LJAST
Libyan Journal of Applied Science and Technology
مجلة ليبيا للعلوم التطبيقية والتقنية

Volume 10 Issue 02
December 2022
ISSN 2958-6119

Using ARIMA model to forecast the Reported Crimes in Libya from 2020 to 2025

Adel Ewhida[1]
[1]Tripoli University, Faculty of Science, Department of Statistics
P.O. Box 13219, Tripoli, Libya

Abstract

A time series is a collection of data that was compiled chronologically. Most time series can be predicted based on their present and historical values. Software programs like Spss and Minitab are made specifically to process time series data. The SPSS and Minitab software programs can be used to create the autoregressive integrated moving average (ARIMA) model, a time series forecast technique. In this paper, the forecasting process using the ARIMA model is illustrated using the SPSS and Minitab software. For instance, the predicted period for reported crimes (RC) in Libya is from 2020 to 2026.

Key words: Time series, Autoregressive integrated moving average model, Reported crimes in Libya.

1. Introduction

In time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model, which fitted to time series data either to better understand the data or to forecast future points in the series. The ARIMA model was proposed by Box and Jenkins in the 1976. In general, a stationary sequence can establish a metrology model. As for a non-stationary sequence, it should be converted to a stationary sequence with difference operation. ARIMA models has been used to forecast the London daily gold price (see, Abdullah L., 2014; Ediger SA., 2006) and primary energy demand on fossil fuel (Widowati et al., 2016). The ARIMA models also has been used for predicting production area and productivity of cotton crop in Syria (see, Almohammad S, Jasem I, Lubboss M., 2018). Also to predict the amount of production in the 2nd steel plant of the Libyan Iron and Steel Company in Misurata (see Al Shafi MA. and Blhaj A., 2020). In this paper, we uses the ARIMA model in forecasting the Reported Crimes

(RC) of Libya because it is well used for short term and medium term forecasting and it is suitable for various data.

2. Methodology

The statistical method was used by analyzing the time series, and working on its stability by means of eliminate the effect of the general trend and variance, then diagnose the ARIMA(p, d, q) model, which combines the two methods of autoregressive and moving average (see, Zhang PG., 2003), where:

- **p** is the number of autoregressive terms,
- **d** is the number of non-seasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.

The steps to adopt ARIMA method are (see, Jenkins B et al., 1994):

a. Model identification

The first step that must be done is to investigate whether the time series data is stationary or not. If the data is not stationary, it must be converted to a stationary series by finding the first difference of the variable as follows (Mandala QS., 1977):

$$Y_t^* = \Delta Y = Y_t - Y_{t-1}.$$

However, if the first differences do not result in a stationary series, the differences process can be repeated until a stationary series is obtained. The differences required to convert the series into a stationary series have been experienced with the degree of integration, so it is said that the series is integrated from the degree (d).

b. Identification of parameters (ACF and PACF)

To use the ARIMA model, the value of d, the number of remaining lag values (q), and the dependent lag value (p) used in the model must be specified. The main tools used to determine q and p are ACF (autocorrelation function) and PACF (partial autocorrelation function).

c. Selection of the best ARIMA model

As a result of stationary identification and identification of ACF and PACF, there will be several models of ARIMA. Therefore, the suitability of the model will be checked by studying the characteristics of the model residues.

d. Forecasting

After obtaining the best model, the future period can be forecasted. In different cases, forecasting by this method is more reliable than forecasting using other time series methods.

3. Data description

The Reported Crimes (RC) is the core indicator of national security. It is an important index to measure the overall security situation of some country. It reflects the country's security strength, structural layout. In 2019, the Bureau of Statistics in Ministry of the Interior, the Criminal Investigation Service, reported data on the reported crimes from 1969 to 2019 (See Table 1).

Table 1: Reported crimes in Libya during 1969 to 2019

Year	Reported Crimes	Year	Reported Crimes	Year	Reported Crimes
1969	46622	1987	29756	2005	74210
1970	34606	1988	33679	2006	72483
1971	37749	1989	36112	2007	75622
1972	40914	1990	35234	2008	80003
1973	43508	1991	38070	2009	76889
1974	42897	1992	39900	2010	65726
1975	46026	1993	44400	2012	26923
1976	39759	1994	45166	2013	21665
1977	34890	1995	51101	2014	13503
1978	34219	1996	49113	2015	10183
1979	32764	1997	49494	2016	9387
1980	32281	1998	48300	2017	11266
1981	29134	1999	53226	2018	14656
1982	30256	2000	52526	2019	14151
1983	30734	2001	54361		
1984	30967	2002	50069		
1985	29156	2003	53807		
1986	30849	2004	55443		

4. Stationarity test

The RC data series during 1969-2019 is plotted is Figure 1, which shows the pattern in which the number of reported crimes develops during the period under study, in order to identify the main features of the series such as the general direction, dispersion, stationary and autocorrelation. The curve of the crime series shows the existence of a general direction which is increase over the period under study, which means that the series is not static in the mean. It is also clear existence of a positive autocorrelation between the observations of the series.

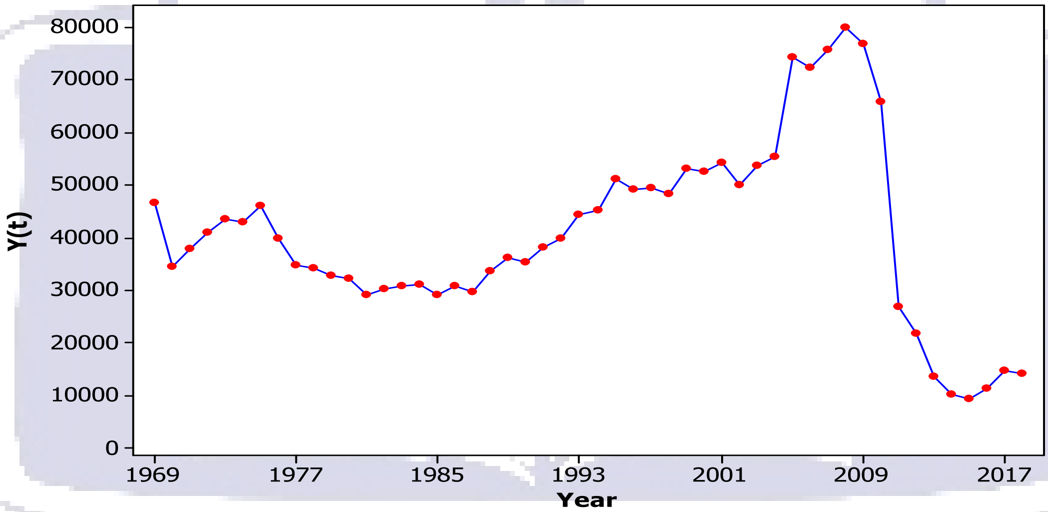
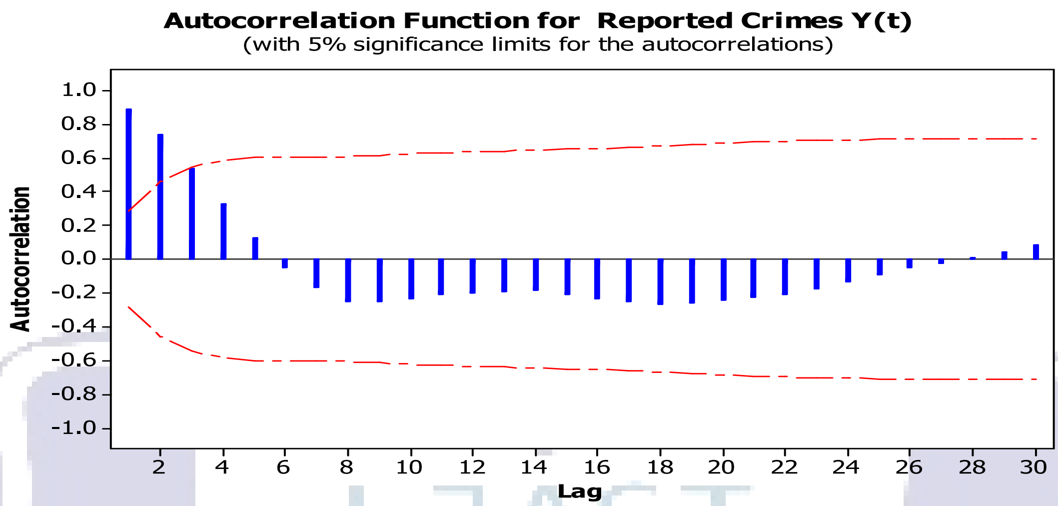


Figure 1: Time series plot for Reported crimes during 1969 to 2019

Therefore, if we imagine a straight line in the middle of the data, and one of the observations is above the line, the following observation tends to lie above the line as well, and vice versa. We also observe a scattering change in the series, and thus the series is not static in the variance. To verify this, the estimated autocorrelation function of the series was calculated and plotted in Figure 2. It is noted that the estimated autocorrelation coefficients slowly fade first to zero, but soon they appear again and their positive values increase with time, which indicates that the original series does not remain stationary.



Log	Car	T	LBQ	Log	Car	T	LBQ
1	0.89	6.28	41.88	16	-0.23	-0.72	124.82
2	0.73	3.24	71.12	17	-0.25	-0.77	129.85
3	0.53	1.97	86.79	18	-0.27	-0.79	135.59
4	0.33	1.13	92.88	19	-0.26	-0.77	141.21
5	0.12	0.41	93.75	20	-0.24	-0.71	146.32
6	-0.05	-0.18	93.98	21	-0.23	-0.66	150.95
7	-0.16	-0.55	95.56	22	-0.21	-0.59	154.90
8	-0.25	-0.83	99.44	23	-0.18	-0.51	157.94
9	-0.25	-0.82	103.38	24	-0.13	-0.38	159.68
10	-0.23	-0.75	106.91	25	-0.09	-0.27	160.63
11	-0.21	-0.67	109.87	26	-0.05	-0.15	160.91
12	-0.20	-0.64	112.61	27	-0.02	-0.06	160.97
13	-0.19	-0.59	115.14	28	0.01	0.03	160.98
14	-0.18	-0.57	117.52	29	0.04	0.12	161.19
15	-0.21	-0.64	120.65	30	0.09	0.24	162.15

Figure 2: Autocorrelation function graph of the reported crimes series y_t

Further, taking the first order difference of natural logarithm of the original RC data to eliminate its non-stationary and obtaining the DLRC sequence

$$Z_t = \log y_t - \log y_{t-1} \quad ; t = 2,3,\dots,n. \tag{1}$$

Figure 3 shows the pattern in which the DLRC sequence develops during the period under study. It is noted that, the Z_t sequence looks is stationary. The autocorrelation function, shown in Figure 4, confirms this fact, as it appears to be fading away quickly, thus being the smallest order of differences necessary to static the series is one ($d = 1$).

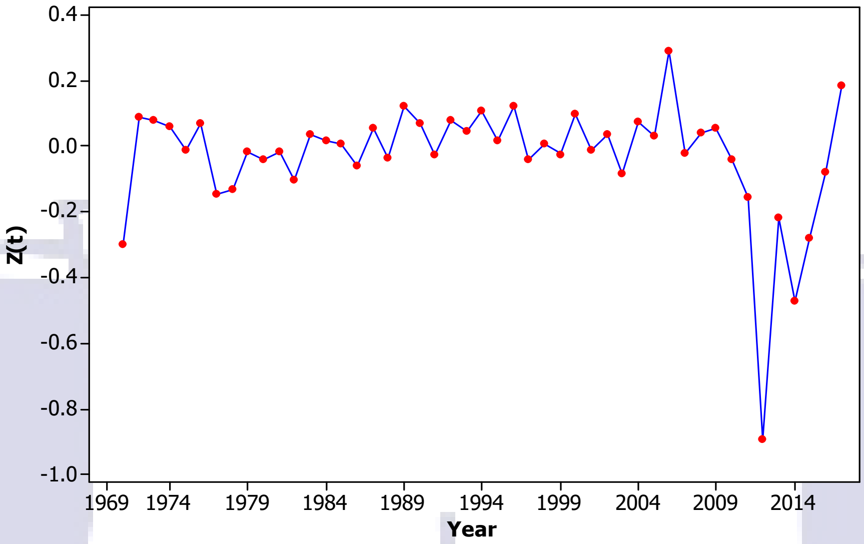
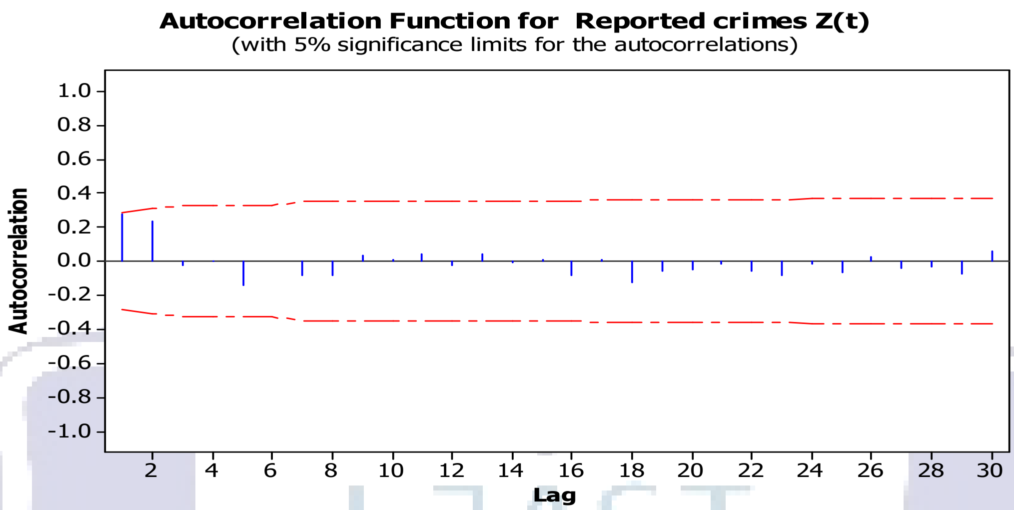


Figure 3: Time series plot for the first order difference of natural logarithm of the original RC data during 1969 to 2019

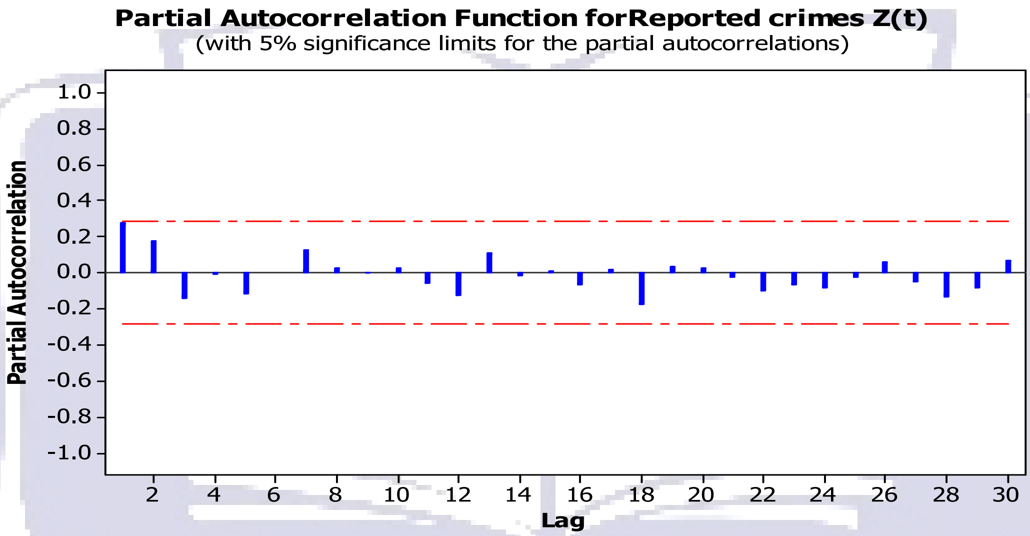


Log	Car	T	LBQ	Log	Car	T	LBQ
1	0.27	1.91	3.86	16	-0.08	-0.47	15.09
2	0.23	1.53	6.78	17	0.01	0.03	15.09
3	-0.02	-0.15	6.81	18	-0.13	-0.72	16.42
4	-0.003	-0.02	6.81	19	-0.05	-0.31	16.67
5	-0.14	-0.88	7.95	20	-0.05	-0.29	16.92
6	-0.31	-1.87	13.36	21	-0.02	-0.09	16.94
7	-0.08	-0.49	13.79	22	-0.06	-0.32	17.25
8	-0.08	-0.47	14.19	23	-0.08	-0.47	17.93
9	0.04	0.21	14.28	24	-0.02	-0.11	17.97
10	0.01	0.05	14.29	25	-0.06	-0.35	18.38
11	0.04	0.22	14.39	26	0.02	0.12	18.43
12	-0.02	-0.13	14.42	27	-0.04	-0.23	18.63
13	0.04	0.24	14.55	28	-0.03	-0.17	18.74
14	-0.01	-0.05	14.55	29	-0.08	-0.43	19.50
15	0.01	0.07	14.57	30	0.06	0.33	19.99

Figure 4: Autocorrelation function graph of the reported crimes series Z_t.

5. Model Identification

After determining the smallest order of differences that can achieve the sequence's static, we identify the order of the autoregressive p and the order of the moving averages q. Since, the methodology of Box and Jenkins depends on estimating these two orders values basic on the autocorrelation and the partial autocorrelation functions. It can be seen from Figure 4 that the autocorrelation coefficient of the DLRC sequence is significantly non-zero when the lag order is 1. It is basically in the confidence band when the lag order is greater than 1, so q can be taken 1. In Figure 5, the partial autocorrelation coefficient is significantly non-zero when the lag order is equal to 1, so p=1 can be considered.



Log	PAC	T	Log	PAC	T
1	0.27	1.91	15.09	-0.07	-0.48
2	0.17	1.21	15.09	0.01	0.09
3	-0.14	-0.97	16.42	-0.18	-1.24
4	-0.01	-0.06	16.67	0.03	0.20
5	-0.12	-0.81	16.92	0.03	0.18
6	-0.28	-1.98	16.94	-0.03	-0.18
7	0.12	0.85	17.25	-0.09	-0.69
8	0.02	0.15	17.93	-0.06	-0.44
9	0.00	0.00	17.97	-0.09	-0.59
10	0.02	0.17	18.38	-0.03	-0.18
11	-0.06	-0.39	18.43	0.06	0.43
12	-0.13	-0.89	18.63	-0.05	-0.34
13	0.11	0.74	18.74	-0.14	-0.95
14	-0.02	-0.12	19.50	-0.08	-0.57
15	0.01	0.05	19.99	0.07	0.47

Figure 5: Partial autocorrelation function graph of the reported crimes series Z_t

6. Model establishment and inspection

The estimated results with the ARIMA model are as follows:

Table 2: Estimation results of the ARIMA(1,1,1) model

ARIMA Model for Z_t				
Estimates at each iteration				
Iteration	SSE	Parameters		
0	0.846	0.100	0.100	0.092
1	0.311	-0.050	0.250	-0.007
2	0.296	-0.200	0.184	-0.002
3	0.289	-0.350	0.114	0.002
4	0.288	-0.430	0.067	0.003
5	0.288	-0.455	0.045	0.003
6	0.288	-0.465	0.034	0.003
Unable to reduce sum of squares any further				
Final Estimates of Parameters				
Type	Coef	SE Coef	T	P
AR(1)	-0.473	0.273	-1.73	0.090
MA(1)	0.024	0.308	0.08	0.938
Constant	0.003	0.011	0.28	0.779
Differencing: 1 regular difference				
Number of observations: Original series 49, after differencing 48				
Residuals: SS = 0.283595 (back forecasts excluded) MS = 0.006302 DF = 45				
Modified Box-Pierce (Ljung-Box) Chi-Square statistic				
Lag	12	24	36	48
Chi-Square	12.1	15.3	21.9	
DF	9	21	33	
P-Value	0.210	0.808	0.929	

Table 2 shows the final estimates of parameters of the model ARIMA (1, 1, 1). It can be seen from the t statistic of the model coefficients and its P-value that the parameter estimates of all explanatory variables of the model are not significant at the significance level of 0.05. Therefore, the final model of the DLRC sequence is ARIMA(1,1 ,0), as shown in table 3.

Table 3: Estimation results of the ARIMA(1,1,0) model

ARIMA Model for Z_t				
Estimates at each iteration				
Iteration	SSE	Parameters		
0	0.792	0.100	0.100	0.092
1	0.498	-0.050	0.250	0.058
2	0.357	-0.200	0.184	0.032
3	0.299	-0.350	0.114	0.012
4	0.289	-0.462	0.067	0.003
5	0.288	-0.486	0.045	0.003
6	0.288	-0.491	0.034	0.003
Relative change in each estimate less than 0.0010				
Final Estimates of Parameters				
Type	Coef	SE Coef	T	P
AR(1)	-0.4923	0.1316	-3.74	0.001
Constant	0.00321	0.0113	0.28	0.028
Differencing: 1 regular difference				
Number of observations: Original series 49, after differencing 48				
Residuals: SS = 0.283729 (back forecasts excluded) MS = 0.006168 DF = 46				
Modified Box-Pierce (Ljung-Box) Chi-Square statistic				
Lag	12	24	36	48
Chi-Square	12.0	15.3	22.0	
DF	10	22	34	
P-Value	0.284	0.849	0.945	

Therefore, the specific form of the model is

$$\Delta \text{LRC} = 0.00321 - 0.4923 + 0.388\ \text{LRC}_{t-1} \quad (2)$$

and the result is shown in figure 6.

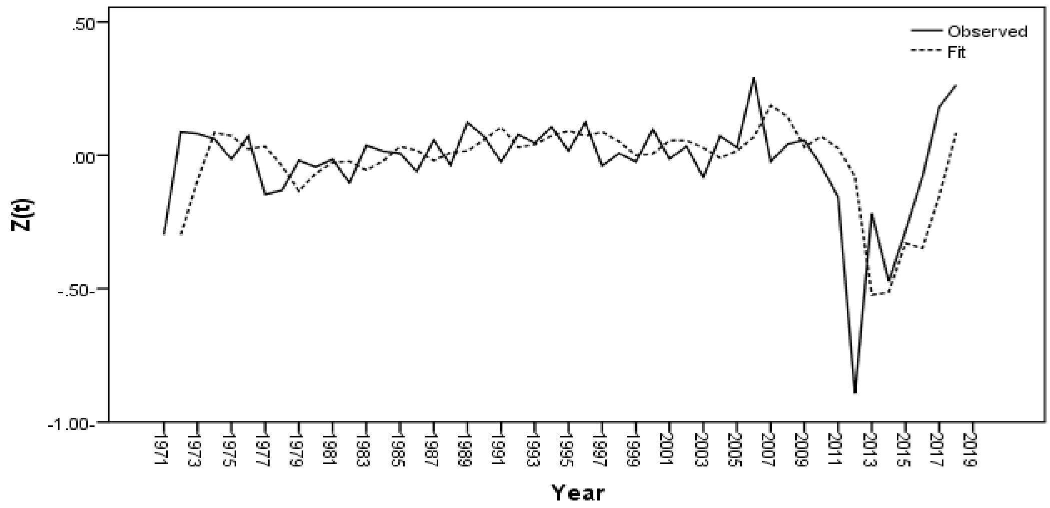


Figure 6: Actual series, fitted series of the DLRC sequence**.**

The autocorrelation function graph of the residual series is shown in Figure 7. It can be seen that the residual is a purely random errors, indicating that the model is valid.

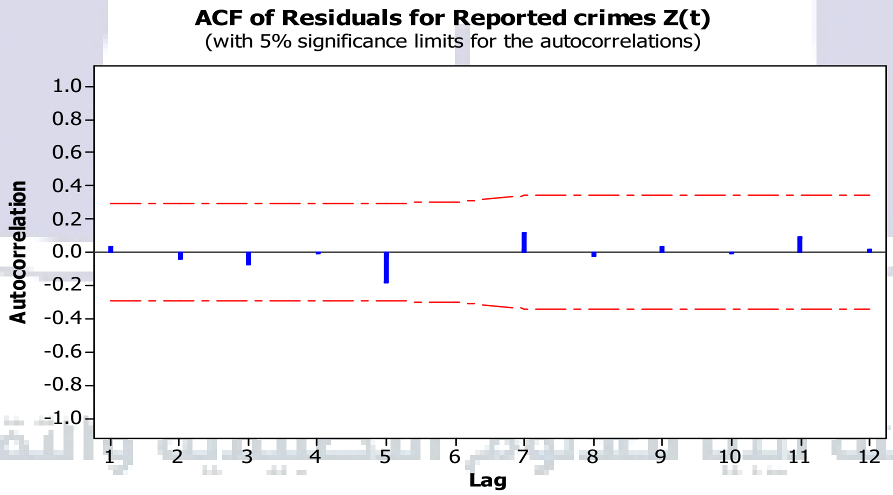


Figure 7: ACF of Residuals for Reported crimes Z(t)

7. Data forecasting

The model is used to analyze the fitting effect with the RC value in 2019. The forecast value in 2019 is 16201. The actual value is 14151. It can be seen that the forecast value is close to the actual result, indicating that the model has a good fitting effect. The mode is used to predict the RC values from 2020 to 2025. The results are listed in Table 7.

Table 7: RC forecast from 2020 to 2025.

Year	Forecasts	Lower Bound Confidence Interval	Upper Bound Confidence Interval
2020	14902.08	12775.61	17382.49
2021	15233.1	10988.35	21117.58
2022	15851.94	9289.899	27049.18
2023	16404.16	7643.536	35205.74
2024	17076.99	6180.305	47185.98
2025	17782.36	4901.063	64519.13

8. Summary

In this research, ARIMA models were then used to predict the number of the reported crimes in Libya, where two ARIMA models were proposed and their compatibility with the data was verified, namely ARIMA(1,1,1) and ARIMA(1,1,0). The results showed the suitability of the second model for estimating the number of the reported crimes in Libya, it was found that the second model ARIMA(1,1,0) is better. So this model was adopted for prediction. All results indicate the ability of our model to predict the number of the reported crimes from 2020 to 2025.

Acknowledgments

I would like to thank the Bureau of Statistics in Ministry of the Interior, the Criminal Investigation Service of Libya for providing data on the reported crimes from 1969 to 2019.

References

1. Abdullah L. 2014. Int J Adv Appl Sc 1, 153-8
2. Almohammad S, Jasem I, Lubboss M. 2018. Applying ARIMA Models for Forecasting the Production of Cotton Crop in Syria. Syrian Journal of Agricultural Research ;

5(1): 39 – 51.

3. Al Shafi MA, Blhaj A. 2020. The use of ARIMA models to predict the amount of production in the 2nd steel plant of the Libyan Iron and Steel Company in Misurata. Journal of Science, Misurata University; 10: 86 – 89.
4. Box GEP., Jenkins GM. 1976. Time Series Analysis, Forecasting and Control. Holden – Day, San Francisco.
5. Ediger SA. 2006. Energy Policy 35, 1-8
6. Jenkins B, George EP, Gwilym M, Reinsel GC. 1994. Time Series Analysis: Forecasting and Control (New Jerse: Prentice Hall, Inc)
7. Mandala QS. 1977. Econometrics. McGraw- Hill, International book Company, New York 1977. Pp 516.
8. Widowati, Putro SP, Koshioc S, Oktaferdian V. 2016. 2nd International Symposium on Aquatic Products Processing and Health (ISAPPROSH) Aquatic Procedia 7, 277 – 84
9. Zhang PG. 2003. Neurocomputing 50, 150-75