Forecast the Gross Value Added in Construction Sector of Bulgaria with SARIMA Model

Assist. Prof. Plamen Yankov
University of Economics - Varna, Varna, Bulgaria
yankov.plamen@ue-varna.bg

Prof. Dr. Julian Vasilev
University of Economics - Varna, Varna, Bulgaria
vasilev@ue-varna.bg

Assoc. Prof. Dr. Pavel Petrov
University of Economics - Varna, Varna, Bulgaria
petrov@ue-varna.bg

Dr. Liliya Mileva
University of Economics - Varna, Varna, Bulgaria
l.mileva@ue-varna.bg

Chief. Assist. Prof. Dr. Svetlana Todorova
University of Economics - Varna, Varna, Bulgaria
svetlana.todorova@ue-varna.bg

Abstract

Construction is an important sector for national economies because it contributes with relatively high gross value added (GVA). The purpose of this study is to forecast GVA in a short-term period based on seasonal ARIMA models. Quarterly time series data from 2010 to 2020 are used for modelling and forecasting. Stationarity is achieved after differencing both - seasonal and non-seasonal component of the data. Based on autocorrelation plots SARIMA $(0,1,1)(0,1,0)_4$ *model is selected as most accurate. Ljung-Box test for the absence of autocorrelation confirms that the model is adequate and suitable to forecast. The current study is conducted as part of the research project BG05M2OP001-1.002-0002-C02 "Digitalization of Economy in a Big Data Environment".*

Keywords: SARIMA, gross value added, forecast, construction

JEL Code: E170

Introduction

The construction sector is a major driving force for national economies globally (Alaka *et al.*, 2020). The construction sector in Bulgaria produces about 15% of total GDP and employs about 200,000 employees. It is typical of construction projects that they are characterized by high investment and are usually accompanied by high financial risk (Chua *et al.*, 1997). For these reasons, the dynamics of the main indicators in the construction sector are the subject of frequent research, both at business and at the macroeconomic level (Ignatova, 2015).

Digital transformation also creates advantages at the macro level. The main one is to increase the accuracy of forecasts and the adoption of adequate policies in a particular sector of the economy. Currently, big data play an important role in compiling models for predictive analysis (Ngo, Hwang and Zhang, 2020).

To estimate the computational model in this research is used statistical package for econometric analysis – Gretl. This software is free for use and it is available for the Windows and macOS platforms at: http://gretl.sourceforge.net/

1. SARIMA Model Briefly

The autoregressive integrated moving average (ARIMA) and its seasonal equivalent (SARIMA) model is often used to forecast important macroeconomic indicators such as gross domestic product (GDP), inflation rates, employment rates (Ghazo, 2021) and yields from different industries (M. Ngungu, Jacob and Noah, 2018). Our study demonstrates that a trend and seasonal pattern exist of the gross value added of the construction sector. The SARIMA model takes both overall trends and seasonal changes into account, which is widely used in modelling time series data.

Time series forecasting with the ARIMA or SARIMA models is a quantitative technique based on the assumption that some aspects of past processes will continue to be repeated in the future. These aspects include trend and seasonality.

In the literature, the SARIMA model is also widely known as $(p, d, q)(P, D, Q)_s$ where each of the letters is an abbreviation for a specific mathematical component. The meaning of each component is detailed shown in the Table 1 and Figure 1 below.

Table 1. Mathematical components of the SARIMA models.

Component	Full name	Description
AR	*Autoregression*	A model that uses the dependent relationship between an observation and some number of lagged observations
I	*Integrated*	The use of differencing of raw observations in order to make the time series stationary
MA	*Moving average*	A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.
S	*Seasonal*	Total number of periods per one season

$$SARIMA = \underbrace{(p, d, q)}_{\text{non-seasonal component}} \underbrace{(P, D, Q)s}_{\text{seasonal component}} \quad s = \text{number of periods per season}$$

Figure 1. Components of the SARIMA model; source - author

The methodology of Box and Jenkins for time series forecasting is strictly followed in the current paper. It can be summarized in several specific stages as follows:

1. Identification. This stage includes 2 sub-stages, known as stationary check and model selection.
2. Estimation and diagnostic. At this stage the residues are checked for normality and in the presence of more than one suitable model the most accurate one is selected based on lower Akaike Information, Schwarz Bayes or Hannan-Quinn criterion.
3. Forecasting. This is the last step, which involves forecasting future values in a short period of time. SARIMA is a method that is suitable for short-term forecasts.

2. Data Source

The data in this study is accessed from Bulgarian National Statistical Institute. Time series include 44 observations for the Gross value added (GVA) of Bulgarian construction sector. Each line represents a separate quarter for an interval of time between 2010 – 2020 year, a total of 11

years at constant 2015 prices. Summary statistics of the analysed time series is shown in Table 2. The data applied in the study is retrieved from the National Statistical Institute which is also available at: https://nsi.bg/en/content/12587/quarterly-data .

Table 2. Summary statistics of the time series; source - own study.

Estimator	Estimate
Number of observations	44
Mean	623 510 000
Minimum	459 170 000
Maximum	1 129 900 000
Standard deviation	160 860 000

3. Model Identification

The dynamics of gross value added (GVA) in the construction sector of Bulgaria for 2010 - 2020 is illustrated in Figure 2.

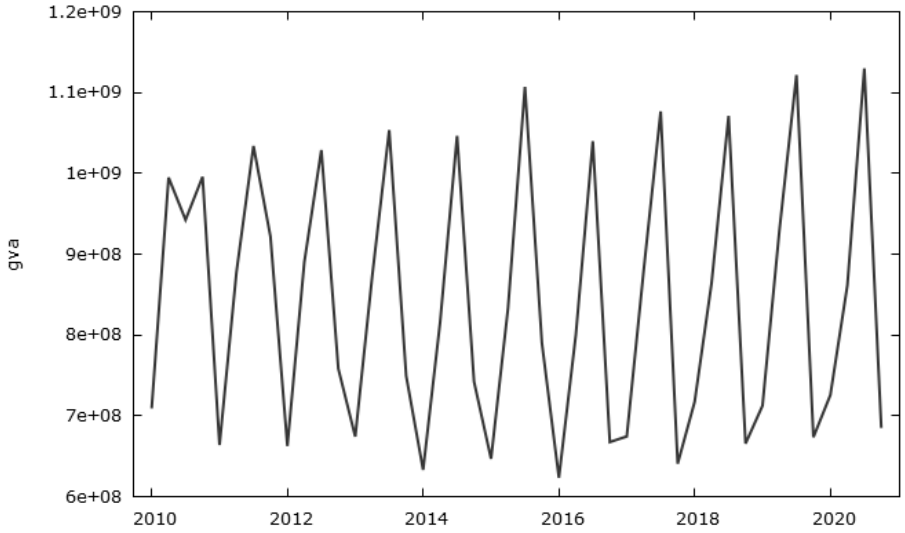

Figure 2. Time series plot of the quarterly GVA observations; source - own study

For time series forecasting, the observations need to be stationary. The time series data often consist of a trend and seasonality which is an evidence for non-stationary. A common assumption in the time series analysis is that the data is stationary. This means that the mean, variance and autocorrelation structure do not change over the time. Data with seasonality, as in the present case, is non-stationary by presumption (Wang, Li and Lim, 2019). Also, time series with trends is not stationary. In general, the presence of trend and seasonality is easily highlighted from the dynamics of the data chart. The automatic decomposition of the data provides a much more accurate idea of the presence of seasonal recurrence, as well as the emergence of a trend (Bonaros, 2021). In this way, the presence of trend and seasonality stands out much more clearly as could be noticed in Figure 3.

For this reason, the decomposition of time series is an important step preceding the actual forecasting (Nau, 2014). SARIMA models are extremely used in modelling data that show seasonal component such as weekly, monthly, quarterly and annual (M. Ngungu, Jacob and Noah, 2018). From the graph above it is cleared that seasonality occurred in every 3rd quarter where a significant

spike is noticeable. Oppositely every 1st and 4th quarter repeatedly show the lowest values in the analyse time series.

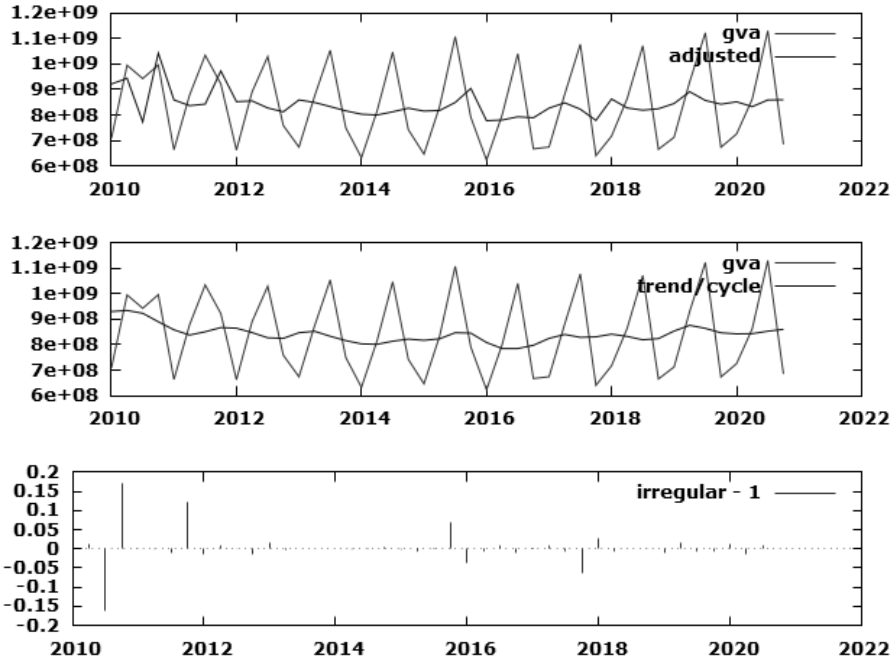

Figure 3. Seasonal decomposition of time series; source - own study.

The autocorrelation ACF plot in Figure 4 also suggests the presence of seasonality in the series. The ACF plot decreases slowly which is an indicator for non-stationary. Seasonality is also well distinguished from the ACF plot where the persistence of regularly repeated spikes indicates that the values at these lags are significant. These are the lags at 4th, 8th, 12th lag, etc.

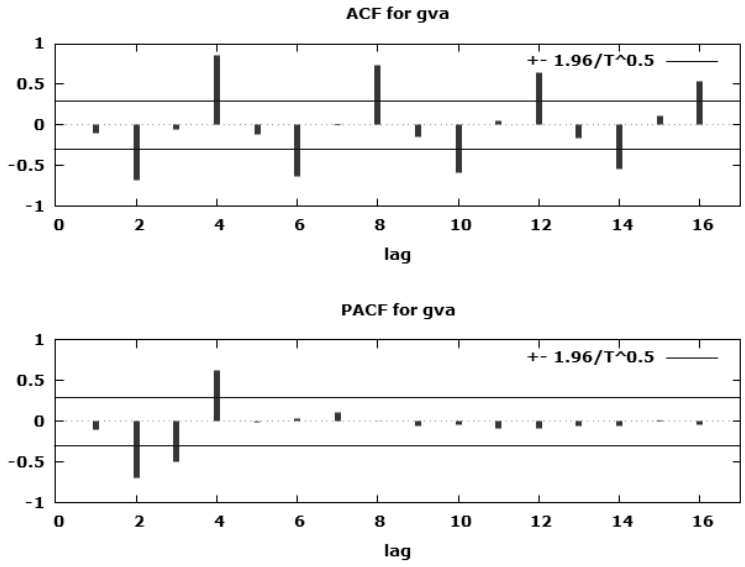

Figure 4. Autocorrelation function (ACF) and partial autocorrelation functions (PACF) for Gross value added in the construction sector of Bulgaria; source – own study.

The Augmented Dickey-Fuller statistical test for stationary is a common test which indicates whether a given time series is stationary or not (Dickey and Fuller, 1979). The ADF test has two hypotheses as follow:

$H_0 =$ The process is not stationary

$H_1 =$ The process is stationary (series is not stationary)

The ADF test is applied, and the result (p-value of 0.93) confirms the assumption that the series have a unit root. Applying a difference is required action which transforms the data from non-stationary to stationary. The ordinary difference is also known as "first difference" meaning difference at lag one. This transformation eliminates the trend.

Its mathematical representation is:

$$y'_t = y_t - t_{t-1}$$

After the first difference of the time series data the ADF test is repeated. The result from the ADF test proved that the data are stationary with a p-value equal to 0.0001. The new data is presented in Figure 5.

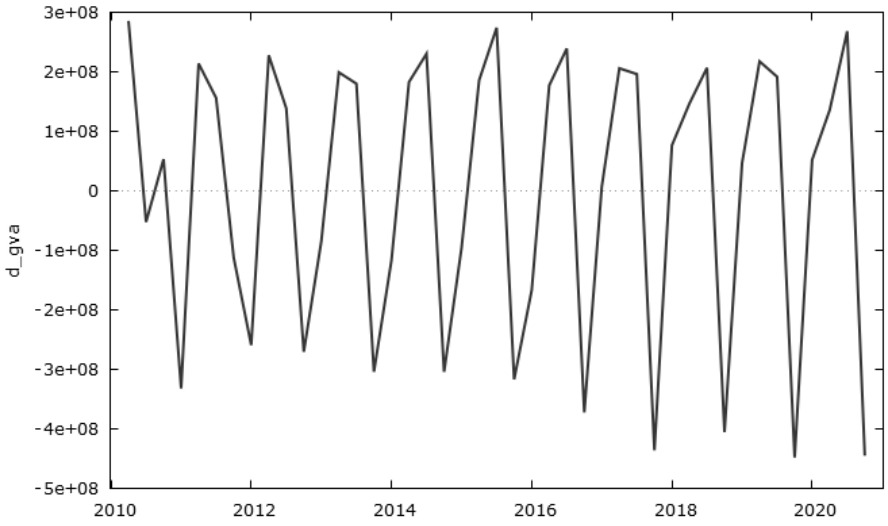

Figure 5. First differencing of the time series.

Seasonal time series are also subject to seasonal difference. The seasonal difference mathematically is described with the formula above:

$$\Delta_s Y_t = y_t - y_{t-s} \Delta_s Y_t = y_t - y_{t-s}$$

After seasonal difference is applied the ADF test confirms stationary with p-value of 0.0041. Figure 6 illustrates the seasonal differenced time series.

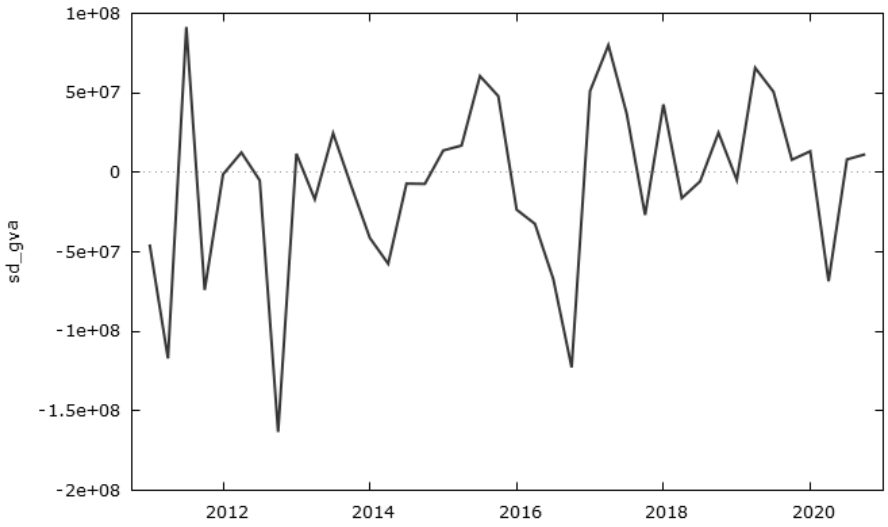

Figure 6. First seasonal differences of the time series; source - own study.

The next step includes the autoregressive and moving average components identification. This is done by examining the non-seasonal and seasonal differences ACF and PACF shown respectively in Figure 7 and Figure 8.

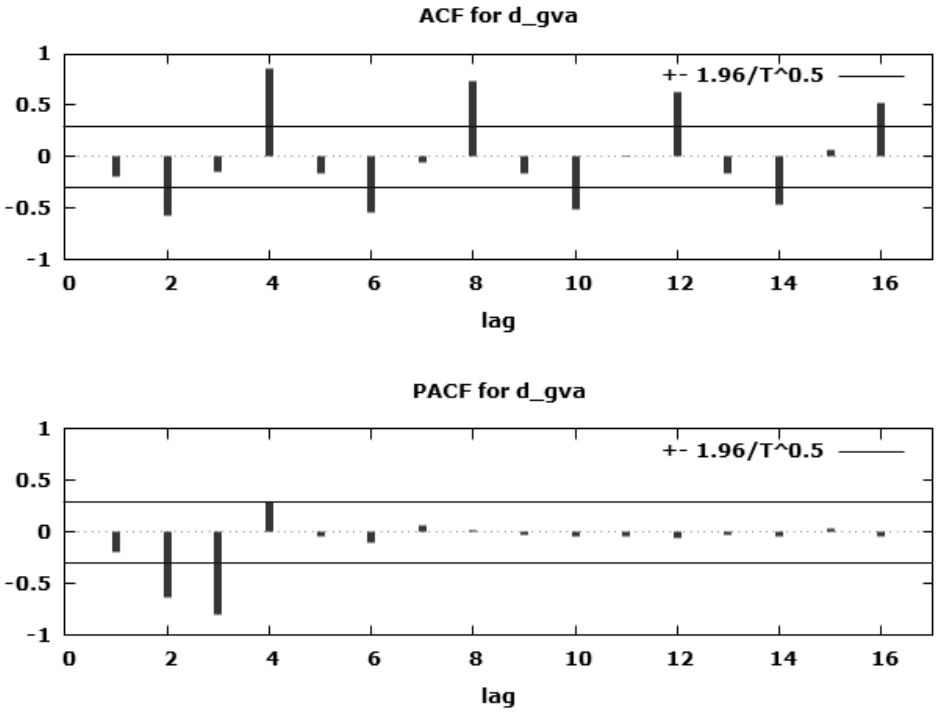

Figure 7. Differenced autocorrelation function (ACF) and partial autocorrelation functions (PACF) for historical construction GVA data; source: own study

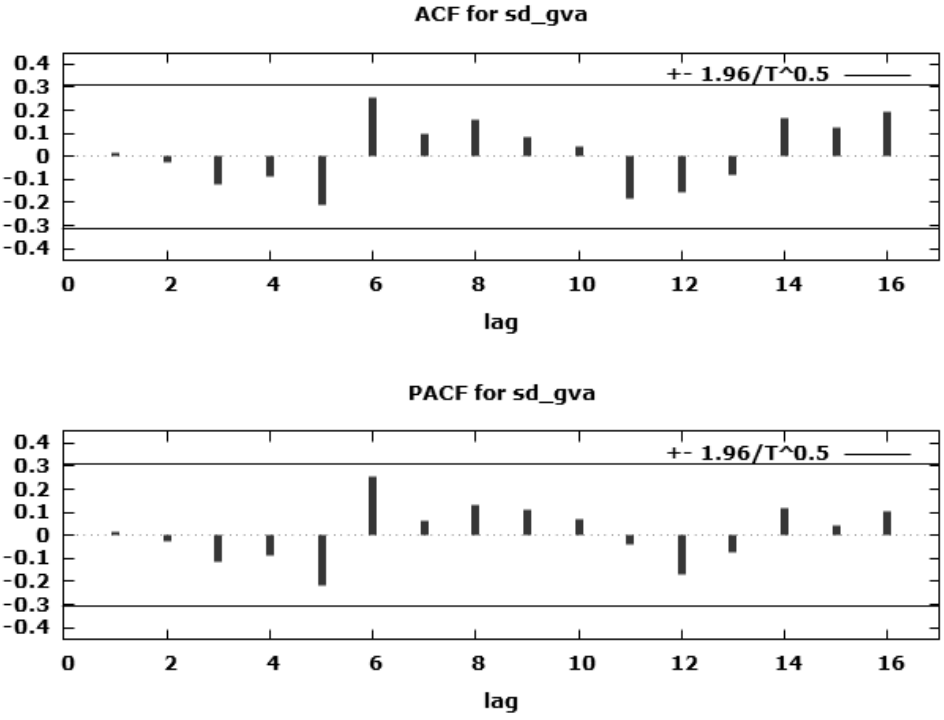

Figure 8. Seasonally differenced autocorrelation function (ACF) and partial autocorrelation functions (PACF) for quarterly construction GVA data; source: own study

In the specific case of this study, only SARIMA $(0,1,1)(0,1,0)_4$ model satisfies the conditions for significance of the coefficients shown in Table 3. Several other models, including SARIMA (0,1,1) (0,1,1)4, SARIMA (1,1,1) (0,1,0)4, are also tested, but their AR and MA coefficients are insignificant with a p-value exceeding the significance level α = 0.05. For this reason, they are not considered in detail. In the presence of several models that meet the conditions for significance of the parameter estimates, it is recommended to use the criteria of Akaike Information (Akaike, 1974), Schwarz Bayes (Schwarz, 1978) and Hannan-Quinn (Hannan and Quinn, 1979).

Table 3. Parameter estimates for SARIMA (0,1,1) (0,1,0)4

Component	θ_1
coefficient	-0.928487
p-value	<0.0001

The most commonly used indicators to assess the accuracy of the forecast are mean absolute percentage error (MAPE), mean absolute error (MAE) and root-mean-square error (RMSE). They are presented in Table 4.

Table 4. Comparative table of the presentation of valid models.

Model	**AIC**	**BIC**	**HQ**	**MAE**	**RMSE**	**MAPE**
$(0,1,1)\ (0,1,0)_4$	1506.969	1510.296	1508.163	4.2676e+007	55 093 000	5.14

4. Model Diagnostic

The residuals of the model are an important indicator that graphically shows how accurately the model captures the constant and the trend in the values that reflect the past process and serve to forecast the future periods. Both ACF and PACF of the residuals at lags up to 16 represented in Figure 9 show the absence of significant autocorrelation within the 95% confidence interval (Mahan, Chorn and Georgopoulos, 2015).

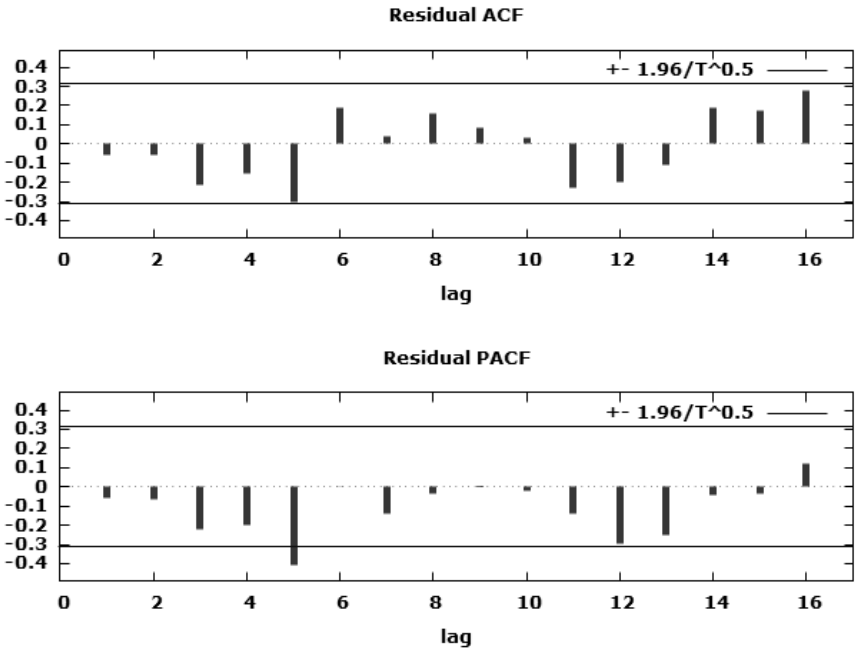

Figure 9. ACF and PACF plot of residuals; source – own study.

The presence of normal distribution of the residuals is tested with the Jarque-Bera test. The null hypothesis is that residuals are normally distributed and against the alternative the residuals are not normally distributed. The JB test of normality fails to reject the null hypothesis which proves the residuals are normally distributed (p-value = 0.2849). The result of the JB test is confirmed by a few other similar tests for normality. Table 5 and Figure 10 illustrate all 3 tests for normality as well as the condition that the model is adequate for GVA forecasting.

Table 5. Tests for normality of the residuals; source - own study.

Normality test	t-statistics	p-value
Jarque-Bera	2.5112	0.2849
Shapiro-Wilk W	0.9748	0.5191
Lilliefors	0.0932	≈0.52

Ljung-Box test for autocorrelation up to lag order 12 is 16.7042 and p-value = 0.1169. The higher p-value shows that residuals are normally distributed (α = 0.05).

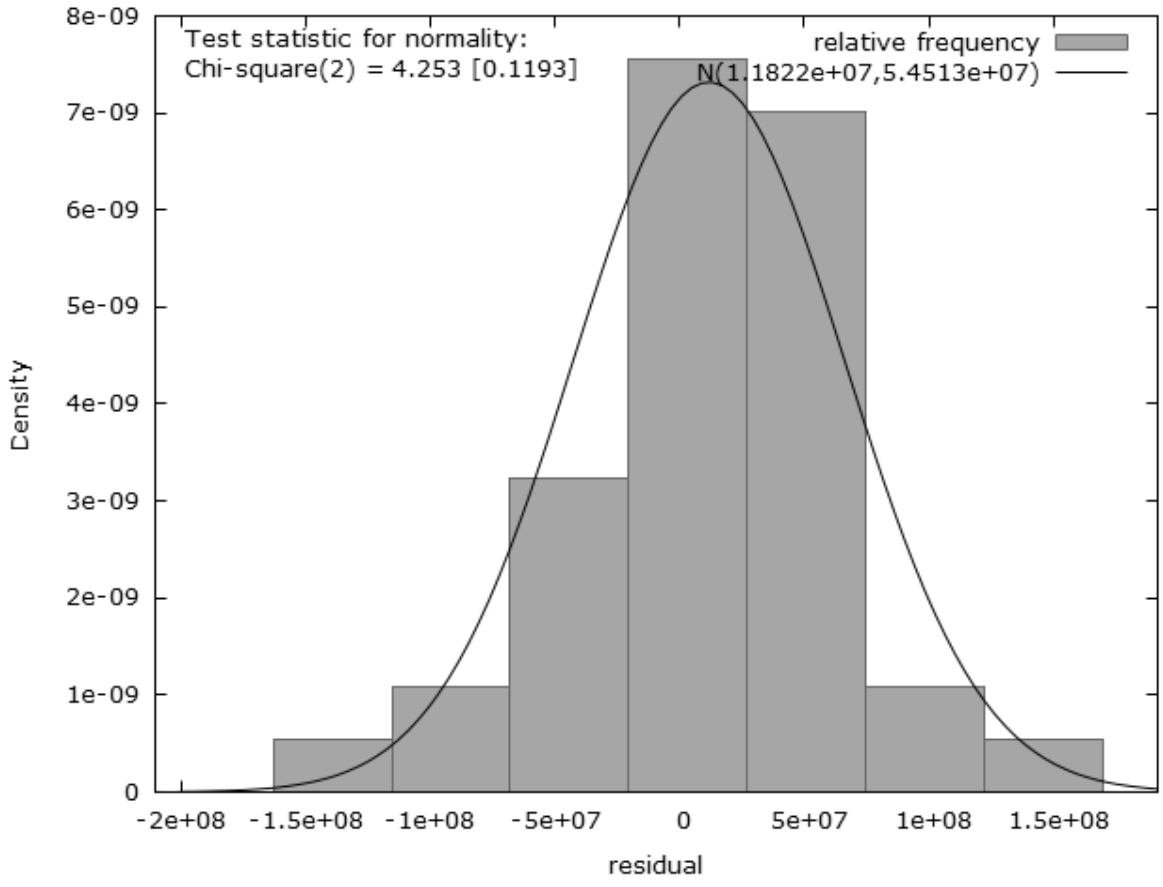

Figure 10. Test for normal distribution of the residuals; source – own study;

5. Forecasting

In the final step, SARIMA $(0,1,1)(0,1,0)_4$ is applied as shown in Figure 11. The forecasted values start from the first quarter of 2021 year and include a 2-year period. They are shown on the graph with a line for the minimum and maximum values as a result of the 95% confidence interval.

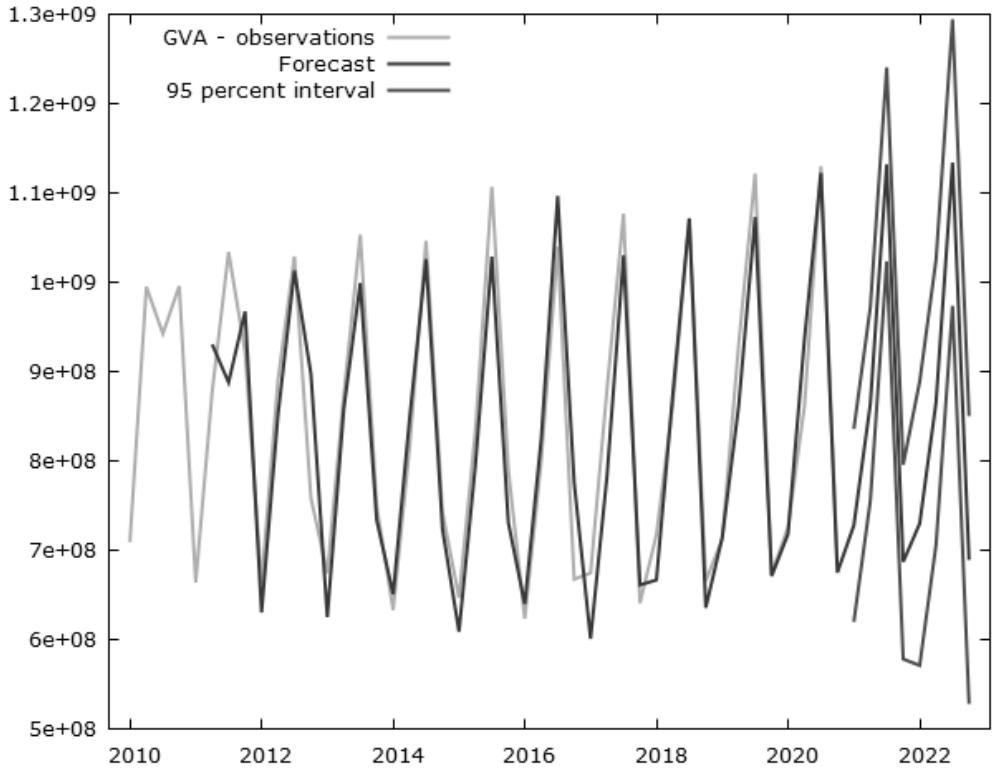

Figure 11. Forecast from SARIMA $(0,1,1)(0,1,0)_4$

The forecasted values are also present in Table 6, in numerical form.

Table 6. Forecasted values of the real GVA of construction (base period: 2015), based on SARIMA $(0,1,1)(0,1,0)_4$; source - own study

Period	Forecast	Lower 95% Limit	Upper 95% Limit
2021-Q1	727 911 000	619 931 000	835 891 000
2021-Q2	863 729 000	755 473 000	971 985 000
2021-Q3	1 131 960 000	1 023 420 000	1 240 490 000
2021-Q4	687 166 000	578 361 000	795 971 000
2022-Q1	729 969 000	571 143 000	888 794 000
2022-Q2	865 787 000	706 212 000	1 025 360 000
2022-Q3	1 134 010 000	973 693 000	1 294 330 000
2022-Q4	689 224 000	528 161 000	8.50286e+008

Meanwhile, the National Statistical Institute announces preliminary data from the first 2 quarters of 2021 for the GVA in the construction shown in Table 7. Preliminary data fall within the minimum and maximum values of the forecast discussed in this study. For the second quarter, the deviation is about 4.5%, which is a good accuracy.

Table 7. Forecast and actual data comparison of the Real GVA of construction (base period: 2015)

Period	Forecast	Preliminary data	Lower 95% Limit	Upper 95% Limit
2021-Q1	727 911 000	634 227 000	598 414 550	874 095 510
2021-Q2	863 729 000	905 454 000	738400.42	1052646.58

Conclusion

Forecasts play an important role in the national and sectoral policymaking. This study examines the dynamics in the gross added value (GVA) for 11 years covering 44 time series in the period 2010 - 2020. The data have all the characteristics of seasonal time series. Based on this finding, a forecasting method with SARIMA is applied. In the course of building the model, all important stages are followed according to the Box-Jenkins methodology, which include identification, calculation and forecasting. The residuals from the model show a normal distribution, which is indicative of the adequacy of the model to respond to deviations in the data. The generated forecasts with the computational model SARIMA $(0,1,1)(0,1,0)_4$ are evaluated according to several criteria, including MAPE, MAE and RMSE. Finally, the actual data for the first 2 quarters of 2021 year published by the NSI are compared with the current forecasts form the SARIMA model. The results confirm the good fit of the model to the time series. The research findings could provide government authorities and business practitioners with better process planning and decision making.

References

1. Akaike, H. (1974) 'A New Look at the Statistical Model Identification'.
2. Alaka, H. A. *et al.* (2020) 'A framework for big data analytics approach to failure prediction of construction firms', *Applied Computing and Informatics*. The Authors. doi: 10.1016/j.aci.2018.04.003.
3. Bonaros, B. (2021) *Time Series Decomposition In Python*. Available at: https://predictivehacks.com/time-series-decomposition/.
4. Chua, D. K. H. *et al.* (1997) 'Neural networks for construction project success', *Expert Systems with Applications*, 13(4), pp. 317–328. doi: 10.1016/S0957-4174(97)00046-8.
5. Dickey, D. A. and Fuller, W. A. (1979) 'Distribution of the Estimators for Autoregressive Time Series with a Unit Root'.
6. Ghazo, A. (2021) 'Applying the ARIMA Model to the Process of Forecasting GDP and CPI in the Jordanian Economy', *International Journal of Financial Research*, 12(3), p. 70. doi: 10.5430/ijfr.v12n3p70.
7. Hannan, E. J. and Quinn, B. G. (1979) 'The Determination of the Order of an Autoregression'.
8. Ignatova, N. (2015) 'Динамика на основните икономически показатели в сектор „ Строителство " в България за периода 2009-2013 г .', pp. 128–142.
9. M. Ngungu, M., Jacob, O. and Noah, W. (2018) 'Modeling Agricultural Gross Domestic Product of Kenyan Economy Using Time Series', *Asian Journal of Probability and Statistics*, (November), pp. 1–12. doi: 10.9734/ajpas/2018/v2i124563.
10. Mahan, M. Y., Chorn, C. R. and Georgopoulos, A. P. (2015) 'White Noise Test : detecting autocorrelation and nonstationarities in long time series after ARIMA modeling', (July). doi: 10.25080/Majora-7b98e3ed-00f.
11. Nau, R. (2014) 'The mathematical structure of ARIMA models', 1(c), pp. 1–8.
12. Ngo, J., Hwang, B. and Zhang, C. (2020) 'Automation in Construction Factor-based big data and predictive analytics capability assessment tool for the construction industry', *Automation in Construction*. Elsevier, 110(December 2019), p. 103042. doi: 10.1016/j.autcon.2019.103042.
13. Schwarz, G. (1978) 'Estimating the Dimension of a Model'.
14. Wang, S., Li, C. and Lim, A. (2019) 'Why Are the ARIMA and SARIMA not Sufficient', pp. 1–10. Available at: http://arxiv.org/abs/1904.07632.