Date of publication JUN-30, 2025, date of current version MAR-05, 2025.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.2.4021

Building an ARIMA Model for Predicting Time Series in Python

GAUKHAR ABDENOVA[1], ZHANAT KENZHEBAYEVA[2], HANNA MARTYNIUK[3]

[1]Department of Mathematical and Computer Modeling, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Department of Computer Science, The Caspian University of Technology and Engineering named after Sh.Yessenov JCS, Aktau, Kazakhstan
[3]Department of System Analysis and Information Technologies, Mariupol State University, Kyiv, Ukraine

Corresponding author: Hanna Martyniuk (e-mail: ganna.martyniuk@gmail.com).

ABSTRACT This study showcases the practical application of the Box-Jenkins model, specifically ARIMA, to predict forthcoming values of a short-term economic indicator in the Republic of Kazakhstan. Data extracted from the Bureau of National Statistics website, covering the period from January 2009 to July 2021, served as the foundation for this analysis. Leveraging the Python programming language, the authors constructed the ARIMA model and conducted thorough time series analysis to uncover temporal patterns within the data. Validation of the model's performance was carried out using data from August 2021 to July 2022. The article presents a comprehensive methodology for model development, encompassing data preprocessing, parameter estimation, and model evaluation stages. Emphasis is placed on the necessity of regular data updates to uphold the accuracy of forecasts, underscoring the practical significance of this study within the domain of time series modeling and forecasting methodologies. As a result, using the constructed model, future values of the series were obtained and a comparison of the predicted values with real data was carried out. To check the error, the mean absolute error in percent (MAPE) was calculated, which was 7.2%. Checking the residual errors showed that the residuals have a normal distribution. This research contributes valuable insights into the application of advanced statistical techniques for economic forecasting, particularly in dynamically evolving contexts like Kazakhstan's economy.

KEYWORDS non-stationary time series; seasonal data; short-term forecasts; ARIMA.

I. INTRODUCTION

A forecast is a prediction of some future event or events. Forecasting is an important problem that covers many areas [6-9], including business and industry, public administration, economics, environmental sciences, medicine, social sciences, politics and finance. Most forecasting tasks involve the use of time series data [5, 11-14].

A time series is a time-oriented or chronological sequence of observations of a variable of interest. The velocity variable is collected at regular intervals, as is usually the case in most time series and forecasting applications. Many business forecasting applications use daily, weekly, monthly, quarterly or annual data. For example, the total sales volume of a product during a month; or it could be statistics that somehow reflect the activity of a variable over a period of time, for example, the daily closing price of a particular stock on a stock exchange. The reason forecasting is so important is that forecasting future events is an important contribution to many types of planning and decision-making processes.

II. STATE-OF-THE-ART

ARIMA is a type of model known as the Box-Jenkins method. Short for "Autoregressive Integrated Moving Average", is a forecasting algorithm based on the idea that information in past values of a time series can only be used to predict future values. The ARIMA model is becoming a popular tool that data scientists use to predict future demand [1, 7, 10, 16, 26]. The Box-Jenkins approach starts with the assumption that the process that generated the time series can be approximated using the ARIMA model. The process of forecasting according to the Box - Jenkins method consists of the following three stages:

1. Identification: Evaluate whether the time series is stationary, and if not, how many differences are required to make it stationary. Determine the parameters of the AR and MA model for the time series.
2. Estimation: Using this data to train model parameters (i.e., coefficients).
3. Diagnostic verification: Evaluation of the selected model in the context of available data and verification of the area where the model can be improved.

Thus, ARIMA is a powerful forecasting tool that enables time series analysis and supports informed decision-making based on its predictions.

III. STATIONARY TIME SERIES

Definition. A time series is called stationary (in a broad sense) if

1. $E(x_t)$=const (the average is constant over time);
2. $cov(x_t, x_{t+h}) = \gamma(h)$ (the covariance depends only on the lag h).

The concept of a stationary time series means that its average value does not change over time, i.e., the time series has no trend. In addition, the covariance between different elements of a time series (as between random variables) depends only on how far they are from each other in time. The value h, which characterizes the time difference between the elements of the time series, is called a lag variable or lag. Since [4]

$$\gamma(0) = cov(x_t, x_t) = Var(x_t), \quad (1)$$

then the variance of the stationary time series also does not change with time.

Let us consider the main models of stationary time series:

- Moving average model MA (Moving Average);
- Auto regression model AR (Auto Regression);
- General mixed ARMA auto regression-moving average model.

For the convenience of representing various models, the (formal) lag operator L is often used:

$$L(x_t) \stackrel{\text{def}}{=} x_{t-1}. \quad (2)$$

$$L^2(x_t) = L\big(L(x_t)\big) = L(x_{t-1}) = x_{t-2}. \quad (3)$$

Therefore,

$$L^k(x_t) = x_{t-k}, \quad (4)$$

and formally put $L^0(x_t) = x_t$.

Let us start right away with the general view of the ARMA (p, q) model

$$\begin{aligned} x_t &= \mu \\ &+ \sum_{j=1}^{p} \phi_j x_{t-j} + u_t \\ &+ \sum_{s=1}^{q} \theta_s u_{t-s} \quad u_t \sim WN(0, \sigma_u^2), \\ &\phi_p, \theta_q \neq 0 \end{aligned} \quad (5)$$

We write using the lag operator L:

$$x_t = \mu + \sum_{j=1}^{p} \phi_j L^j x_t + u_t + \sum_{s=1}^{q} \theta_s L^s u_t. \quad (6)$$

$$\left(1 - \sum_{j=1}^{p} \phi_j L^j\right) x_t = \mu + \left(1 + \sum_{s=1}^{q} \theta_s L^s\right) u_t \quad (7)$$

Now we introduce two polynomials of degree p and q:

$$\phi(z) = 1 - \sum_{j=1}^{p} \phi_p z^p; \quad (8)$$

$$\theta(z) = 1 + \sum_{s=1}^{q} \theta_s z^s = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q. \quad (9)$$

Then the model can be formally written as

$$\phi(\mathrm{L})x_t = \mu + \theta(\mathrm{L})u_t. \quad (10)$$

$\phi(\mathrm{L})x_t$ is called the autoregressive part of the ARMA model, $\theta(\mathrm{L})u_t$ – part of the moving average. The ARMA model defines a stationary series ⟸⟹ the stationarity condition is met: all the roots of the autoregressive polynomial

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \quad (11)$$

Let us consider special cases of the general ARMA model:

- MA (q) = ARMA (0, q) – moving average model;
- AR (p) = ARMA (p, 0) – the auto regression model.

Model MA (q) (q – lag order):

$$x_t = \mu + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}, \quad u_t \sim WN(0, \sigma_u^2), \theta_q \neq 0. \quad (12)$$

The time series takes into account only external shocks up to the order of q. The model sets a stationary series x_t under any circumstances $\{\theta_j\}_1^q$, because φ(z) ≡ 1 has no roots [15]**Помилка! Джерело посилання не знайдено.**. Using the lag operator, it can be written as

$$x_t = \mu + \theta(L)u_t. \quad (13)$$

Function

$$\rho(h) = corr(x_t, x_{t+h}), \quad (14)$$

is called the autocorrelation function (ACF) of a stationary time series. It is obvious that it is also an even function of the lag variable and p (0) = 1. For the autocorrelation coefficient, it is obvious:

$$corr(x_s, x_t) = \rho(s - t). \quad (15)$$

For an arbitrary stationary series, there is a limit to the autocorrelation function

$$\lim_{h \to \pm\infty} \rho(h) = 0. \quad (16)$$

This means that as the time lag increases, the elements of the time series become "less correlated". This can be interpreted as follows: with the growth of time t, the time series "forgets its past states", because $corr(x_s, x_t) = \rho(s - t) \to 0$, $t \to +\infty$, if s fixed [12].

As for ACF, p (h) = 0 at |h| > q, i.e., the series "forgets" past values with lags greater than the order of the model.

The AR model of order p has the following form:

$$x_t = \mu + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + u_t, \quad u_t \sim WN(0, \sigma_u^2), \phi_p \neq 0. \tag{17}$$

It can be seen that the current value in this model depends on the past values before the p lag and on the current external shock [10].

Using the lag operator, the model can be written as

$$\phi(L)x_t = \mu + u_t. \tag{18}$$

Along with the autocorrelation function, a partial autocorrelation function (PACF) is also considered, which is a particular correlation coefficient between the levels of a time series x_t and x_{t+h} with the exclusion of the influence of intermediate levels $x_{t+1}, \ldots, x_{t+h-1}$.

$$\rho_{part}(h) = corr(x_t, x_{t+h} | x_{t+1}, \ldots, x_{t+h-1}) \tag{19}$$

Obviously, $\rho_{part}(0) = 1, \rho_{part}(1) = \rho(1)$. And the peculiarity of the particular autocorrelation function is that

$$\rho_{part}(h) = 0 \text{ при } |h| > p \text{ и } \rho_{part}(p) = \phi_p. \tag{20}$$

III. NON-STATIONARY TIME SERIES

A. DIFFERENTIATION OF THE SERIES

In order to move to a stationary series, the differentiation operation is often used. The operation of differentiation or the first finite difference of a series is denoted as

$$\Delta \boldsymbol{x_t} = \boldsymbol{x_t} - \boldsymbol{x_{t-1}} = (\boldsymbol{1} - \boldsymbol{L})\boldsymbol{x_t}, \tag{21}$$

where L is the lag operator.

The idea is to consider its increment in one period instead of the original series.

Second-order differentiation:

$$\Delta^2 x_t = \Delta(\Delta x_t) = x_t - 2x_{t-1} + x_{t-2}.$$

Differentiation of arbitrary order:

$$\Delta^2 x_t = \Delta(\Delta^{k-1} x_t).$$

Using the lag operator, we can formally write

$$\Delta^2 x_t = (1 - L)^2 x_t = (1 - 2L + L^2) x_t$$

and in the general case

$$\Delta^{\boldsymbol{k}} \boldsymbol{x_t} = (\boldsymbol{1} - \boldsymbol{L})^{\boldsymbol{k}} \boldsymbol{x_t}. \tag{22}$$

B. ARIMA MODEL

For non-stationary series, which can be represented as stationary with respect to increments, we introduce the following class of models. $x_t \sim ARIMA(p, k, q)$, if

1. $x_t \sim I(k)$ integrated order k;
2. $\Delta^k x_t \sim ARMA(p, q)$.

Formally, we will consider *ARMA (p, q) = ARIMA (p, 0, q)*. Let $x_t \sim ARIMA(p, k, q)$ [7]. This means that $\Delta^k x_t$ is stationary *ARMA (p, q)*, and can be represented as

$$L^n x_t = x_{t-n} \tag{23}$$

$$\left(1 - \phi_1 L - \cdots - \phi_p L^p\right)\Delta^k = \mu + \left(1 + \theta_1 L + \cdots + \theta_q L^q\right)u_t \tag{24}$$

and all the roots of the autoregressive polynomial

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \tag{25}$$

modulo more than 1 (stationarity condition). Then using $\Delta^k x_t = (1 - L)^k x_t$, we get

$$\phi(L)(1 - L)^k x_t = \mu + \theta(L)u_t, \tag{26}$$

where $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z_q$. We introduce a polynomial

$$\gamma(z) = \phi(z) = 1 - \gamma_1 z - \cdots - \gamma_{p+k} z^{k+p}. \tag{27}$$

It has a unit root of multiplicity k, and the remaining p roots modulo more than 1.Then for the series $x_t \sim ARIMA(p,\ k, q)$

$$\gamma(L)x_t + \mu + \theta(L)u_t \Rightarrow x_t = \mu + \sum_{j=1}^{k+p} \gamma_j x_{t-j} + u_t + \sum_{s=1}^{q} \theta_s u_{t-s}. \tag{28}$$

Now let us move on to the implementation of the ARIMA model.

IV. MODEL PREDICTION IN PYTHON

A. FORECASTING A SHORT-TERM ECONOMIC INDICATOR

Data from the state website were selected for forecasting stat.gov.kz - Bureau of National Statistics of the Republic of Kazakhstan [17]. The data under study is an indicator of a short-term economic indicator (CEI). KEI is calculated to provide operational output indicators for basic industries, such as agriculture, trade, industry, construction. These figures are 60% of the total GDP. The indicators are presented in a monthly interval in the period from 2009 to 2021. The total number of indicators is 146. The selected data were presented in tabular form in Excel format. Figure 1 shows the beginning of the data that was taken from the table.

date	price
2009-01-01	1039.1
2009-02-01	1083.5
2009-03-01	1202.5
2009-04-01	1293.4
2009-05-01	1378.1
2009-06-01	1325.5
2009-07-01	1828.5

Figure 1. A fragment of data. At current prices for the net month, billion tenge

We need to develop a test kit for statistical data research and evaluation of candidate mathematical models.

This involves two steps:

1. Determining the sample of statistical data to be checked.
2. Development of a mathematical model and the use of a mathematical method for evaluating the model [8].

B. ANALYSIS OF STATISTICAL DATA FOR THE DEVELOPMENT OF THE ARIMA MODEL

On the website stat.gov.kz contains statistical indicators of the economic indicator by month. Therefore, it is advisable to use part of the sample to build a mathematical model, and part to check the predicted values. Usually, they take the last 12 months of statistical data for this. The data for verification contains indicators from August 2021 to July 2022, and the data for building the model contains indicators from January 2009 to July 2021.

Figure 2. Data graph of the short-term economic indicator

From the graph we can draw the following conclusions:

- Indicators of the indicator increase from year to year. This means that the time series has a pronounced trend.
- There do not seem to be any obvious outliers.
- There are relatively large fluctuations from year to year, up and down.
- The fluctuations in later years seem to be greater than the fluctuations in earlier years.
- Trend means that the sampling of statistical data is almost non-stationary [9].

Time series have the character of seasonality, for example, sales are always low at the beginning of the year and high at the end of the year. To track seasonality, the ARIMA seasonal model is used – ARIMA (p, d, q) (P, D, Q)S. Here (p, k, q) are the non–seasonal parameters described above, and (P, K, Q) follow the same definitions, but apply to the seasonal component of the time series. The parameter s defines the periodicity of the time series (4 – quarter periods, 12– year periods, etc.) [3].

We can also visualize our data in Figure 3 using a method called time series decomposition, which allows us to break our time series into three separate components: trend, seasonality, and noise [6]. The following graph shows:

- Current row.
- Trend chart of the series.
- Indicator of the seasonality of the series.
- Remains of the row.

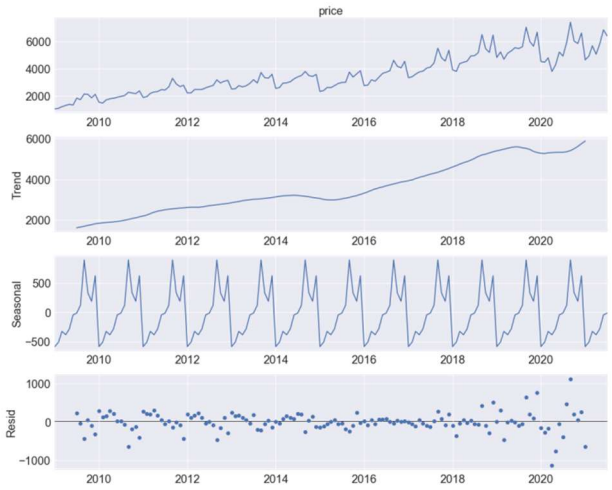

Figure 3. Decomposition of a series

C. BUILDING THE ARIMA MODEL. THE BOXING–JENKINS METHODOLOGY.

The Box-Jenkins methodology for selecting an ARIMA model for this series of observations consists of three stages:

1. Identification - using the data and all related information to help select the subclass of the model that will best be able to summarize the data.
2. Evaluation - using data to train model parameters.
3. Diagnostic verification - evaluation of the fitted model in the context of available data and verification of areas where the model can be improved [5]. Identification is divided into the following steps:

1. Evaluation of the time series for stationarity. If the series is not stationary, then the number of comparisons necessary to make it stationary is calculated.
2. Identification of ARMA model parameters for data.

It is necessary to consider some recommendations regarding this step:

- Test root modules in order to check the series for static. Repeat the tests after each round of comparison.
- It is necessary to avoid excessive comparison, as this will lead to additional correlation and additional complexity.
- Thus, two diagnostic graphs were used for the AR and MA configuration, which help to select the p and q parameters for ARMA or ARIMA:
- Autocorrelation Function (ACF): The graph shows the sum of the correlation of observations with delay values (lags).
- Partial Autocorrelation Function (PACF): The graph shows the sum of the correlation for data with lag values that are not taken into account for the previous lagging values of observations.

It was found that both graphs are depicted as a histogram showing 95% and 99% confidence intervals in the form of horizontal lines. The bars that cross these intervals are correspondingly more significant.

The estimation step involves the use of numerical methods in order to minimize losses or errors.

The idea of the diagnostic check is that at this step, evidence of non-compliance of the model with the data should be searched.

The first check is to analyze the compliance of the model with the data. This usually means that the model is more complex than it should be and captures random noise in the training data. Retraining is a serious problem for time series

forecasting, as it negatively affects the ability of the model to generalize, which leads to a decrease in forecast performance on out-of-sample data [1].

Particular attention should be paid to the performance of both the training sample and the test sample. That is why we should carefully select the data to prepare the model.

The forecast of residual errors provides great opportunities for diagnosis. An overview of the error distribution helps to align the bias of the model. Errors from an ideal model will resemble white noise, that is, a Gaussian distribution with an average value of zero and symmetric variance [16]**Помилка! Джерело посилання не знайдено.**.

In order to obtain a more objective assessment of stationarity, econometricians have developed tests based on the verification of statistical hypotheses. One such stationarity test is the extended Dickey–Fuller test. The following code in Figure 8 allows to implement this in Python. From the statsmodels.tsa.stattools package, to work with statistical data, we use the function for the Dickey–Fuller test [2]**Помилка! Джерело посилання не знайдено.**:

```
dftest = adfuller(train, autolag = 'AIC')
print("1. ADF : ",dftest[0])
print("2. P-Value : ", dftest[1])
```

```
1. ADF :  0.46052084494494716
2. P-Value :  0.9836224557617228
```

Figure 4. Dickey-Fuller test calculation code

From this test, we can see that the p-value is greater than 0.05. This indicates the non-stationarity of the series. Following the Box–Jenkins method, we lead to a stationary series, that is, we differentiate it.

$$\Delta_{12}x_t = x_t - x_{t-12} = (1 - L^{12})x_t$$

$\Delta\Delta_{12}x_t = (1 - L - L^{12} + L^{13})x_t = x_t - x_{t-1} - x_{t-12} + x_{t-13}$.

$L^{12}x_t = x_{t-12}$, L – lag operator.

Dickey-Fuller test after differentiation:

1. ADF : -4.15700
2. P-Value: 0.00078

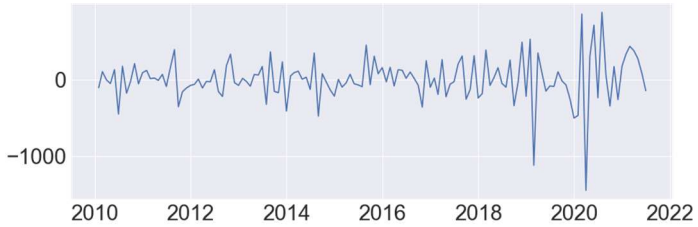

Figure 5. Differentiated series

The series has become stationary, that is, the p-value is below 0.05. Now we identify the parameters AR and MA. For this purpose, ACF and CHAKF graphs were constructed. The graph of the autocorrelation function is called an autocorrelogram and describes the parameter of the moving average q. And the graph of a part autocorrelation function is called a part autocorrelogram and shows the autocorrelation parameter p. plot_acf and plot_pacf functions from the statsmodels module are used to plot the autocorrelation function (ACF) and private autocorrelation function (CACF). graphics.tsaplots [13].

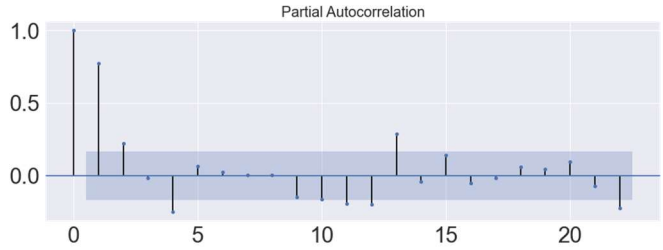

Figure 6. Part autocorrelogram

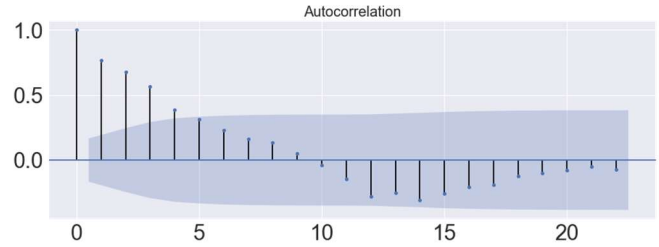

Figure 7. Autocorrelogram

Based on the calculations given above, the seasonal model ARIMA (p, k, q) (P, K, Q, S) is constructed. The series has the following parameters:

Non-seasonal parameters:

1. p is the auto regression parameter;
2. k is the order of differentiation, equal to 1;
3. q is the parameter of the moving average;

Seasonal parameters:

1. P is the parameter of seasonal auto regression;
2. K is the order of seasonal differentiation, equal to 1;
3. Q is the parameter of the seasonal moving average;
4. S is the period of seasonality, equal to 12.

ARIMA's seasonal model includes both non-seasonal and seasonal factors. Equation for seasonal ARIMA (p, k, q) (P, K, Q, S):

$$\phi(L)\Phi(L^S)\Delta x_t = \mu + \theta(L)\Theta(L^S)u_t \quad (29)$$

$$\begin{aligned}\left(1 - \phi_1 L - \cdots - \phi_p L^p\right)\left(1 - \beta_1 L^S - \cdots - \beta_p L^{pS}\right)(1 - L)^k(1 - L^S)^K x_t = \\ = \mu + (1 + \theta_1 L + \cdots + \theta_q L^q)(1 + \omega_1 L^s + \cdots + \omega_q L^{qS})u_t,\end{aligned} \quad (30)$$

where

ϕ – autoregression polynomial,

θ – moving average polynomial,

L – lag operator,

$\Phi(L^S)$ – seasonal autoregression polynomial,

$\Theta(L^S)$ – the polynomial of the seasonal moving average,

β – seasonal auto regression coefficients,

ω - seasonal moving average coefficients.

To integrate different combinations of parameters, a grid search is used. For each combination of parameters, the SARIMAX () function from the statsmodels module can select a new seasonal ARIMA model and evaluate its overall quality.

The optimal set of parameters will be the one in which the necessary criteria are most productive [14].

Following these instructions, we select the ARIMA model. Now it is possible to use the parameter triplets defined above to automate the process of evaluating ARIMA models in various combinations. In statistics and machine learning, this process is known as parameter grid search (grid search, or hyper parameter optimization).

When evaluating and comparing statistical models corresponding to various parameters, it is taken into account how well a particular model corresponds to the data and how accurately it is able to predict future data points. We use the AIC (Akaike Information Criterion) value, which is suitable for working with ARIMA models based on statsmodels.

V. RESULTS

As a result of selecting the model, the best one according to the Akaike criterion is ARIMA (2,1,1) (2,1,0,12). When selecting seasonal ARIMA models (as, indeed, any other models), it is important to diagnose the model to make sure that none of the assumptions made by the model has been violated. The plot_diagnostics object allows quickly diagnose the model and investigate any unusual behavior.

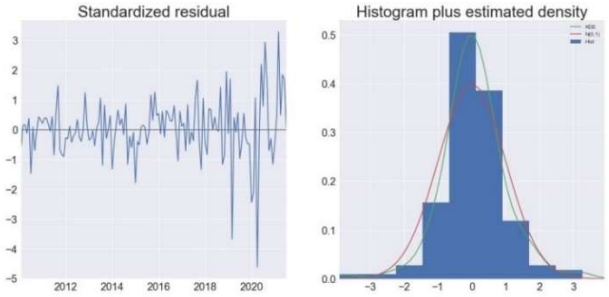

Figure 8. Histogram and calculated density

A good final check of the model is checking for residual forecast errors. Ideally, the distribution of residual errors should be Gaussian with zero mean. We can verify this by plotting the residuals using a histogram and density graphs. The main task is to make sure that the residuals of the model are uncorrelated and distributed with a zero mean value. If the ARIMA seasonal model does not satisfy these properties, it means that it can still be improved.

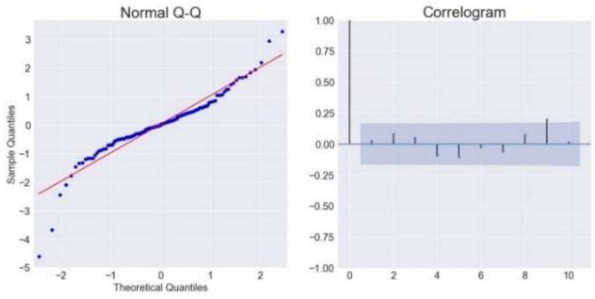

Figure 9. Correlogram

The graph shows that the distribution is normal and has an average of about zero. These graphs allow us to conclude that the selected model is (satisfactorily) suitable for analyzing and predicting time series data. We have chosen a satisfactory model, but some parameters of the ARIMA seasonal model can be improved. For example, a grid search considers a limited set of parameter combinations; to find the best model, we can expand the search.

The constructed ARIMA model can be used to predict future time steps. First, we need to compare the predicted values with the real values of the time series, which will help us understand the accuracy of the forecasts. When preparing the data, the last year was cut out to test the model. Let us assume that it is July 2021 and the forecast will be implemented for a year ahead, that is, for 12 periods.

To predict the model, the get_forecast and predicted_mean methods are used.

Figure 10. Comparison graph of test and predicted data

The graph shows that the predicted values almost repeat the test series. To test this, the average absolute error in percentages (MAPE) was calculated. The code is shown in Figure 11.

$$MAPE = 100\% \frac{\sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|}{n}, \quad (31)$$

where $A_t -$ *actual value*, F_t – the predicted value.

```
from sklearn.metrics import mean_squared_error
from math import sqrt

# ARIMA(2,1,1)(2,1,0,12)
mse = mean_squared_error(test, fc_series.dropna())
print('MSE: %.3f' % mse)
rmse = sqrt(mse)
print("RMSE: %.3f" % rmse)

mape = np.mean(np.abs(fc_series - test)/np.abs(test))
print("MAPE: %.3f" % (mape * 100) + '%')
```

```
MSE: 472673.863
RMSE: 687.513
MAPE: 7.260%
```

Figure 11. Error calculation code

VI. CONCLUSIONS

The ARIMA model (also known as Box–Jenkins models) is a very powerful and flexible class of models for analyzing and predicting time series. Over the years, they have been very successfully applied to solve many research and practical problems. The advantage of this model is the formalized and most thoroughly developed methodology, following which we can choose the model that is most suitable for each specific

time series. In addition, point and interval forecasts follow from the model itself and do not require separate evaluation.

In the course of this work, problems related to the requirement for data series are identified: at least 40 observations are required to build an adequate ARIMA model, and about 6-10 seasons are required for the seasonal ARIMA model, which is not always possible in practice. And there is also no simple way to adjust the parameters of the ARIMA model. It is necessary to periodically rebuild the model almost completely, and sometimes choose a completely new model.

As a result, the following tasks were completed:

1. Scientific papers on the research topic were analyzed;
2. The theoretical foundations of time series forecasting are described, AR, MA, ARMA, ARIMA models are considered;
3. The possibilities of the Python language for data analysis are considered and the program code for calculating and constructing the model is written;
4. To build the model, a sample of short-term economic indicator data from the state website of the Bureau of National Statistics of the Republic of Kazakhstan was selected. According to the Box–Jenkins approach, the time series is identified as a seasonal non-stationary series. The parameters of the ARIMA seasonal model (p, k, q)(P, K, Q, S) are selected in accordance with the Akaike criterion.

Based on the constructed model, the future values of the series are obtained and a comparison is made between the predicted values and real data. To check the error, the average absolute error in percent (MAPE) is calculated. The margin of error is 7.2%. Checking the residual errors shows that the residuals have a normal distribution.

References

[1] Aasim, S. N. Singh, and A. Mohapatra, "Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting," *Renewable Energy*, vol. 136, pp. 758–768, 2019. https://doi.org/10.1016/j.renene.2019.01.031.

[2] A. Lavanya, et al., "Assessing the performance of Python data visualization libraries: a review," *Int. J. Comput. Eng. Res. Trends,* vol. 10, issue 1, pp. 28-39, 2023. https://doi.org/10.22362/ijcert/2023/v10/i01/v10i0104.

[3] J. Banaś and K. Utnik-Banaś, "Evaluating a seasonal autoregressive moving average model with an exogenous variable for short-term timber price forecasting," *Forest Policy and Economics*, vol. 131, p. 102564, 2021. https://doi.org/10.1016/j.forpol.2021.102564.

[4] W. W. S. Wei, "Dimension reduction in high dimensional multivariate time series analysis," In: Zhang, L., Chen, DG., Jiang, H., Li, G., Quan, H. (eds) *Contemporary Biostatistics with Biopharmaceutical Applications. ICSA Book Series in Statistics*, Springer, Cham. https://doi.org/10.1007/978-3-030-15310-6_3.

[5] D. J. Pedregal, "Time series analysis and forecasting with ECOTOOL," *PloS one*, vol. 14, issue 10, e0221238, 2019. https://doi.org/10.1371/journal.pone.0221238.

[6] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed., OTexts, 2018. https://doi.org/10.32614/CRAN.package.fpp2.

[7] E. Dave, A. Leonardo, M. Jeanice, and N. Hanafiah, "Forecasting Indonesia exports using a hybrid model ARIMA-LSTM," *Procedia Computer Science*, vol. 179, pp. 480–487, 2021. https://doi.org/10.1016/j.procs.2021.01.031.

[8] H. Guo, Q. Chen, Q. Xia, C. Kang, and X. Zhang, "A monthly electricity consumption forecasting method based on vector error correction model and self-adaptive screening method," *Int. J. Electr. Power Energy Syst.*, vol. 95, pp. 427–439, 2018. https://doi.org/10.1016/j.ijepes.2017.09.011.

[9] M. K. Ahmadzai and M. Eliw, "Using ARIMA models to forecasting of economic variables of wheat crop in Afghanistan," *Asian J. Econ. Bus. Account.*, vol. 13, issue 4, pp. 1–21, 2020. https://doi.org/10.9734/ajeba/2019/v13i430180.

[10] M. S. Khan and U. Khan, "Comparison of forecasting performance with VAR vs. ARIMA models using economic variables of Bangladesh," *Asian J. Probab. Stat.*, vol. 10, issue 2, pp. 33–47, 2020. https://doi.org/10.9734/ajpas/2020/v10i230243.

[11] M. Elsaraiti, A. Merabet, & A. Al-Durra, "Time series analysis and forecasting of wind speed data," *Proceedings of the 2019 IEEE Industry Applications Society Annual Meeting*, 2019, pp. 1-5. https://doi.org/10.1109/IAS.2019.8912392.

[12] Yu. M. Minaev, O. Yu. Filimonova, and Yu. I. Minaeva, "Forecasting of fuzzy time series based on the concept of the nearest fuzzy sets and tensor models of time series," *Cybernetics and Systems Analysis*, vol. 59, no. 1, pp. 165-176, 2023. https://doi.org/10.1007/s10559-023-00551-9.

[13] A. Zemkoho, "A basic time series forecasting course with Python," *Operations Research Forum*, vol. 4, no. 1, Cham: Springer International Publishing, 2022. https://doi.org/10.1007/s43069-022-00179-z.

[14] R. Adhikari, and R. K. Agrawal, "An introductory study on time series modeling and forecasting," *arXiv preprint arXiv:1302.6613*, 2013.

[15] W. Polasek, "Time series analysis and its applications: With R examples," pp. 323-325, 2013. https://doi.org/10.1111/insr.12020_15.

[16] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A comparison of ARIMA and LSTM in forecasting time series," *Proceedings of the 2018 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2018, pp. 1394-1401. https://doi.org/10.1109/ICMLA.2018.00227.

[17] Agency for Strategic Planning and Reforms of the Republic of Kazakhstan Bureau of National Statistics, "The short-term economic indicator," Dataset, 2022. [Online]. Available at: https://stat.gov.kz/api/getFile/?docId=ESTAT108589.

[18] Ch. Faloutsos, et al., "Forecasting big time series: old and new," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 2102-2105, 2018. https://doi.org/10.14778/3229863.3229878.

[19] Y. Liang, et al., "Foundation models for time series analysis: A tutorial and survey," *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6555-6565. https://doi.org/10.1145/3637528.3671451.

[20] Y. Zou, et al., "Complex network approaches to nonlinear time series analysis," *Physics Reports*, vol. 787, pp. 1-97, 2019. https://doi.org/10.1016/j.physrep.2018.10.005.

[21] K. M. Bazikova, G. A. Abdenova, and G. E. Sagyndykova, "Linear stochastic distributed model of money accumulation in the form of a state space," *KazNU Bulletin. Mathematics, Mechanics, Computer Science Series*, vol. 110, no. 2, pp. 128–138, 2021. https://doi.org/10.26577/JMMCS.2021.v110.i2.11.

[22] A. Abdenov, G. Abdenova, Z. Kenzhebayeva, and A. Mukhanova, "Labour productivity prediction estimate in manufacturing environment using regression and dynamic models," *Mater. Today: Proc.*, vol. 16, pp. 254–261, 2019. https://doi.org/10.1016/j.matpr.2019.05.087.

[23] A. Abdenov, G. Abdenova, and D. Kulbayev, "Estimation of equation coefficients of free and forced string vibrations in continuous medium with consideration of dynamic noise and measurement noise," *Mater. Today: Proc.*, vol. 16, pp. 336–342, 2019. https://doi.org/10.1016/j.matpr.2019.05.099.

[24] A. R. Musin, "Comparison of the quality of forecast models of the foreign exchange market using Kalman filtering and traditional time series models," *Internet J. "SCIENCE"*, vol. 9, no. 3, 2017.

[25] A. S. Sorokin and A. R. Musin, "On the issue of using the Kalman filter in econometric models," *Science and Practice*, no. 1(25), pp. 71–76, 2017.

[26] A. Khulood, B. Zafar, and A. Mueen, "Time series forecasting using LSTM and ARIMA," *International Journal of Advanced Computer*

Science and Applications, vol. 14, no. 1, pp. 313-320, 2023. https://doi.org/10.14569/IJACSA.2023.0140133.

GAUKHAR ABDENOVA, ***Ph.D., associate professor of the department of "Mathematical and computer modeling" of the ENU named after L.N.Gumilev. Scientific interests: Fundamentals of mathematical modeling, Modeling of natural phenomena, Modern methods of mathematical modeling, Mathematical packages in modeling, Computer methods for analyzing financial data, Econometrics, Operations research in economics, Hedging methods, Modeling risk situations, Designing information systems.***

ZHANAT E. KENZHEBAEVA, ***candidate of technical sciences, Acting Associate Professor "Computer science", Yessenov University. Education: Aktau State University named after Sh. Yessenova, 2002. Scientific interests: System analysis, information management and processing, Information Communication Technologies***

HANNA MARTYNIUK*, PhD, Associate Professor. Education: National Aviation University, 2011. Position: Associate Professor of the Department of System Analysis and Information Technologies, Mariupol State University. Scientific interests: information assurance of noise measurement; statistical models of information signals; statistical methods for measuring the characteristics of random processes and fields; methods of signal simulation and measurement data.*