ARIMA MODEL FOR DAY-AHEAD ELECTRICITY MARKET PRICE FORECASTING

1. INTRODUCTION

Electricity is an energy commodity that is much different from other commodities such as oil, natural gas, coal, etc. because it is not physically storable locally in large quantities. Storing electricity at a grid-scale is the desired grid characteristic; however, it is not widely used because it is not economically feasible and uncompetitive at the moment . Demand response or demand-side management techniques have been important tools in improving the voltage profile, system efficiency, and stability, and for matching the stochastic output power of renewable sources [2]. A majority of demand response programs are based on electricity price signals as forecasts may improve system stability, as prices become more volatile, the balance of the grid is compromised but not only, the importance of the electricity price forecasting may have wide usage. Electricity price forecasting (EPF) literature started to develop at the beginning of the 2000s , the main methods of electricity price forecasting may be divided into five groups: multi-agent, fundamental, reduced-form, statistical, and computational intelligence models. EPF may support decisions for new power plans implementation and reliable cost-based analysis may secure a clear vision of the future for long-term corporative and national energy strategies development. In addition, a forecast may secure bank financing for new projects, help traders to hedge portfolio risk of market prices, and allow big industrial consumers to plan properly their costs. However, decisions on increased RES production investigations may push the prices down because of the sufficient production, but on the other side, it should be considered that low prices in the long term are not sustainable, because they will make the generation not profitable. Electricity markets in the EU became completely liberalized in the first decade of the 2000s and different market participants such as producers, traders, and system operators (TSOs and DSOs) took part in the different types of organized market places. Different wholesale markets are organized on different time horizons in advance of the actual moment of production and consumption. In the long term, producers and consumers can trade large blocks of power in the futures or forward markets years before actual delivery. On the other hand, spot markets allow the trades same‐day hours ahead or next-day delivery. The most important market segment is the day-ahead market, which usually closes at noon for next-day delivery. Compared to the forward markets, a more precise estimate of the demand for the day ahead is possible based on weather data, wind, water, and sun conditions, actual and forecast load, events that might influence demand, planned and unplanned plant outages, and prices of CO2 emissions, electricity futures, Brent crude oil, goal futures, and gas futures. Therefore, this paper are used several approaches to analyse the Bulgarian IBEX hourly electricity price dynamics in the day-ahead market

2. PROPOSED METHODOLOGY

It can be considered that time series modelling is different from more traditional classification and regression predictive modelling approaches. The temporal nature adds order to the observations. This imposed order means that important assumptions about the consistency of those observations need to be handled specifically. In the domain of machine learning [7], a specific collection of methods and techniques may be used particularly well suited for predicting the value of a dependent variable according to time. This paper will be covered one domain of machine learning, which contains a specific collection of methods and techniques particularly well suited for predicting the value of a dependent variable according to time Auto‐Regressive Integrated Moving Average (ARIMA) [4].

The general process for ARIMA modelling includes visualization and aggregation of the time series data. First, the data to be used for forecasting will be using statistical tools, and patterns for seasonality will be checked using autocorrelation analysis. The data will be decomposed and made stationary if needed and further analysis will be done on patterns and residuals. For variable analysis correlation,

scatterplots will be used as the main tools. Each variable and its possible effects on the data will be discussed briefly. Time series will be analysed for three components: trend, seasonality, and noise. The modelling input-process-output approach that is used describing the structure of an information processing defining both inputs and outputs as one united mechanism and part of the system modelling process is shown in fig. 1.

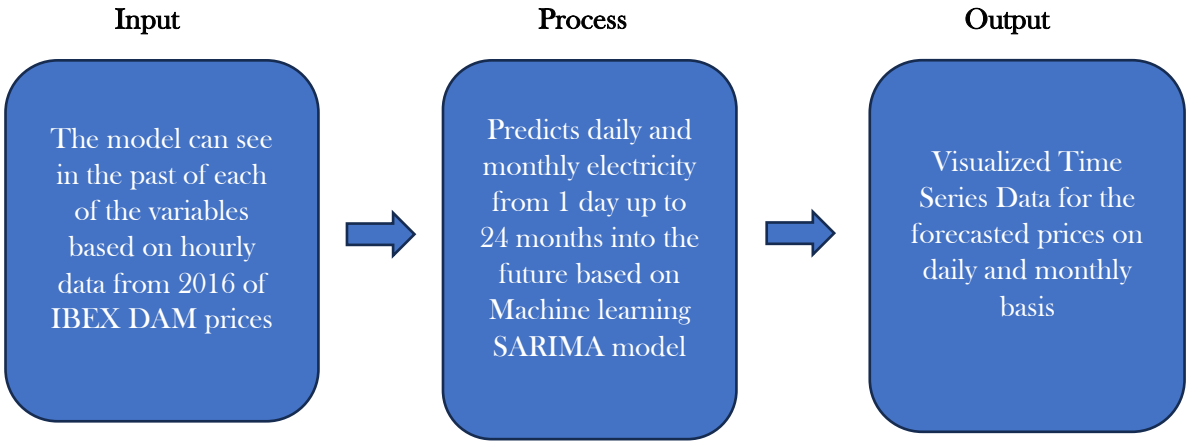

Fig. 1. Systemic input-process-output modelling approach process

As it is widely known, time-series data may be characterized as stationary and non-stationary time series data and may also be differentiated by different patterns as dependencies on time, and whether they are showing trend or seasonal effects and their consistency over time, which may be measured with the mean or the variance of the observations. Statistical modelling

methods to be effective, require the time series to be stationary. Therefore, non-stationary time series data will be made stationary by identifying and removing trends and removing seasonal effects. After collecting the data, understanding the data better is required preliminary data analysis. This will enable the process of finding good models that fit the used data well. For the data analysis, tools such as correlation graphs, autocorrelation functions, comparative variable analysis, decomposition, and residual analysis will be used. Autoregressive (AR) models investigate if past values affect current values. AR models are used in time-varying processes. Therefore, a linear regression model may be built that attempts to predict the value of dependent variable days and months ahead, given the values it had on previous days and months as p is a parameter of how many lagged observations to be taken into consideration, as it is shown by the following formula (1):

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{1}$$

Differencing is a transformation applied to time-series data to make time-series data stationary. This allows the properties not to depend on the time of observation, eliminating trend and seasonality and stabilizing the mean of the time series. Differencing will be done a couple of times if needed until data becomes stationary (2).

$$y_t^* = y_t' - y_{t-1}' = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2} \tag{2}$$

Moving Average Model (MA) assumes the value of the dependent variable on the current day depends on the previous day's error terms, as the formula shows (3):

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{3}$$

where μ is the mean of the series, the θ1, ..., θq are the parameters of the model, and the εt, εt−1,..., εt−q are white noise error terms. The value of q is the order of the MA model. ARIMA Box-Jenkins model adds differencing to an ARMA model [5]. Differencing subtracts the current value from the previous and can be used to transform a time series into stationary data (4).

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} \ldots - \theta_q e_{t-q} \tag{4}$$

Where p is the number of autoregressive terms (AR order), d is the number of nonseasonal differences (differencing order) and q is the number of moving-average terms (MA order). If forecasting results of the predicted data are not good using ARIMA, Seasonal ARIMA(SARIMA) will be applied. SARIMA model is used when the time series exhibits seasonality and is similar to ARIMA models, but there are a few parameters that need to be added to account for the seasons. The general form of the seasonal model SARIMA is given by (5):

$$\text{ARIMA}\,(p,d,q) \times (P,D,Q)S \tag{5}$$

with p = non-seasonal AR order, d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern.

3. DATA ANALYSIS

The forecasting range will be 24h with an hourly frequency. The baseline model will use historical prices for future values prediction. The data that will be used form is the public source: IBEX, https://ibex.bg/en/ containing real hourly data set from 20.02.2016 to 05.03.2022 Bulgarian day-ahead market hourly prices as shown in fig. 2.

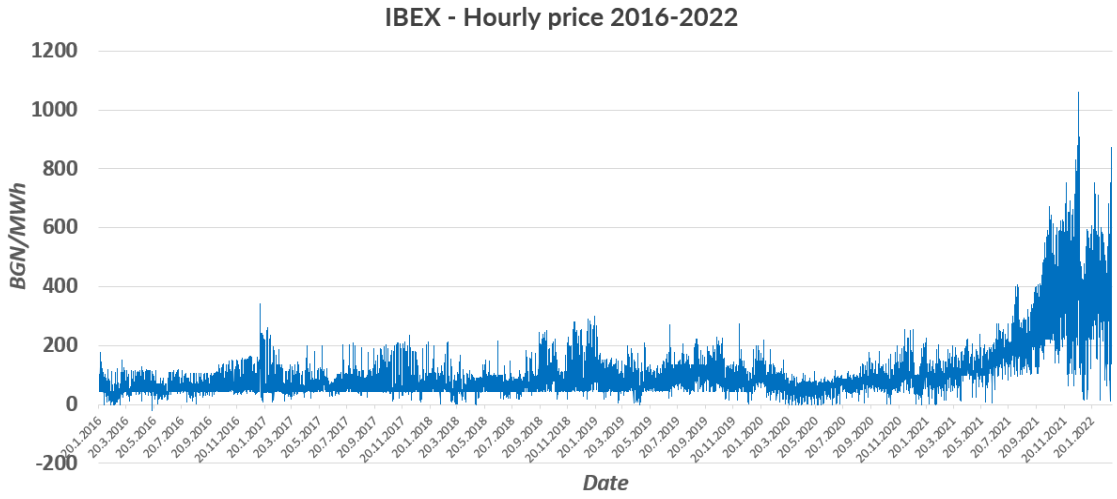

Fig. 2. Hourly means

Data will be grouped on an hourly, daily, and monthly basis and then aggregated as electricity prices follow patterns over daily, weekly and seasonal timeframes as shown in fig.3. The main goal is to predict the average daily electricity price at the Bulgarian Power Exchange for days and months ahead.

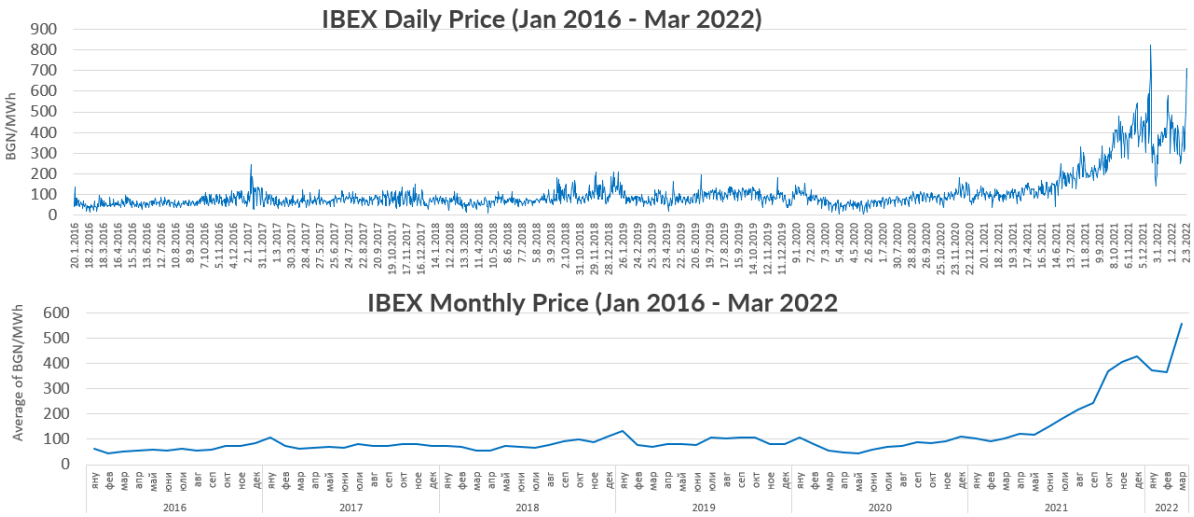

Fig. 3. Daily and monthly means

Expanding forecast means that after forecasting the next 24 h, that raw data is included in our training period for forecasting the day after. For the evaluation of the forecasts, scaled error models will be preferred over percentage error models. After comparing the error values and evaluating the forecasts, ways to improve the forecasting performance of the models will be investigated using backward feature elimination. This model will provide the removal of variables that are irrelevant to the forecasts. An attempt at combining the forecasts will be made to gather more accurate forecasts will be realized.

The data has been loaded into Data Frame, cleaned up, converted hourly data into daily and monthly data, visualized, and plotted the time series data to detect patterns. From fig. 4, it can be seen, that there is a trend in time that suggests that the data are not stationary. To be sure we will use an Augmented Dickey-Fuller test.

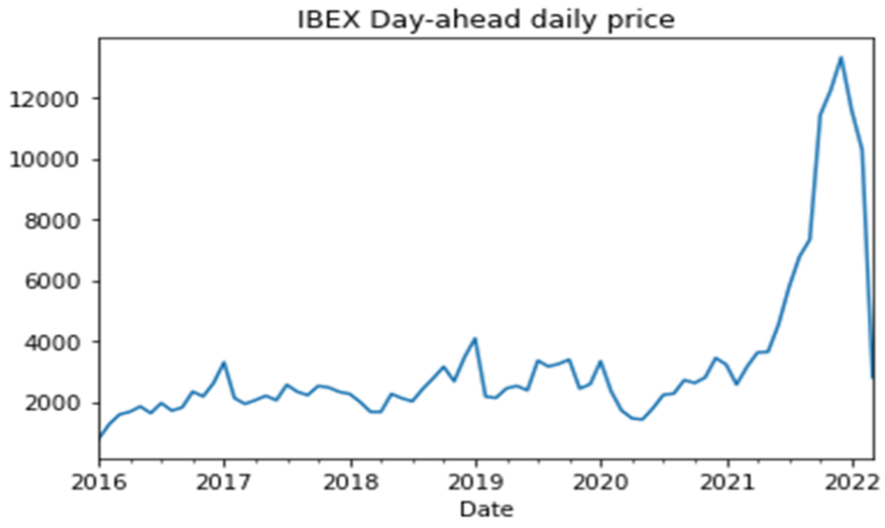

Fig. 4. Monthly means and rolling mean and standard deviation.

3.1. Autocorrelation of the Time Series

To further analysis the data and confirm the daily and monthly seasonality observed in the previous section autocorrelations functions can be inspected. Autocorrelation is the correlation, the linear relationship, between the function and the delayed version of the function. Before the model is built, the data need to be checked if the time series is stationary. There are two primary ways to determine whether a given time series is stationary rolling statistics and augmented Dickey-Fuller Test. Plot The rolling mean and rolling standard deviation is plotted. The time series is stationary if they remain constant with time. The time series from Dickey Fuller Test is considered stationary if the p-value is lower than 0,05. Stationary data means data that has no trend concerning time. Inspecting fig. 5 and taking the mean of the residuals, it is seen that there is no bias and the mean is almost zero. The residuals follow a normal distribution however they are very high at some points and seem to have some correlation. Some explanations for this could be that the initial data have a lot of spikes, there are holiday effects that are not included in this model, and other external effects such as excess electricity production that lead to jumps.

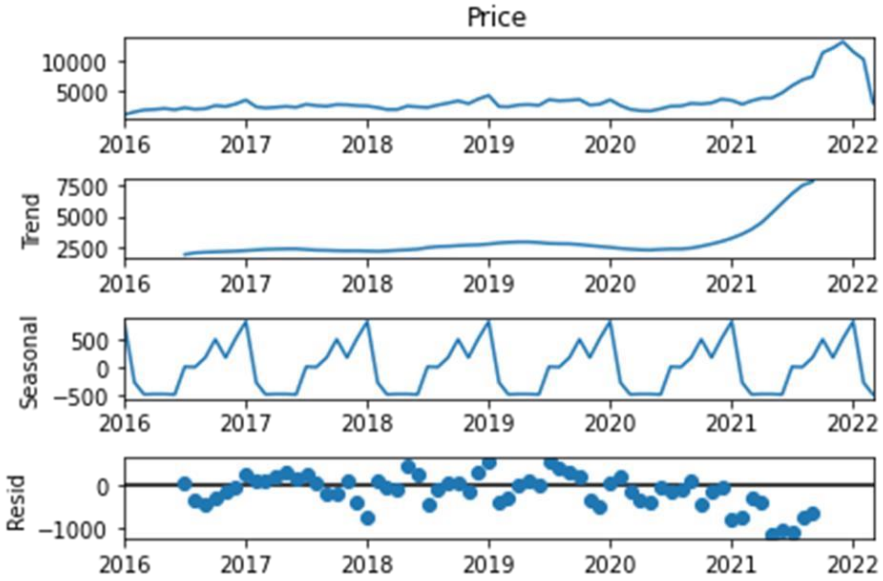

Fig. 5. Decomposition of data

Below in table 1 are showed the results from the Augmented Dickey-Fuller Test to test whether the series is stationary or not post Differencing.

Parameter	Augmented Dickey-Fuller Test Results:
ADF Test Statistic	1.850218578480598
p-value	0.9984435851817096
#Lags Used:	27
Number of Observations Used:	2209

Table 1. Augmented Dickey-Fuller Test results

From the data above, can be considered that there is weak evidence against the null hypothesis, the time series has a unit root, indicating it is non-stationary.

4. FORECASTING MODELS RESULTS

4.1. Daily data analysis, differencing, autocorrelation, error check, and results

To make the data stationary, it will be differentiated. Differencing in statistics is a transformation applied to time-series data to make it stationary. This allows the properties not to depend on the

time of observation, eliminating trend and seasonality and stabilizing the mean of the time series. Seasonal First Difference was made and the results are shown in fig. 6. Results from the second Augmented Dickey-Fuller Test to test whether the series is stationary or not post Differencing are shown in table 2. The results now are showing that data has no unit root and is stationary.

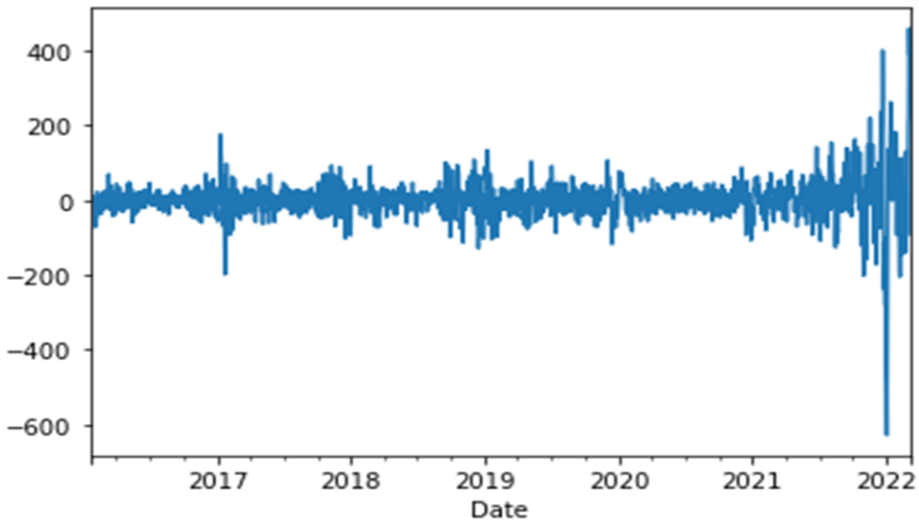

Fig.6. Seasonal First Difference daily data

Parameter	Augmented Dickey-Fuller Test Results:
ADF Test Statistic	-9.491961306426585
p-value	3.632537140497319e-16
#Lags Used:	27
Number of Observations Used:	2209

Table 2. Augmented Dickey-Fuller Test results.

Identification of an AR model is done with the PACF. The identification of an MA model is often best done with the ACF rather than the PACF. P, D, Q where P is AR model lags, D is differencing and Q shows the MA lags. For non-seasonal data P=1 (AR specification), D=1 (Integration order), Q=0 or 1 (MA specification/polynomial). In the graphs below (fig. 7), each spike (lag) that is above the dashed area considers being statistically significant.

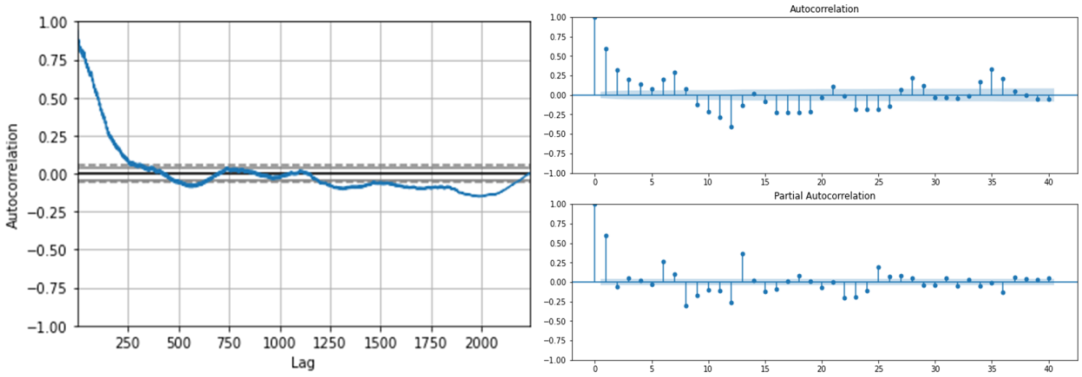

Fig. 7. Differencing and Autocorrelation

Figure 8 shows the ARIMA forecasts against actual data for the complete forecasting period. It can be seen that the forecasting is not good using the ARIMA model, since the time series exhibits seasonality. Therefore, we will implement Seasonal-ARIMA

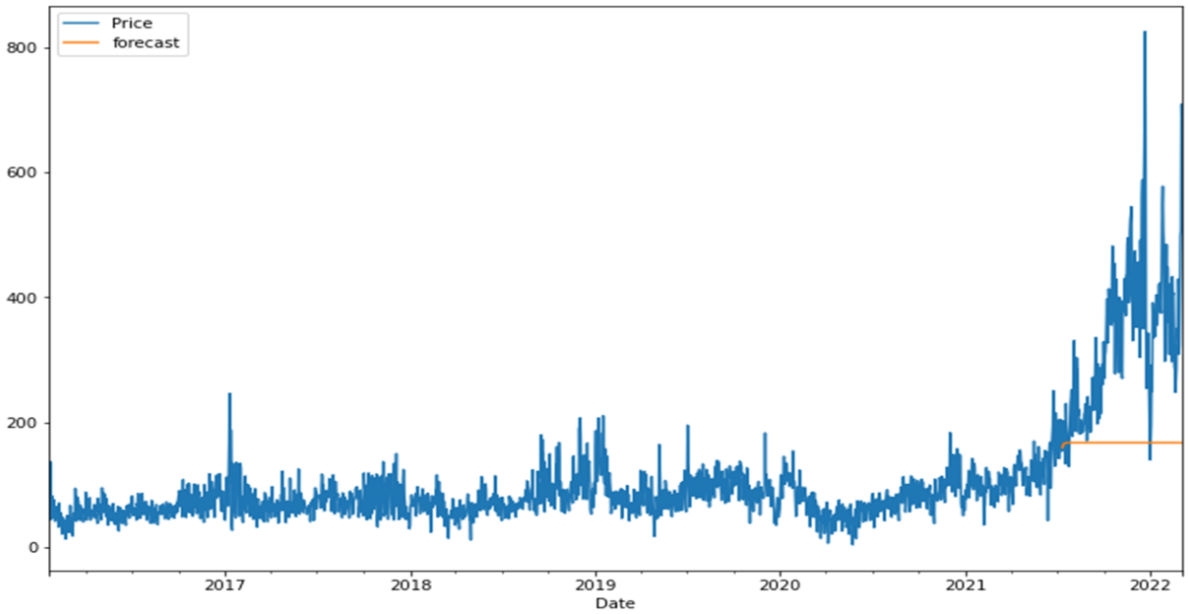

Fig. 8. Applied ARIMA model for daily predictions showing not good results

SARIMA model is used when the time series exhibits seasonality. This model is similar to ARIMA models [6], we just have to add a few parameters to account for the seasons. SARIMA is written as ARIMA(p, d, q)x(P, D, Q)m. Seasonal differencing takes into account the seasons and differences between the current value and its value in the previous season eg: The difference for the daily data would be the monthly value by the number of the days in the month. We can see here that the residuals are white noise and they are also normally distributed. We can thus go ahead with this model. Therefore, we will now predict the test data and then check for the accuracy of the real data with the following parameters Order=(1, 1, 1), seasonal order=(1,1,1,31))

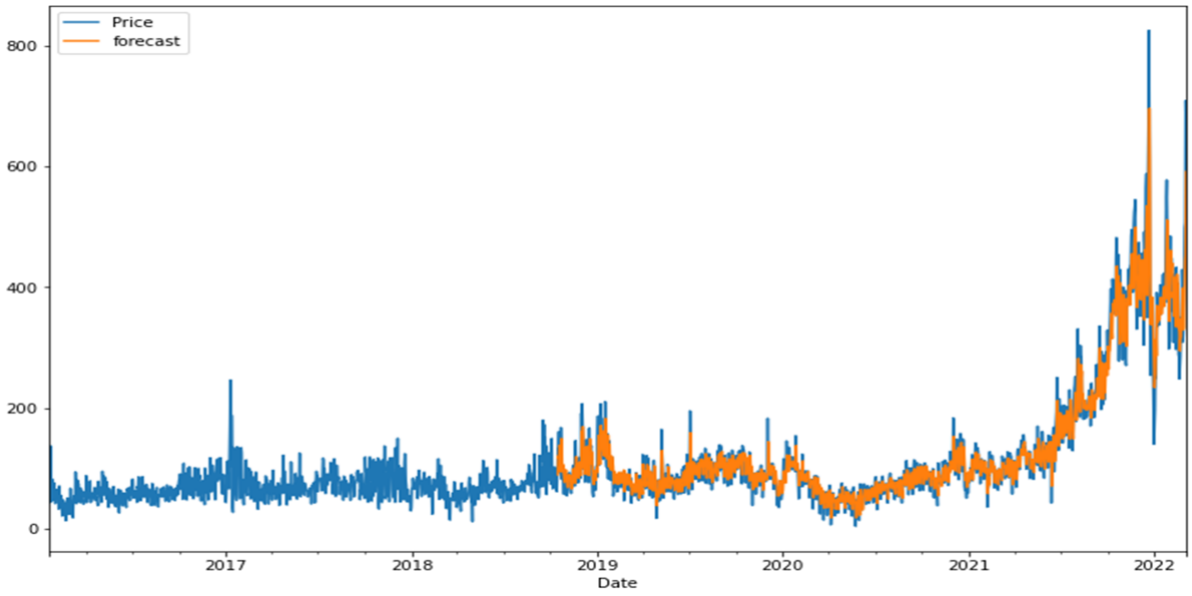

Fig. 9. Applied SARIMA model for daily predictions, error check between real price data and price forecasted data

We can see that the prediction closely follows the actual value as from 2237 days we start from a value of 1000 and end at a value of 2300. Figure 9 are compared these values with the average value of the test series to check if the magnitude of the error is acceptable. It can be seen that

forecasting using SARIMA gave good results since the data exhibited seasonality. Therefore, we may continue with SARIMA Time Series Forecasting

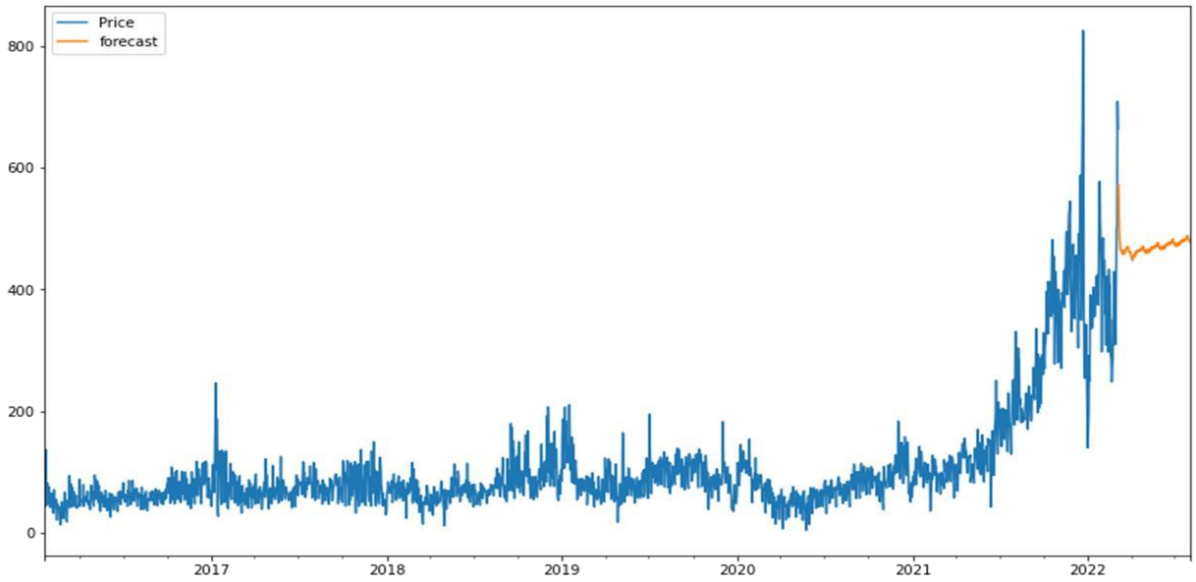

Fig. 10. Daily Data forecasting results with SARIMA for 150 days ahead, historical price, and future.

4.2. Monthly data analysis, differencing, autocorrelation, error check, and results

Seasonal first and second differences and autocorrelation were made and the results are shown in fig. 10 and table 3. Results from the third Augmented Dickey-Fuller Test to test whether the series is stationary or not post Differencing are shown in table 3

Parameter	Augmented Dickey-Fuller Test Results:
ADF Test Statistic	0.20509122633685525
p-value	0.9725457039909695
#Lags Used:	7
Number of Observations Used:	55

Table 3. Augmented second Dickey-Fuller Test results

The second ADF Test Statistic shows weak evidence against the null hypothesis, the time series has a unit root, indicating it is non-stationary. Therefore, second differentiation is made. The results now are showing that data has no unit root and is stationary. The seasonal second difference was made and the results are shown in fig. 11

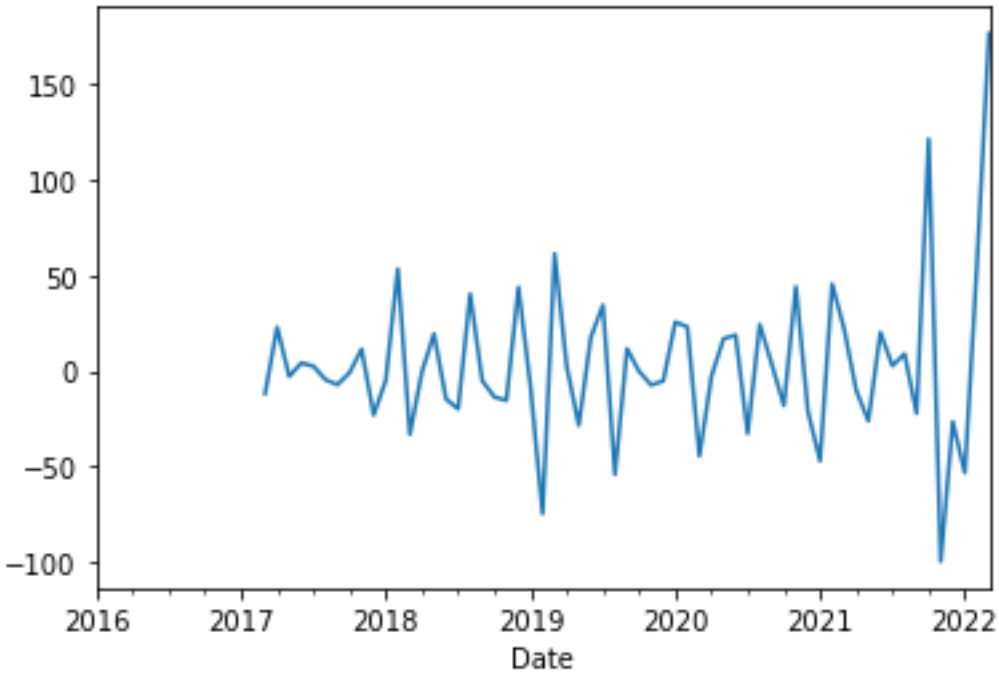

Fig. 11. Seasonal second difference monthly data

Parameter	Augmented Dickey-Fuller Test Results:
ADF Test Statistic	-7.3183330481968865
p-value	-7.3183330481968865
#Lags Used:	5
Number of Observations Used:	55

Table 4. Augmented third Dickey-Fuller Test results

Identification of an AR model is done with the PACF. For non-seasonal data P=1 (AR specification), D=1 (Integration order), Q=0 or 1 (MA specification/polynomial). In the graphs below (fig. 12), each spike(lag) that is above the dashed area considers being statistically significant

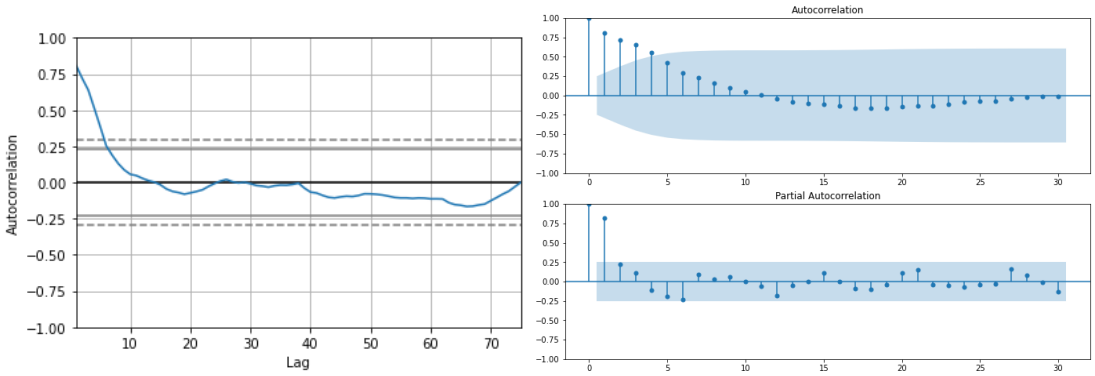

Fig. 12. Differencing and Autocorrelation on monthly data

Figure 13 shows the ARIMA forecasts against actual data for the complete forecasting period. It can be seen that the forecasting is not good using the ARIMA model, since the time series exhibits

seasonality. Therefore, we will implement Seasonal-ARIMA

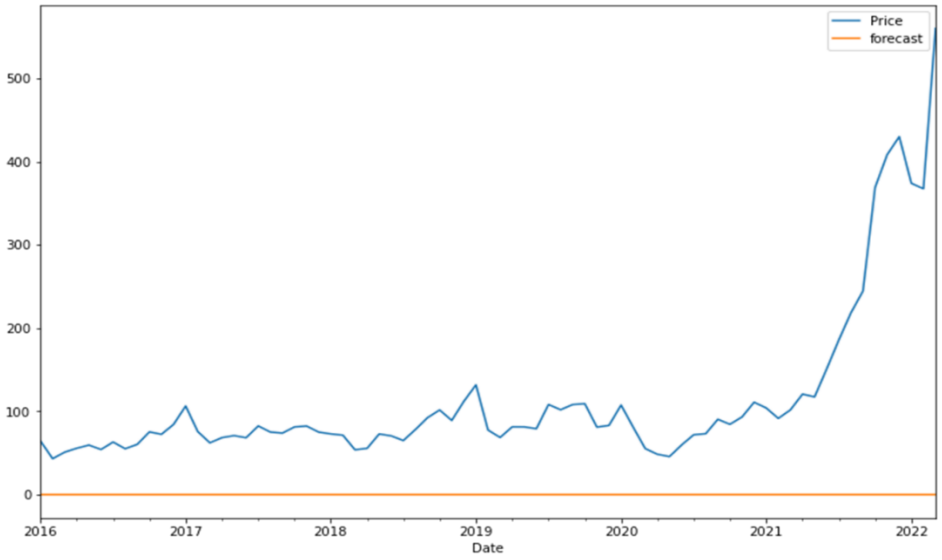

Fig. 13. Applied ARIMA model for monthly predictions showing not good results

SARIMA model is applied and shown in fig. 14 to predict values for the forecast data and then check for the accuracy with the real data with the following parameters: order=(1, 1, 1), seasonal order=(1,1,1,12)) for ARIMA(p, d ,q)x(P, D, Q)m values

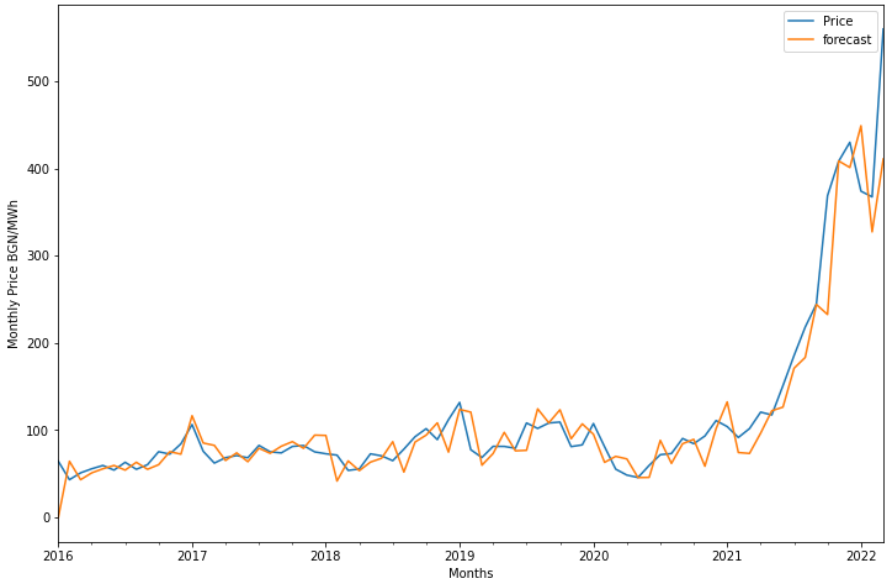

Fig. 14. Applied SARIMA model for monthly predictions, error check between real price data and price forecasted data

Figure 15 shows monthly data for forecasted price results with SARIMA ahead based on historical prices and real historical prices. We can see that the forecasted price closely follows the actual value. Figure 14 compares these values with the average value of the test series to check if the magnitude of the error is acceptable. It can be seen that forecasting using SARIMA gave good results since the data exhibited seasonality. Therefore, we may continue with SARIMA Time Series Forecasting

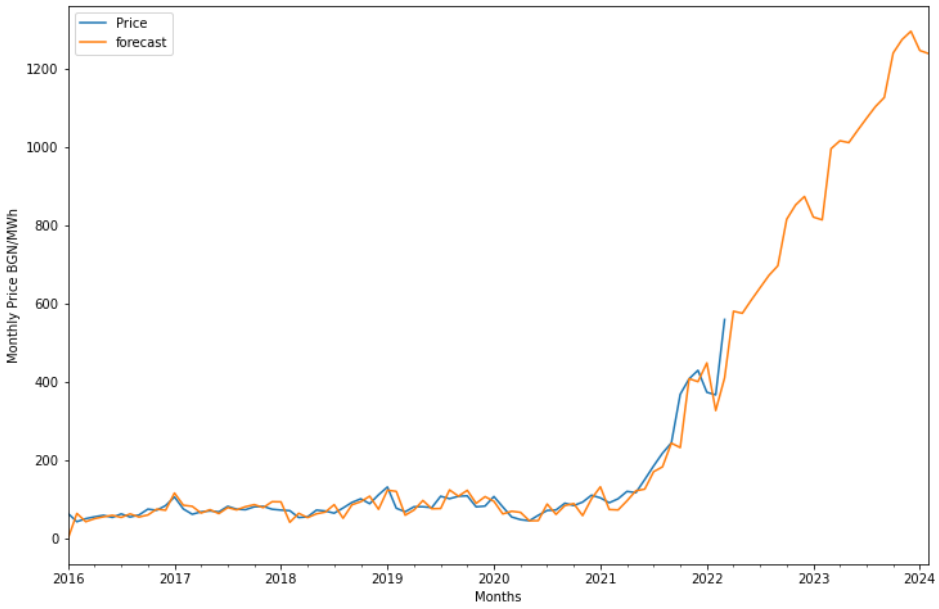

Fig. 15. Monthly data forecasted results with SARIMA for 60 months ahead, historical price, and future predictions

CONCLUSIONS: -

The ability to make predictions based on historical observations creates a competitive advantage considering the complexity of the power sector nature and the need for a systemic approach in power sector decision-making. Auto-Regressive Integrated Moving Average is a domain of machine learning and may be used as a well-suited method and technique for predicting the value of a dependent variable according to time. Observations from a non‐stationary time series show seasonal effects, trends, and other structures that depend on the time index. Forecasting results for electricity prices are not good using ARIMA, since the time series exhibits seasonality. ARIMA method together with the discrete wavelet transform method may be more suitable in other files as well as predicting future electrocardiographic (ECG) [8, 9] or photo plethysmo graphic (PPG) signals [10, 11] from previous ECG for improving the accuracy of prediction. For electricity prices, results show that forecasting using SARIMA gave good results since the data exhibited seasonality. Thereafter, it may be considered that identified and implemented SARIMA is a suitable forecasting method for the volatile nature of electricity prices. The error evaluation method is determined depending on the data properties, and individual forecasting methods are therefore compared. To improve the forecasts in the future, a combination of Trigonometric Seasonal Box-Cox Transformation and Artificial Neural Networks (ANN) methods may be used for seasonal naïve forecasts.