = 1$, $q = 2$. On the other hand, both graphs show certain periodic patterns and the lags indicate a possible seasonal autoregressive component, therefore the order of the seasonal component is $P = 2$, $D = 1$, $Q = 0$. Finally, we consider the prior distributions of each of the parameters, in this case we select priors that are not very informative. In this way the complete model is:

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

Table **??** shows a summary of the posterior distributions of each of the parameters. The statistic $\hat{R}$ of each of them indicates that the chains converge, and the effective sample sizes (ESS) are values greater than the total number of iterations, indicating a feasible sample size for the effective representation of the parameters, therefore, we accept the approximation of the obtained posteriors.

In Figure 5 it is observed that the chains appear to be stationary indicating convergence, in addition there is no multi-modality in the posterior distribution of the parameters, therefore we can accept the estimates of the model and continue with the checking of fit in data. Figure 6 shows a summary of the residuals of Model 1. The upper graph shows the series of residuals, which presents a slight negligible cyclicality and does not present volatility. The ACF and PACF graphs at the bottom of Figure 6 show a low correlation, maintaining the confidence intervals, therefore, the residuals appear stationary. The histogram and the quantile graph (intermediate) show that the posterior mean of the residuals have a symmetric distribution and without heavy tails indicating normality. Therefore, we conclude that the residuals follow a Gaussian white noise, satisfying the assumptions for Model 1. We compare the model obtained with two alternative models of different orders to those proposed for the initial model,

	mean	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

Table 1: Summary of the posterior distributions of each of the parameters for Model 1. The statistics presented are the posterior mean (mean), Standard error (SE), credibility intervals at 90%, effective sample size (ESS) and the potential scale reduction ($\hat{R}$.)

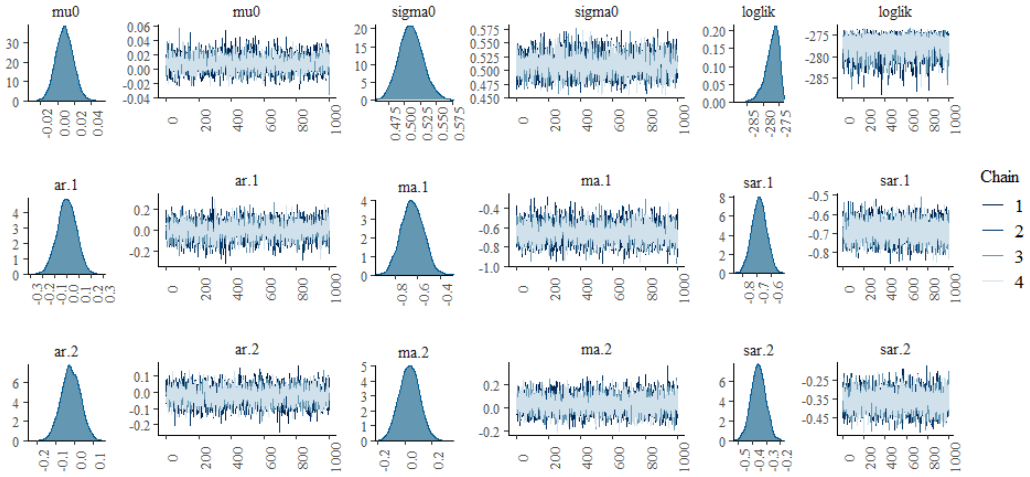

Figure 5: A posterior density graphs and simulated chains for each of the estimated parameters.

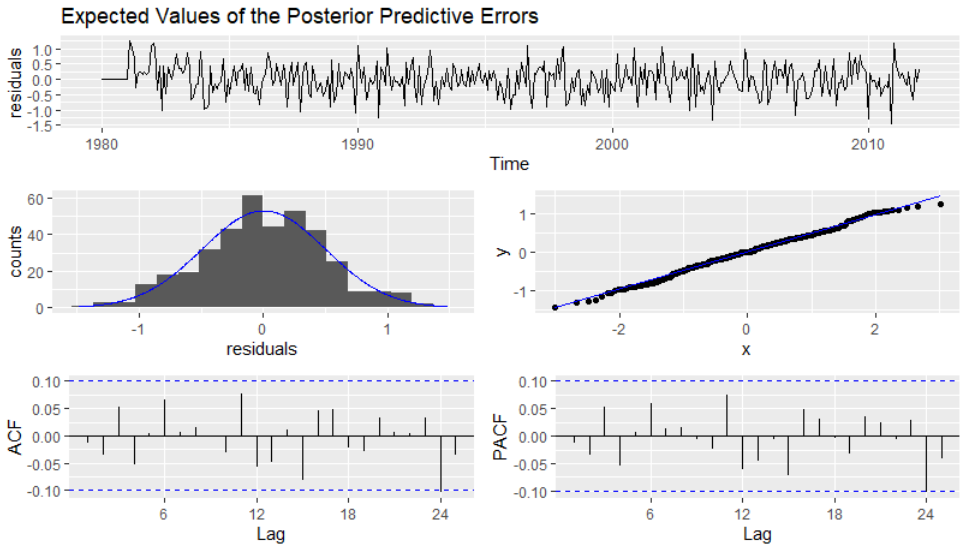

Figure 6: The series of residuals (top). The histogram and quantile graph (middle part). Graphs of autocorrelation and partial autocorrelation of the model residuals (bottom).

which are defined in the following two equations:

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

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

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

Table 2: Comparative summary of the precision in each model, where Model 1 shows the best results and Model 3 shows the worst.

Model 2 has prior distributions slightly different from the other models, for the auto-regressive and moving average parameters (ar, ma, sma) a $Beta(2,2)$ distribution was selected transformed in such a way that the values obtained are in the range $[-1,1]$, guaranteeing the stationarity of the process. The rest of the parameters will remain with the same distributions as Model 1. It is important to note that the same procedure was used to estimate and diagnose Models 2 and 3, validating each of the stages of the process. We proceed to compare the models using cross-validation, which estimates the *elpd*

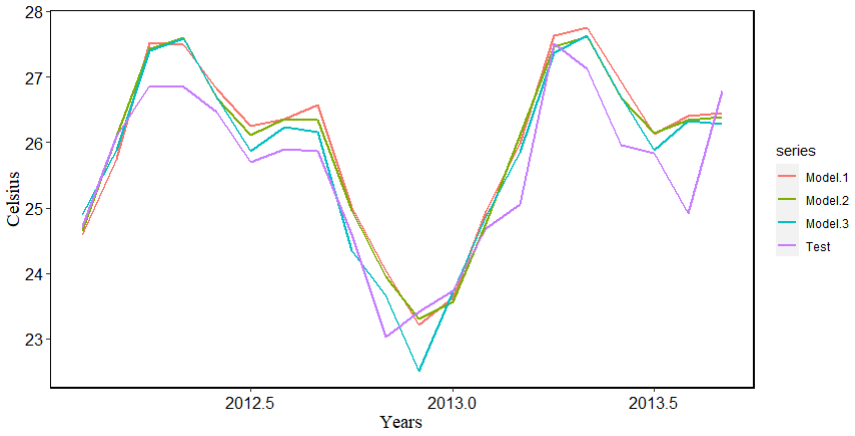

Figure 7: Comparison of predictions of the models with the test set.

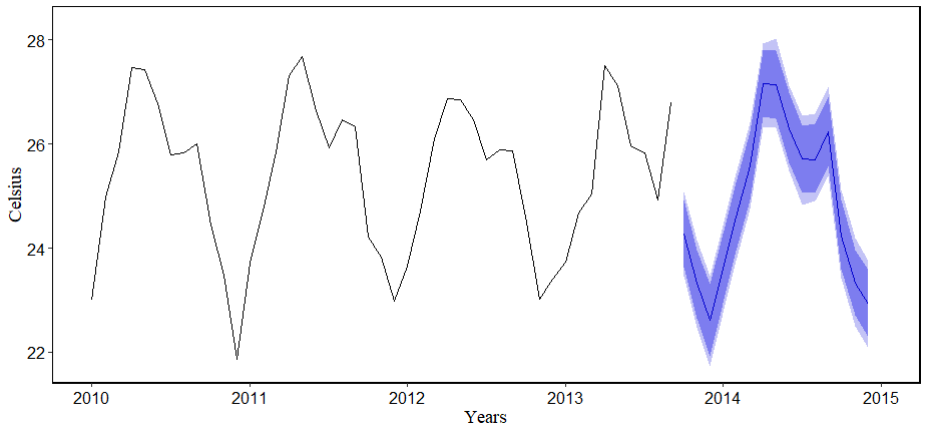

Figure 8: Forecast generated by Model 1.

(Expected log predictive density) measure to establish the predictive capacity of each model. Table 2 shows the difference between *elpds* , where Model 1 has a higher predictive capacity than the other two. Additionally, the upper part of Figure 8 presents the predictions made by each model comparing the results with the data of the test set, it is observed that each of the models generate very reliable predictions that are similar to each other. Finally, we select Model 1 to predict the next 15 months, the predictions are made from the month of September 2013 and the results are displayed in the lower part of Figure 8.

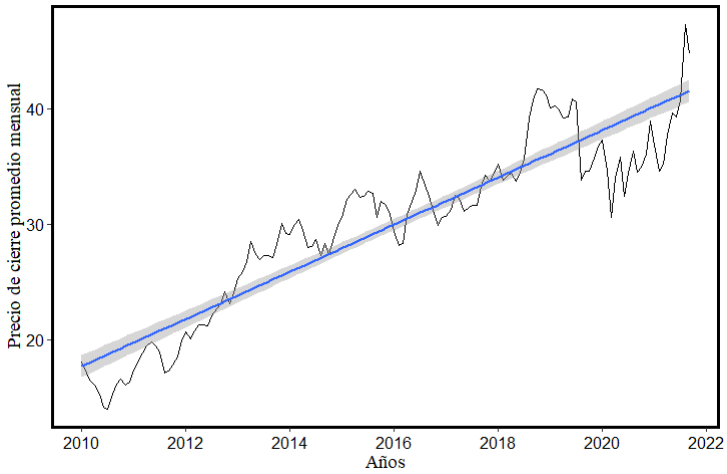

Figure 9: Average closing price in dollars from January 2010 to September 2021.

Closing price on Pfizer stocks

For the second example we study the closing price of the stocks in the pharmaceutical company Pfizer, in this case the data set is obtained from Kaggle, the data shows the main attributes of the shares from June 1972 to September 2021 with 4 monthly measurements. For convenience, we will analyse the monthly average of the closing price since January 2010. The final series with which we will work consists of 141 observations, where the first 134 will be the training set and the rest the test set. Figure 9 shows that the data have an increasing trend that is modeled with a SARIMA model with deterministic trend Hyndman and Athanasopoulos (2021b). On the other hand, in Figure 10 we observe that the data stabilizes from the first difference, it is also important to note the high volatility that occurs in the period between 2019 and 2021. Based on these facts and taking as reference the ACF and PACF graphs that indicate a possible autoregressive model, we will define the model to be used:

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

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

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

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

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

The model y_t is a regression model with its errors η_t, in this way we infer in each one of the parameters of the SARIMA model and the regression parameter. Table 3 shows a summary of the posterior distributions of the parameters. The statistic $\hat{R}$ indicates convergence for each of the parameters, which is visually confirmed in Figure 11, the graphs of the posterior distributions do not show multi-modality and the chains seem stationary, which in this context are sufficient conditions to suppose that a feasible inference was obtained. Then, when looking at Figure 12 it is shown in the series of residuals that

	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

Table 3: Summary of the posterior distributions of each of the parameters for Model 1. The statistics presented are the posterior mean (mean), Standard error (SE), credibility intervals at 90%, effective sample size (ESS) and the potential scale reduction ($\hat{R}$.)

the model does not correctly explain the period between 2019 and 2021 this due to the high volatility in those years, which is also represented in the density graph and quantiles showing the presence of heavy tails, however, given the low autocorrelation shown in the ACF, PACF graphs and the rest of the series of residuals, this is a feasible model for the adjustment of the data. On the other hand, the

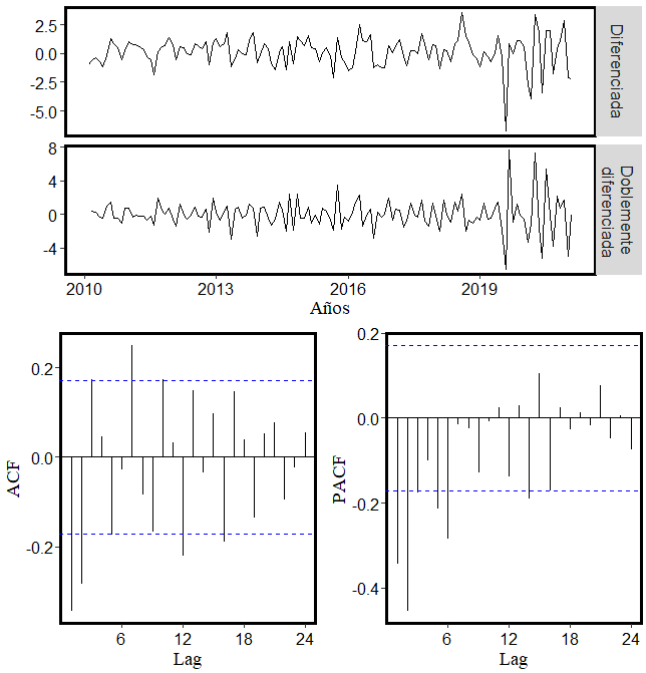

Figure 10: The upper graphs show the non-seasonally differentiated and doubly differentiated data. The lower graphs show the graphs of the ACF and PACF of the doubly differentiated data.

autocorrelation plots show that an increase in the order of the autoregressive model could improve the results. Now, in the comparison of models, two other regressions with ARIMA errors are proposed, which are:

$$\text{Model 2: } y_t = \beta_1 t + \eta_t$$
$$\eta_t \sim ARIMA(1,1,0)$$
$$\text{Model 3: } y_t = \beta_1 t + \eta_t$$
$$\eta_t \sim ARIMA(1,0,0)$$

For Model 2, we try to analyse the data without a seasonal component and for Model 3 a component $AR(1,0,0)$ is proposed without making a difference, which indicates that the errors are based solely on the autoregressive adjustment, so the comparison is based specifically on the order of the errors since the prior distributions of the three models are the same. Figure 13 shows that Model 3 shows the worst results, however, Models 1 and 2 show similar predictions. On the other hand, when applying CV-LOO and calculating the precision differences in the models, it is obtained that Model 1 presents a higher precision than Model 2, therefore the final prediction is made using the first model. Observing Figure 13 shows a prediction with linear behaviour, however, in the context of the observations and the increasing trend, it is possible that the real values of the future are greater than the prediction obtained, however, it is a good first prediction that can be optimized by including exogenous variables.

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

Table 4: Comparative summary of the precision in each model, where Model 1 shows the best results and model 3 shows the worst.

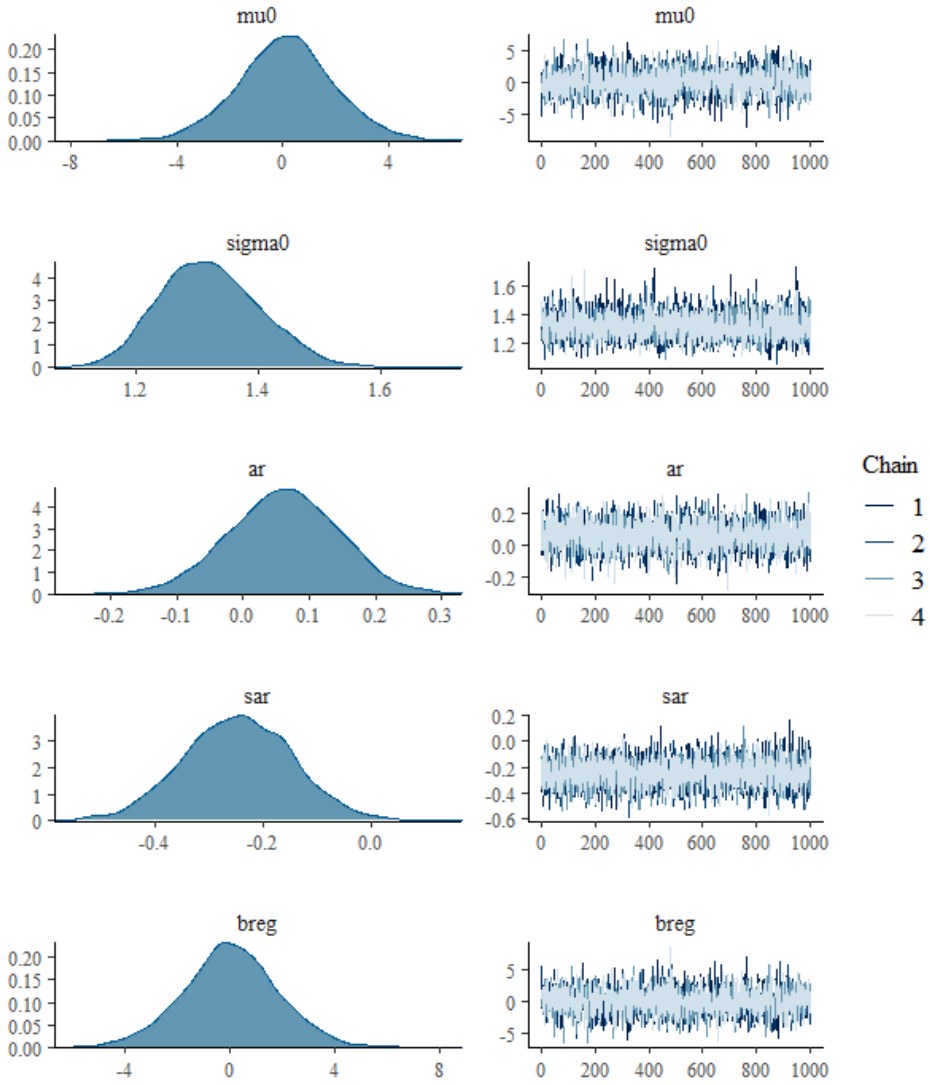

Figure 11: Plots of the posterior densities and the simulated chains for each of the estimated parameters.

Conclusions

In this study we introduce a new methodology for the analysis and prediction of time series with SARIMA models in a Bayesian approach. The proposed methodology allows to carry out an adequate process of inference, checking and selection of models, for the analysis of time series with SARIMA models. We have presented the performance and applicability of the methodology through two examples where both presented satisfactory results.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705. [p3]

O. Bogdan and C. Stefan, Cristian. Time series forecasting using neural networks. *CoRR*, abs/1401.1333, 2014. URL http://arxiv.org/abs/1401.1333. [p1]

G. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*, chapter 7. Holden-Day, 1970a. [p3]

G. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970b. [p1]

R. R. Chaired. Unit root tests. University of Washington. URL https://faculty.washington.edu/ezivot/econ584/notes/unitroot.pdf. [p3]

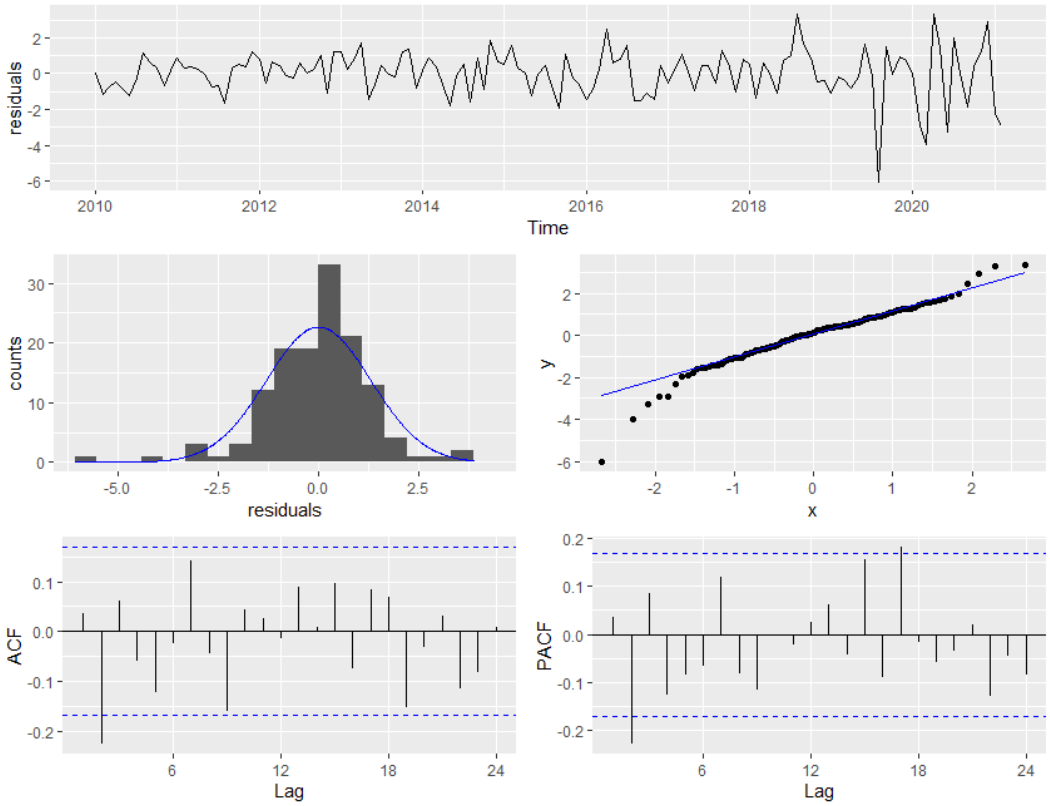

Figure 12: The series of residuals (top). The histogram and quantile graph (middle part). Graphs of autocorrelation and partial autocorrelation of the model residuals (bottom).

C. Dimitriou-Fakalou. Yule-walker estimation for the moving-average model. *International Journal of Stochastic Analysis*, 2011:151823, Aug 2011. ISSN 2090-3332. doi: 10.1155/2011/151823. URL https://doi.org/10.1155/2011/151823. [p3]

J. Durbin and S. Koopman. *Time Series analysis by state space methods*. Number Second Edition. Oxford University press, 2012. [p1, 3]

T. W. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. ISSN 00063444. URL http://www.jstor.org/stable/2336512. [p3]

J. Gabry and T. Mahr. bayesplot: Plotting for bayesian models, 2021. URL https://mc-stan.org/bayesplot/. R package version 1.8.1. [p5]

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)*, 182(2):389–402, 2019. doi: https://doi.org/10.1111/rssa.12378. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12378. [p5]

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. URL https://arxiv.org/abs/2011.01808. [p1, 4, 5]

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*, chapter 11. 2021. [p5, 6]

R. J. Hyndman. *ARIMA processes*, chapter Informed Student Guide to Management Science. Hans Daellenbach and Robert Flood, 2002a. URL https://robjhyndman.com/papers/ARIMA.pdf. [p1]

R. J. Hyndman. *Box-Jenkins modelling*, chapter Informed Student Guide to Management Science. Hans Daellenbach and Robert Flood, 2002b. URL https://robjhyndman.com/papers/BoxJenkins.pdf. [p3]

R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*, chapter 8.9. OTexts Melbourne, Australia, 2 edition, 2018. URL https://otexts.com/fpp2/seasonal-arima.html. [p1]

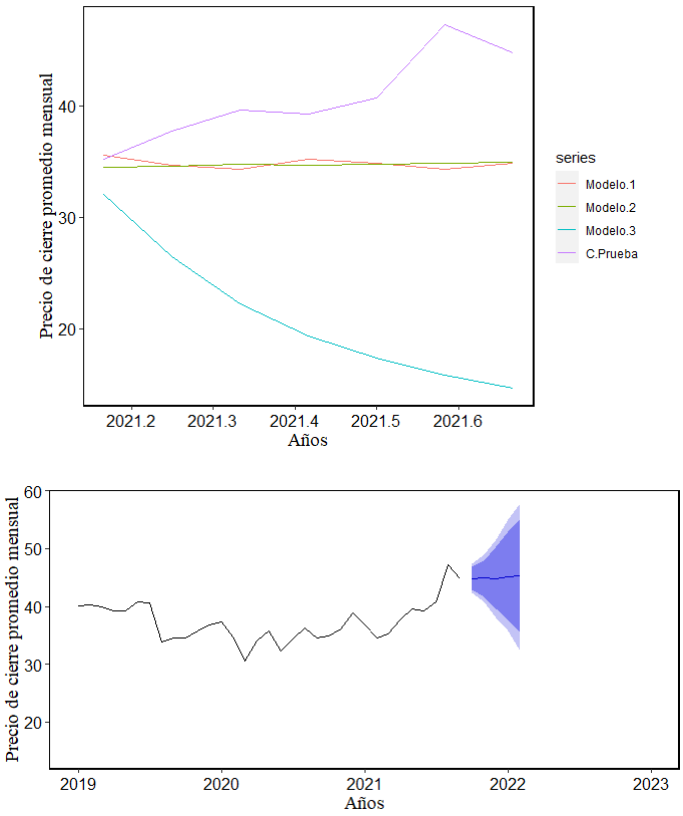

Figure 13: The upper part compares the predictions of the models with the test set. The lower part presents the prediction generated by Model 1, from October 2021 to February 2022.

R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*, chapter 5.10. 3 edition, 2021a. URL https://otexts.com/fpp3/tscv.html. [p5]

R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*, chapter 10. 3 edition, 2021b. URL https://otexts.com/fpp3/dynamic.html. [p9]

Kaggle. Pfizer stock price (all time). URL https://www.kaggle.com/kannan1314/pfizer-stock-price-all-time. [p1, 9]

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*, 54(1):159–178, 1992. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(92)90104-Y. URL https://www.sciencedirect.com/science/article/pii/030440769290104Y. [p3]

G. M. LJUNG and G. E. P. BOX. On a measure of lack of fit in time series models. *Biometrika*, 65(2): 297–303, 08 1978. ISSN 0006-3444. doi: 10.1093/biomet/65.2.297. URL https://doi.org/10.1093/biomet/65.2.297. [p3]

I. N. Lobato and C. Velasco. A simple test of normality for time series. *Econometric Theory*, 20(4): 671–689, 2004. ISSN 02664666, 14694360. URL http://www.jstor.org/stable/3533541. [p3]

M. E. Lopes, L. J. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection, 2015. [p3]

U. Lotrič and A. Dobnikar. Using smoothing splines in time series prediction with neural networks. In *Artificial Neural Nets and Genetic Algorithms*, pages 121–126, Vienna, 1999. Springer Vienna. ISBN 978-3-7091-6384-9. [p1]

M. Magnusson, M. R. Andersen, J. Jonasson, and A. Vehtari. Leave-one-out cross-validation for bayesian model comparison in large data, 2020. URL https://arxiv.org/abs/2001.00980. [p5]

E. Mahdi. Portmanteau test statistics for seasonal serial correlation in time series models. *SpringerPlus*, 5, 2016. URL https://doi.org/10.1186/s40064-016-3167-4. [p3]

A. A. Matamoros, A. Nieto-Reyes, R. Hyndman, M. O'Hara-Wild, and T. A. nortstest: Assessing normality of stationary process, 2021. URL https://CRAN.R-project.org/package=nortsTest. R package version 1.0.3. [p3]

R. Muller and E. Muller. Berkeley earth. URL http://berkeleyearth.org. [p1, 6]

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. URL http://doi.org/10.1098/rsta.2011.0550. [p1]

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*, chapter 3.3. Springer International Publishing, 4 edition, 2017. [p3]

J. S. Speagle. A conceptual introduction to markov chain monte carlo methods, 2020. URL https://arxiv.org/abs/1909.12313. [p1]

R. J. H. Spyros G. Makridakis, Steven C. Wheelwright. *Forecasting Methods and Applications*, chapter 7. 1998. [p2]

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*, 11(2):99–105, 1988. ISSN 0167-6911. doi: https://doi.org/10.1016/0167-6911(88)90082-5. URL https://www.sciencedirect.com/science/article/pii/0167691188900825. [p3]

S. Taylor and B. Letham. Forecasting at scale. *PeerJ Preprints 5:e3190v2*, 2017. URL https://doi.org/10.7287/peerj.preprints.3190v2. [p1]

T. S. D. Team. Stan. URL https://mc-stan.org/. [p1]

A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, Aug 2016. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4. URL https://arxiv.org/pdf/1507.04544.pdf. [p5, 6]

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. URL https://mc-stan.org/loo/. R package version 2.4.1. [p5]

Daniel Dala
Departamento de Estadística
Universidad Nacional Autónoma de Honduras
Honduras
(ORCiD if desired)
andresht6@hotmail.com

Author Two
Affiliation
Address
Country
(ORCiD if desired)
author2@work