Oil-price forecasting based on ARIMA, exponential smoothing, and autoregressive neural network models

Felipe B. P. Araújo[1], José Artur L. C. Marques [2], Allan Kardec D. Barros Filho. [3]

[1]*Dept. of Electrical Engineering, Federal University of Maranhão*
Av. dos Portugueses, 1966 – Vila Bacanga, 65080-805, Maranhão, Brazil
felipe.pimentel@discente.ufma.br
[2]*Dept. of Electrical Engineering, Federal University of Maranhão*
Av. dos Portugueses, 1966 – Vila Bacanga, 65080-805, Maranhão, Brazil
jarturcabral@gmail.com
[3]*Dept. of Electrical Engineering, Federal University of Maranhão*
Av. dos Portugueses, 1966 – Vila Bacanga, 65080-805, Maranhão, Brazil
akduailibe@gmail.com

Abstract. Financial time series are sensitive to exogenous shocks. From this perspective, this work presents a comparative analysis of predictive crude oil prices scenarios, obtained from a historical series of average annual prices. Two approaches were used: first, a combination of classic strategies based on exponential smoothing, and ARIMA models. Second, an autoregressive neural network model. Both approaches are complementary when used for long-term forecasting of oil prices and show good response to volatile data. Therefore, we are able to present an alternative data analysis, in a field where there is a great amount of relevant historical series, using probabilistic and non-linear models in order to observe predictions and make more effective decisions.

Keywords: Exponential Smoothing, ARIMA, Forecasting, Neural Network, Oil Prices.

1 Introduction

Uncertainty surrounds predictions of future events. If financial time series is used in data analysis, the uncertainty may be greater due to the volatility of the data. Therefore, in order to reduce this uncertainty, it is important to use statistical models that produce point forecasts and appropriate confidence intervals. In this perspective, exponential smoothing and ARIMA models could be used with historical oil prices to produce forecasts.

The data can be found in Looney [1] at the annual statistical report called BP Statistical Review of World Energy 2020 released by British Petroleum. The purpose of this work is to predict oil prices between 2020 and 2025, based on the crude oil time series presented in that report. So, the computer simulations were executed using R programming language, which has effective resources to manipulate time series as well as packages with several functionalities, including forecasting based on exponential smoothing, ARIMA, and autoregressive neural network models.

2 Exponential Smoothing

The exponential smoothing method aims to produce forecasts based on past observations. According to Cowpertwait and Metcalfe [2] to achieve this, it is assumed that there may be systematic effects of trend and seasonality when generating the data. So, it is wise to attribute an exponential smoothing parameter to the model when analyzing historical data. Because the model considers the effect of moving averages exponentially weighted in a decreasing manner as the observations get older.

The R package fable provides a great deal of exponential smoothing forecasting models. These models are organized in such a way that the selection of the best method to use is made according to the main components of the time series (eg, trend and seasonal components). Also, the way these components are added into the smoothing method (eg, additively, multiplicatively, or in a damped way) is also to be considered. Therefore, each method is labeled by two letters and they define the type of trend and seasonal component. All methods are presented in Table 1.

Table 1. Classification of exponential smoothing methods

Trend Component	Seasonal Component		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	(N , N)	(N , A)	(N , M)
A (Additive)	(A , N)	(A , A)	(A , M)
A_d (Additive damped)	(A_d , N)	(A_d , A)	(A_d , M)

A third letter was later added to this classification. This letter refers to addictive or multiplicative errors. This transformed the methods from Table 1 (which only generated point forecasts) into more robust statistical models, which can generate forecast intervals as well as point forecasts. So, each state-space model was labeled as ETS (Error; Trend; Seasonal).

With the transformation of the model, an error e_t with a certain probability distribution is added to the smoothing equation. This change influenced the observation equation, the transition equation, and their respective level, seasonal, and smoothing parameters. The identification of the forecast intervals also changes, depending on whether the errors are additive or multiplicative. In other words, these equations, together with the statistical distribution of errors, shape a complete statistical model.

In order to choose the best model to be used with a certain time series, the R language function ETS ( ) uses the principle of maximum likelihood and the corrected Akaike Information Criterion (AIC_c). Since the likelihood is the chance that something will happen, a good model is associated with a high likelihood. Hence, this strategy was used for the BP Statistical Review of World Energy 2020 data and the best model will be used to make the forecasts.

3 ARIMA models

Unlike exponential smoothing models, which use trend and seasonal components, the ARIMA models use autocorrelation analysis and a unit root test to achieve the best possible model for time series data. "ARIMA” stands for Autoregressive Integrated Moving Average and it uses a parameterization approach based on Box and Jenkins strategy [3]. The method determines the ARIMA orders (p, d, q), where **p** is the order of the autoregressive part, **d** is the degree of first differencing involved and **q** is the order of the moving average part. Equation (1) shows the complete model:

$$y_t' = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t. \tag{1}$$

Where ϕ_p is the autoregressive weights, θ_q is the moving average weights, and ε_t represents white noise. The constant c (where μ is the average value of y'_t) is shown in Equation (2)

$$c = \mu(1 - \phi_1 - \cdots - \phi_p), \tag{2}$$

The Arima ( ) function from the R package fable follows the Box and Jenkins approach. First of all, it runs a Kwiatkowski-Phillips-Schmidt-Shin unit root test (KPSS) to determine the value of d, the degree of first differencing, which converts a non-stationary time series into stationary. After the unit root test, the functions of autocorrelation (ACF) and partial autocorrelation (PACF) are plotted so that they may help determine p and q. In case they are insufficient to identify p and q, the Arima ( ) function uses the lowest Akaike Information Criterion and maximum likelihood estimation method to calculate them. When the parameters are found, it is necessary to check the residuals before being able to make predictions. The residuals of the model must behave like white noise, in other words, all autocorrelations must be within the threshold limits. This way a confidence interval can be achieved, in agreement with Hyndman and Khandakar [4].

The residuals are of great importance both in exponential smoothing and ARIMA models. They must behave like white noise because, according to Neusser [5], “The white noise process is therefore stationary and temporally uncorrelated. As the ACF possesses no structure, it is impossible to draw inferences from past observations to its future development, at least in a least square setting with linear forecasting functions". Because of its stochastic behavior, financial time series must be modeled using methods that consider these characteristics, because, in a certain sense, the process must be in statistical balance to produce prediction intervals. This behavior is emphasized by Cryer and Chan [6].

4 Autoregressive neural network model

The use of lagged values in time series as inputs into a neural network to forecast future values, using a linear autoregressive model, can be a good prediction strategy. Since multilayer neural networks are usually feed-forward, each layer is connected to the previous and the next layer. According to Hyndman and Athanasopoulos [7], “The inputs to each node are combined using a weighted linear combination”. Equation 3 shows the linear combination of the inputs.

$$z_j = b_j + \sum_{i=1}^{n} w_{i,j} x_i \ . \tag{3}$$

Then the result is altered by a nonlinear function, such as a sigmoid, as shown in Equation 4

$$s(z) = \frac{1}{1+e^{-x}} \ . \tag{4}$$

A learning algorithm is used to pick random starting weight values from the neural network framework. At every iteration, the network is trained to produce several different random weight values, and then it generates an averaged value. This process continues until the mean square error cost function is minimized. This procedure tends to make the neural network less sensitive to outliers. And the parameters b_j e $w_{i,j}$ are estimated according to data.

R language function NNAR (p, k) is an autoregressive model where **p** is the maximum number of lags used as inputs and **k** is the number of hidden nodes. According to Hyndman and Athanasopoulos [7], “For non-seasonal time series, the default is the optimal number of lags (according to the AIC) for a linear AR (p) model. If k is not specified, it is set to k = (p+1)/2 (rounded to the nearest integer)”. A neural network works iteratively to make predictions. In other words, historical data is used in one step ahead forecasts. For two steps ahead, historical data and results of the first step forecasts are used. This continues until the prediction horizon is achieved.

In addition to the neural network architecture, which simulates biological neuron networks, the prediction interval is calculated differently in comparison to the other models. ARIMA and exponential smoothing models use residuals to calculate the intervals. Neural networks, according to Hyndman and Athanasopoulos [7], “are not based on a well-defined stochastic model, and so it is not straightforward to derive prediction intervals for the resultant forecasts. However, we can still compute prediction intervals using simulation where future sample paths are generated using bootstrapped residuals".

5 Computer simulations

The methodology of this work consists of the analysis of open data presented in the annual statistical report called BP Statistical Review of World Energy 2020, released by British Petroleum. The time series used for prediction represents crude oil prices from 1861 to 2019, as shown in Figure 1.

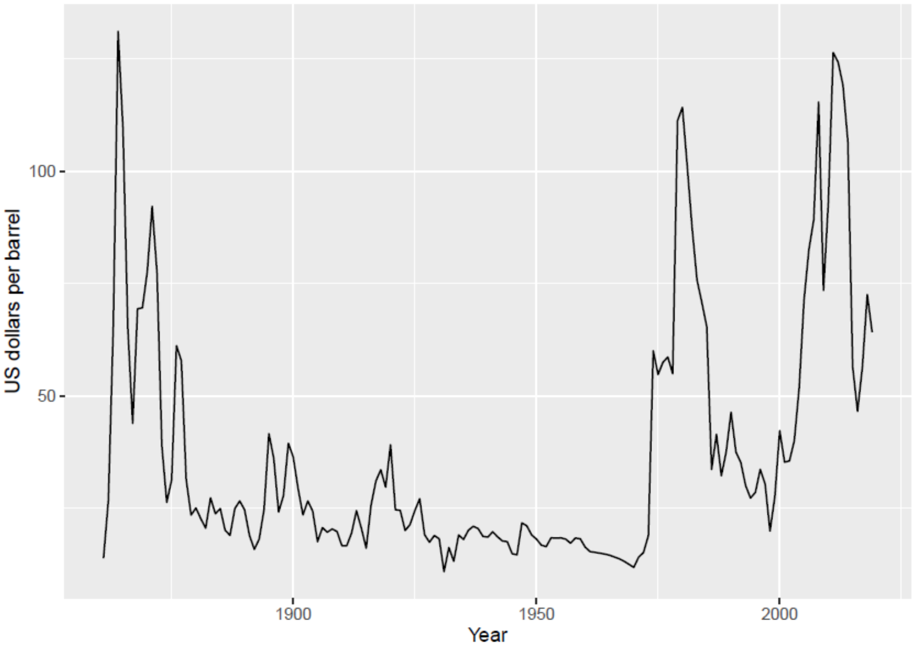

Figure 1. Crude oil prices (1861 – 2019)

The first prediction was made using an exponential smoothing method and the function ETS ( ) from R package fable. The proposed model was an ETS (M, A, N) with AIC$_c$ value of 1525.149 and smoothing parameters: $\alpha = 0.999899$ and $\beta = 0.09173289$. The estimated initial states were: l = - 9.960107, b = 23.31603 and $\sigma^2 = 0.0942$. Point forecasts (blue line) and their respective prediction intervals of 80% (dark blue) and of 95% (light blue) are shown in Figure 2.

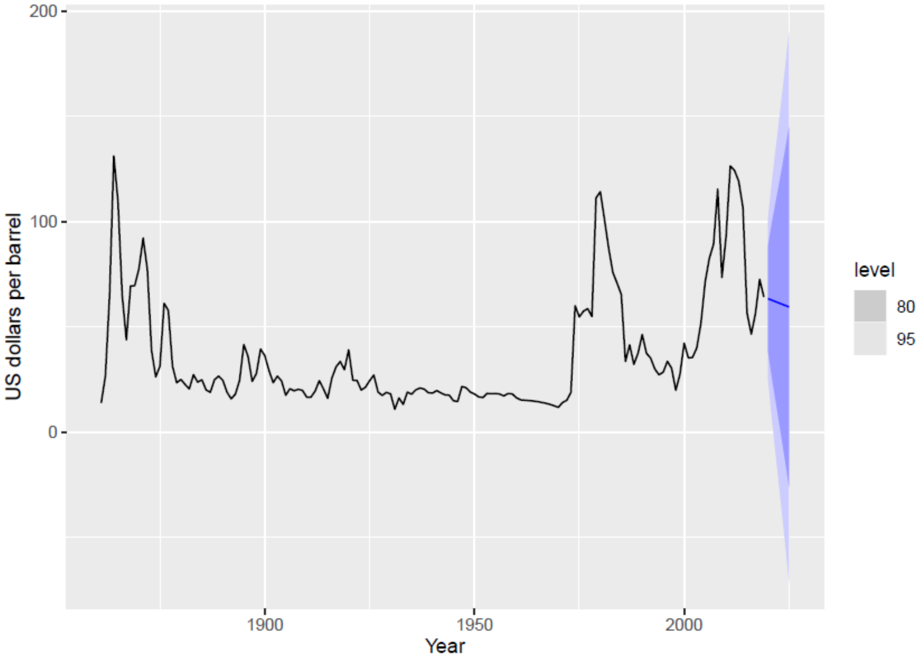

Figure 2. Oil-price forecasting based on exponential smoothing (2020-2025)

The second prediction uses the ARIMA model, as well as another function from R package fable called Arima ( ), in order to adjust its parameters. The proposed model was an ARIMA (3, 1, 3) with AIC$_c$ value of 1260.89 and $\sigma^2 = 159.5$. The residual plot may be seen in Figure 3 and it shows white noise characteristics.

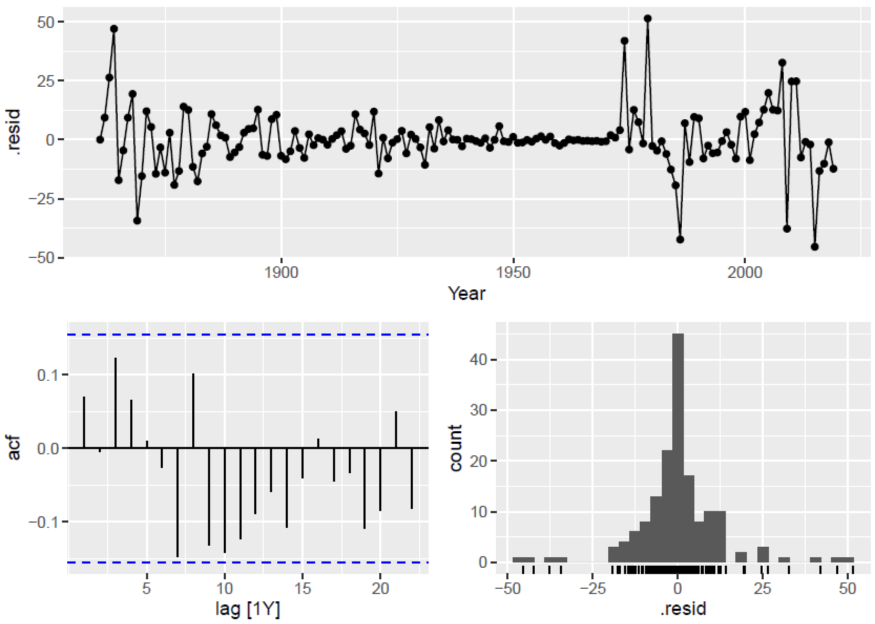

Figure 3. Time series plot of the residuals, ACF and histogram

Once more, the point forecasts (blue line) and their respective prediction intervals of 80% (dark blue) and 95% (light blue) from the ARIMA (3, 1, 3) model may be seen in Figure 4.

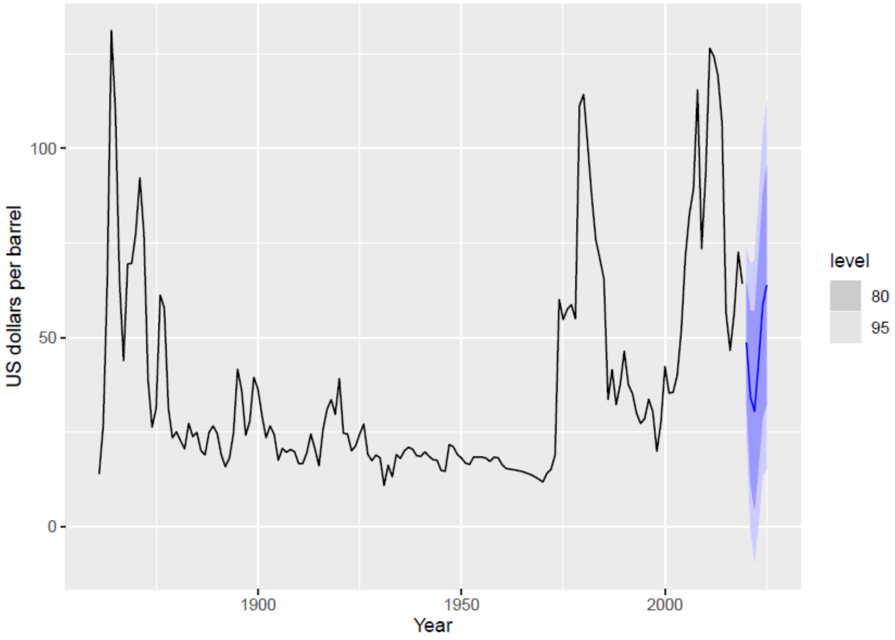

Figure 4. Oil-price forecasting based on ARIMA model (2020-2025)

The last prediction uses an autoregressive neural network model and another function called NNETAR ( ) from the R package fable to adjust its parameters. The proposed model was a NNAR (10, 6) and σ^2 = 14.92. Point forecasts (blue line) and their respective prediction intervals of 80% (dark blue) and 95% (light blue) from NNAR (10, 6) model may be seen in Figure 5.

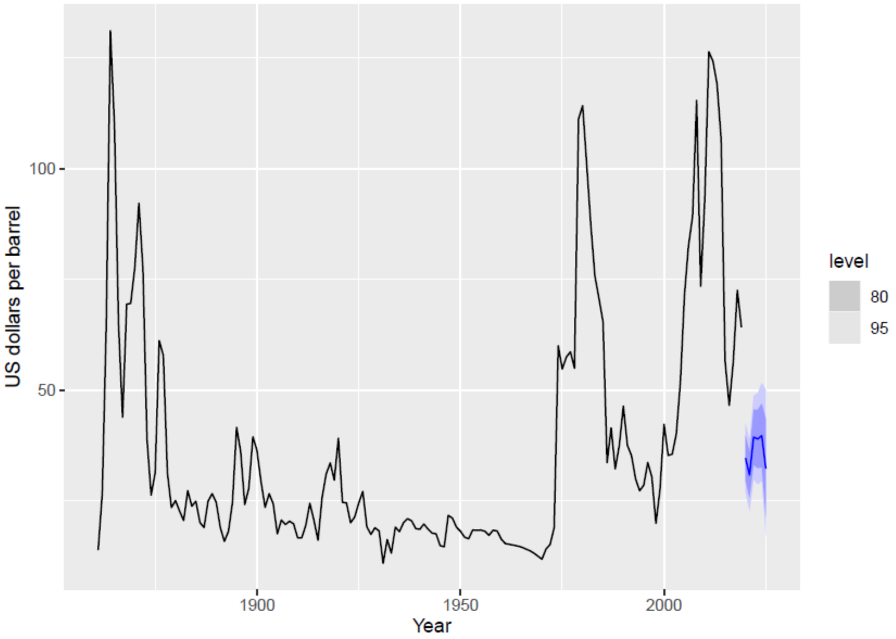

Figure 5. Oil-price forecasting based on autoregressive neural network model (2020-2025)

6 Conclusions

It is remarkable that the average oil price forecast for the period of 2020 to 2025 presents lower values than in 2019. This reduction is less pronounced in the exponential smoothing model, which produces point forecasts of 63 US dollars per barrel in 2020 and 59 US dollars per barrel in 2025. The price values behave like a decreasing line, and the confidence intervals expand as the forecast horizon increases. In this model, the confidence intervals are wider than ARIMA and autoregressive neural network models, confirming that there is a greater uncertainty level.

The ARIMA model also presents lower average oil prices compared to 2019, however, this reduction is more pronounced in 2020 (48 US dollars per barrel), reaching its smallest average value in 2022 (31 US dollars per barrel), and escalating again only in 2023 (44 US dollars per barrel). The higher average value predicted occurs in 2025 (64 US dollars per barrel) with a more accurate forecast interval, thus indicating a smaller degree of uncertainty in comparison to the exponential smoothing method.

When compared to the year 2019, the autoregressive neural network model was the model that indicated the most significant oil price reduction. The average point forecast values are estimated to be between 31 and 39 US dollars per barrel. The point forecast is 35 US dollars per barrel in 2020, 39 US dollars between 2022 and 2024, and 33 US dollars in 2025. The level of uncertainty is the lowest of them all.

It is common to combine both ARIMA and exponential smoothing models to make forecast scenarios. This work demonstrates that all forecasting models showed a decline in crude oil price scenarios for the next few years.

Acknowledgements. First, I would like to thank all co-authors for helping me on my project. Also, I would like to thank the Programa Multidisciplinar de Formação de Recursos Humanos em áreas de interesse do setor do petróleo (PRH 54.1-ANP) for the opportunity to carry out my research on oil and gas sector.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] B. Looney et al, Statistical Review of World Energy 2020. *BP Statistical Review*, 69th edition, 2020.
[2] P. S. Cowpertwait and A. V. Metcalfe, *Introductory Time Series with R*. Springer Science + Business Media. New York: LLC, 2009.
[3] G. E. P. Box, G. M. Jenkins, G. C. Reinsel and G. M. Ljung. *Time series analysis: Forecasting and control*. 5th edition. Hoboken, New Jersey: John Wiley & Sons, 2015.
[4] R. J. Hyndman and Y. Khandakar. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 27(1), 2008.
[5] K. Neusser. *Time Series Econometrics*. Springer Texts in Business and Economics. Bern: Springer International Publishing Switzerland, 2016.
[6] J. D. Cryer and K. S. Chan. *Time Series Analysis with Applications in R*. 2nd edition. Springer Science + Business Media. [S.l.]: LLC, 2008.
[7] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. 3rd edition, OTexts: Melbourne, Australia, 2021.