Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho
Vol. 11, No. 1, pp. 134– 138, January-March 2023

journals.uoz.edu.krd

p-ISSN: 2663-628X
e-ISSN: 2663-6298

FORECASTING THE RATIO OF THE RURAL POPULATION IN IRAQ USING BOX-JENKINS METHODOLOGY

Qais M. Abdulqader [a,*]
[a] Technical College of Petroleum and Mineral Sciences\Zakho, Duhok Polytechnic University, Zakho, Kurdistan Region, Iraq – (qais.mustafa@dpu.edu.krd)

Received: *18 Jan., 2023 /* ***Accepted:*** *18 Feb., 2023 /* ***Published:*** *20 Feb., 2023* ***https://doi.org/10.25271/sjuoz.2022.11.1.1124***

ABSTRACT:
In this paper, the Box-Jenkins methodology has been applied and used to forecast the ratio of Iraq's rural population from 1960 to 2019. A sample size of (60) observations of the annually rural population of Iraq has been taken. A combination of some adequate time series models has been prepared and obtained and some statistical criteria have been used for comparison and model selection. Results of the study concluded that the ARIMA (0,2,1) is an adequate and best model to be used for forecasting the annual ratio of rural population data in Iraq. During the period 2020 to 2030, the ratio of the rural population will keep decreasing gradually, and the percentage of the rural population of Iraq in 2030 will be (27.732).

KEYWORDS: Forecasting, Box-Jenkins, Rural population, ARIMA.

1. INTRODUCTION

Population growth is a major problem for any country, and has a significant impact on its economic and social development. This has huge implications for government plans, economic growth, and social welfare. Population forecasting helps you understand the future trend of economic growth and better allocate existing resources. Thus, the population forecasting procedure can help control the increase and decrease in the population. It can provide plentiful labor resources for all domains of life and prevent unwanted effects on the population, such as aging and impairment [1].
Many researchers have done a lot of papers works on the analysis and forecasting of population data. In [2], the authors made an empirical study to forecast the population of Pakistan from 1951 to 2007 by applying the Box-Jenkins methodology. Results concluded that the ARIMA (1,2,0) is a suitable model and can be applied for forecasting for the next 20 years. The author in [3], applied the Box-Jenkins methodology to forecast the census data of Iraq. The study concluded the continuous increasing trend that the ARIMA (2,2,0) is the best model to be used for forecasting the population of Iraq during the period 2011 to 2020. Authors in [4], used a mathematical model for dynamic population growth to obtain a prediction model for the rural population in Shaanxi province. The study results showed that, for the coming 15 years, the ratio of the rural elderly population in Shaanxi Province will get larger and grow, the ratio of the old-age dependency will increase too, and the ratio of the labor force will decrease. In reference [5], the authors used structural population data for regions, states, and provinces from the Population Census of China. A model of population, development, and environment was used to explain the population change criteria. The analysis concluded that the population of China is expected to increase first and then decrease from 2010 to 2050. In recent years, the authors in [6] proposed the ARIMA model for forecasting the population change in China in the next few years depending on the feature extraction and the analysis of the principal components. The analysis showed a downward trend in the population. In [7], a study was carried out using the logistic model and least square model to do a comparative study for making predictions of population growth between Bangladesh and India at the end of the 21st century. It was found that the projection data from 2000 to 2020 using the established models closely match the real data. In [8], the author compared three methods for forecasting the population of Iraq consisting Markov chains, artificial neural networks, and ARIMA methodology based on some statistical criteria and using real data of the population of Iraq during the period 1977-2007. The ARIMA (1,1,1) model was selected from the comparison as the best one for predictions. based on the specified model, The population of Iraq was forecasted during the period 2008 to 2030. An objective strategy for forecasting multi-regional population growth was put forth by the researchers in [9]. The findings indicate that this work can continue to serve as a neutral reference for urban and regional planning. The ARIMA Model was used by the researchers in [10] to predict the growth of the urban population in the Philippines. It has been demonstrated that the ARIMA(20,1,10) model is the most accurate for predicting the rise of the nation's urban population. In [11], the author used the Malthusian model, Unary linear regression model, Logistic model, and Gray prediction model to offer forecasts of the populations of 210 prefecture-level cities. According to the findings, there is a growing demographic disparity between cities, and the overall urban population tends to increase in middle-tier cities while decreasing in high-tier and low-tier cities.
The research aims first, to review time-series forecasting methods through Box-Jenkins models. Second, to test the possibility of applying the ARIMA method in predicting the percentage of Iraqi citizens who are in the countryside. Third, to determine the optimal model among the ARIMA models for forecasting the proportion of the rural population. Finally, to forecast the proportion of the rural population until 2030. The rest of the study is organized as follows. In section 2, the methodology of the study will be discussed. In section 3, the application of Box-Jenkins on the real data will be done. Finally, in section 4, some conclusions and recommendations will be present.

2. BOX-JENKINS METHODOLOGY

The Box-Jenkins analysis technique refers to an analyzing method of identifying the model, estimating, diagnostic, and using it to forecast by using integrated autoregressive, moving average (ARIMA) time series models. The procedure is suitable for medium to long length time series (i.e; not less than 50

observations). The general equation model ARIMA (P, I, Q) model for y can be presented as [12]:

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \quad (1)$$

Here $\emptyset$ and θ terms indicates the parameters to be find and a is a normal error and the value of its mean is zero and its distribution is independent and identical, P is a full behind value number of y_t it refers to the rank of the dimensions of autoregressive (AR), I represent the number of differences of w, and Q represents the number of full behind values of the error expressions showing the rank of the moving average (MA) dimension of the specified model. The term integrated indicates that to get a forecast for y from the determined model it is required to integrate the forecast y_t.

It must be taken into consideration that the initial identification model imposes us to temporary consideration of ARIMA models that will be efficiently suited and checked [9]. The Box-Jenkins supposes stationarity of the sequential data. A series is said to be strictly stationary when satisfying the two conditions, a fixed mean and variance as a first condition, and the existence of a constant auto-covariance structure as a second condition. If the second condition is not satisfied, then the series has weak stationarity (also called second-order stationarity) [13]. In time series analysis, if the values of the autocorrelations start high and decline quietly, then the series is said to be non-stationary, and the methodology of Box-Jenkins induce taking difference for once or more if needed to gain stationarity. Also, the variance of the inaccuracy of the specified model should be constant which indicates that the variance of values is identical for each subgroup and it is not depending on the time point or time level. When this condition is not met, then a suitable transformation using the Box-Cox test is required to solve un-stabilizing the variance [14].

A. Identification

The identification step for building and selecting the suitable values (P, Q) of the specified ARMA model for understanding the stationarity of time series y_t is executed on the grounds of its properties, including the mean, the function of autocorrelation ACF and the function of partial autocorrelation PACF. For autoregressive process AR, the values pattern of ACF are infinite (i;e. exponential and/or sine-cosine wave decay). While PACF is finite (i;e. cut off at lag P). For moving average process MA, the values pattern of ACF and PACF are the opposite of the AR process. For the ARMA process, the ACF and PACF pattern of their values are infinite (i;e. exponential and/or sine-cosine wave decay) [15].

B. Estimation

In addition to the need for stability in the Box-Jenkins models, it also has to be capable of invertibility. This means that the latest data are more steadily metric than more remote data; the parameters used in the model turn down from the most recent data down to the further past data [16]. Many techniques of estimation have been used in time series analysis such as backcasting, cross-validation, and out-of-sample procedure. Other important estimation methods for building and fitting models using the Box-Jenkins methodology are ordinary least square, maximum likelihood, non-linear estimation, and moments method [12],[17], [18]. To obtain the perfect and best model from a combination of some adequate models, some criteria measure of accuracy and goodness of fit can be used. In this paper, we depend on two important statistics: the Root Mean Square Error RMSE which has the characteristic of being easy handling mathematically, and the Akaike Information Criterion AIC which represents a formula of the variance of the model residuals, disciplined by the number of parameters estimated. The appropriate model is selected based on the lowest value of these criteria [19],[20].

C. Diagnosing

A selected model should be studied cautiously to test the verification of model adequacy. If the adequacy of the model is confirmed, then the residual series should behave as white noise. The residuals of both ACF and PACF can be used to check the proximity of at to white noise. The work is done by studying the autocorrelation scheme of the residuals to consider if there are additional large correlation values that exist. If all the ACF and PACF values are small, then the model is mentioned to be adequate and forecasts are generated. Otherwise, the values of P and/or Q should be adjusted and the re-estimation of the model is needed [21].

It is important to mention for a statistic test to detect autocorrelation or test the combined hypothesis that all m of the r_k coefficients of the correlation is jointly equal to zero using the Box-Pierce test which was introduced and developed by [22]. The test is a way to check for the absence of sequential autocorrelation, up to a specified lag k. The formula can be showed as:

$$Q = T \sum_{k=1}^{m} \hat{r}_k{}^2 \quad (2)$$

Where T represents sample size and m shows maximum lag length. It is be able to say that the test is approximately Chi-square distribution under the null hypothesis that all values of m autocorrelation coefficients are zero [23].

D. Forecasting

Once a suitable model has been chosen depending on the scientific Box-Jenkins technique and its parameters have been estimated successfully, the model has a right to be used and make forecasts. The efficiency of the specified model can only be completely rated after the real data for the forecast period have become obtainable [20]. To achieve successful outcomes, we used Statgraphics 18 for data analysis forecasting.

3. APPLICATION

The data used in this study consist of 60 values of the rural population of Iraq over the period (1960-2019). Data are taken from (macrotrends website [24]). To analyze and predict the total population of the rural area of Iraq over the next ten years (assuming that the decreasing of rural population data is affected only by time, and it is free from outside intervention). Table 1 shows the rural population data during the time interval 1960 to 2019 and the selected data from 1960 to 2015 is for processing and analysis, while the remaining values of 2016 to 2019 were allocated for verification. By using the Statgraphics statistical software, a broken line diagram of the rural population of Iraq from 1960 to 2015, is shown in Figure 1.

Table 1: The rural population of Iraq from 1960 to 2019 in percentage.

Year	Pop.	Year	Pop.	Year	Pop.
1960	57.101	1975	38.621	1990	30.294
1961	55.568	1976	37.602	1991	30.478
1962	54.022	1977	36.596	1992	30.663
1963	52.468	1978	35.822	1993	30.848
1964	50.907	1979	35.148	1994	31.034
1965	49.349	1980	34.479	1995	31.22
1966	48.132	1981	33.817	1996	31.407
1967	47.055	1982	33.161	1997	31.595
1968	45.98	1983	32.512	1998	31.612
1969	44.911	1984	31.868	-	-
1970	43.846	1985	31.233	-	-
1971	42.786	1986	30.604	-	-
1972	41.731	1987	29.983	2017	29.722
1973	40.687	1988	29.928	2018	29.527
1974	39.649	1989	30.111	2019	29.322

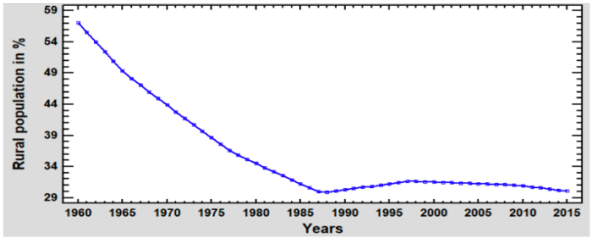

Fig.1. Annual variation of rural population in Iraq (% of total population)

From Fig.1., it can be seen that the series contains a long-term trend of curve decrement from 1960 to 1988, and from 1989 to 2002 the slight increasing events of the rural population ratio have appeared. From 2003 till 2015, the curve tended to decrease gradually which is preliminarily conclude with the non-stationary of the sequence. The two functions so-called the Autocorrelation Function ACF and the Partial Autocorrelation Function PACF are two useful tools to check whether the series is stationary or not. Figure 2 shows the lags and the values of the two functions.

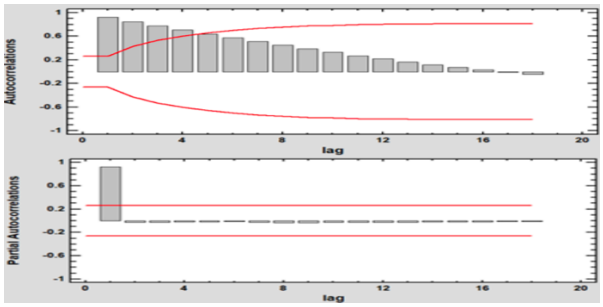

Fig.2. ACF and PACF of the rural population in Iraq during 1960 - 2015

In fig.2., the ACF values show steadily declining from the beginning to the end, and four values are out of the confidence interval. The value of the Box-Pierce statistic with eighteen lags is equal to 269.491 (p-value = 0.00), which is highly significant. While in the case of PACF, only one spike can be seen at lag 1 which is out of the limits. On the other hand, the Box-Cox transformation value was (-0.431) and its interval was (-2.941, 2.042) which includes the value zero. This recommended that the log transformation is a suitable choice to make the series stationary in variance before taking the difference of the series. After taking the log transform on the real data and checking again the ACFs and PACFs several times, we deduced that the series should be differenced twice to get stationarity in the mean. Figures3 and 4 show the transformed series respectively.

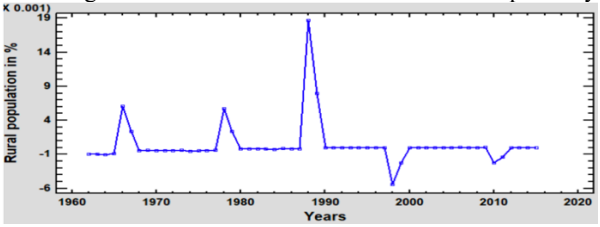

Fig.3. Log of rural population in Iraq after 2nd differencing during 1960-2015.

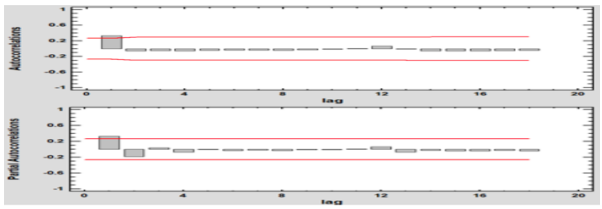

Fig.4. ACF and PACF of log of the rural population in Iraq after 2nd differencing during 1960-2015.

Fig.3. present the trend of the rural population after differencing twice the natural logarithm of the series during the specified period. Fig.3. is unusual as that of Figure 1, and its pattern is approximate stationary and has no more trend. Figure 4 and its upper part is the ACF of the 2nd difference of natural logarithm of the rural population. Except for the first lag, all the spikes at different lags are within the confidence limits. Figure 4 and its lower part is the sample PACF of the same series shows that all its spikes at different lags are inside the confidence interval except one spike at lag 1. On the other hand, the test concerning Box-Pierce takes a value of 8.78021 (p-value = 0.964619), which is greater than 0.05 meaning that the series is random. From figures 3 and 4 and the box-Pierce test, we conclude that the series is stationary, and different stationary models can be applied to this series.

After obtaining stationarity, we proceed to fit a suitable ARMA model to the adjusted series. The study uses the Root Mean Square Error RMSE as a measure of accuracy and Akaike Information Criterion AIC as a measure of the goodness of fit of the model which was mentioned in the section devoted to the theoretical side to select the best model order. Table 2 shows some combinations and adequate ARIMA with the estimated criteria.

Table 2. Some adequate ARIMA models with their criteria values

Model	RMSE	AIC	MAE
ARIMA(0,2,1)	**0.0958**	**-4.6561**	**0.0309**
ARIMA(1,2,0)	0.0975	-4.62092	0.0356
ARIMA(2,1,0)	0.0964	-4.60626	0.0455
ARIMA(0,2,2)	0.0967	-4.60281	0.0316
ARIMA(1,2,1)	0.0966	-4.60249	0.0313

It is clear from the table II that out of five specified models, the ARIMA (0,2,1) model is more appropriate to be adopted and explain the natural characteristic properties of the rural population because it has the lowest values of the RMSE, AIC and MAE comparing to the other models. Table 3. presents the estimated parameter of the ARIMA (0,2,1) model.

Table 3. Parameter estimation of ARIMA (0,2,1)

Parameter	Estimate	Standard Error	t-test	P-Value
MA (1)	-0.422137	0.125993	-3.35047	0.001493

The next step after estimating the parameter of the ARIMA (0,2,1) model is to check for randomness using the residuals graph of the ACF and PACF as shown in figure 5.

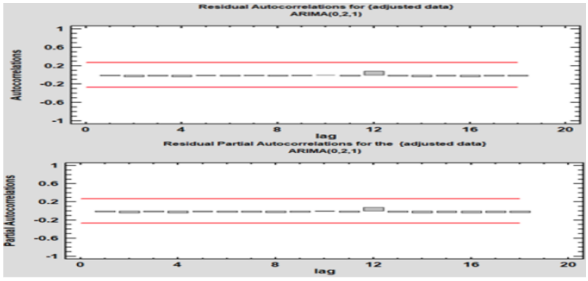

Fig.5. Residuals of the ACF and PACF for the adjusted data

From fig.5., one can see that all the autocorrelation coefficients values of ACF and PACF are non-statistically significant, concluding that the current series may well be white noise. Furthermore, the test statistics for the randomness of residuals using Box-Pierce was equal to (1.4282) and the P-value was (1.000) which exceeds 0.05. Thus, the hypothesis concerning the randomness of the series cannot be rejected at 95% or higher of the confidence level.

After proceeding with the main steps of building an ARIMA model including data preparation, model selection, parameter

estimation, and diagnostics, the last and the important step remains, which is forecasting. As we mentioned before that the rural population values of 2016 to 2019 will be allocated for verification. Table 4 presents the comparison between the forecasted values by the ARIMA (0,2,1) model with the real values for the specified period.

Table 4. The forecasted values by the ARIMA (0,2,1) model with real data during 2016-2019

Year	Real value	Forecast	Lower 95% limit	Upper 95% limit
2010	30.897	30.967	30.774	31.162
2011	30.732	30.915	30.411	31.427
2012	30.568	30.862	29.976	31.775
2013	30.405	30.810	29.483	32.197
2014	30.242	30.758	28.944	32.685
2015	30.079	30.706	28.367	33.237
2016	**29.906**	**29.917**	**29.740**	**30.094**
2017	**29.722**	**29.755**	**29.297**	**30.220**
2018	**29.527**	**29.594**	**28.790**	**30.421**
2019	**29.322**	**29.434**	**28.234**	**30.685**

From table 4, we can record some notes. First, from 2010 to 2019, all the true values and predicted values are very close to each other. Second, all these observed values fall inside the confidence interval. Third, from 2016 to 2019 the percentage of decreasing of the rural population in Iraq concerning the real values is (1.95%) while for the predicted values is (1.61%). Thus, we can say that ARIMA (0,2,1) model is the best and appropriate to be used to forecast annually rural population data, and during the period 2016 to 2030, there will be (7.3%) decrease in the rural population, and the percentage of the rural population of Iraq in 2030 would be (27.732) persons. Table V. shows the forecasted values (in percentages) of the rural population in Iraq with their confidence intervals and the Figure6 presents the forecasts for the log of the annually rural population data from the period 2016 to 2030 depending on the ARIMA (0,2,1) model.

Table 5. The forecasted values by the ARIMA (0,2,1) model during the period 2016-2030

Year	Forecast	Lower 95% limit	Upper 95% limit
2020	29.275	27.641	31.007
2021	29.117	27.016	31.381
2022	28.960	26.367	31.807
2023	28.803	25.699	32.283
2024	28.648	25.016	32.807
2025	28.493	24.322	33.380
2026	28.339	23.620	34.002
2027	28.186	22.913	34.673
2028	28.034	22.204	35.394
2029	27.882	21.496	36.166
2030	27.732	20.790	36.991

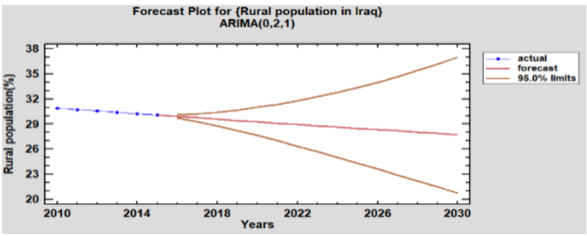

Fig.6. Forecasts for the annually rural population of Iraq from 2016 to 2030 using ARIMA (0,2,1).

4. CONCLUSION

In this paper, with the help of the Statgraphics program as statistical software, the ARIMA (0,2,1) model is determined as an adequate model according to the total rural population data of Iraq country from 1960 to 2015. To test whether the chosen parameters of the model are sensible, the rural population data of Iraq from 2016 to 2019 are validated. The analysis results show that the model has a good fitting effect because the decrease ratio of the rural population of Iraq between the real values and forecasted values during the specified time interval was close to each other and all values were placed within the confidence limits. Finally, we predict the rural population data of Iraq in the next years. the prediction results show that the rural population of Iraq is keeping decrease generally year by year. During the period 2020 to 2030, the percentage of the decrease of the rural population will be (7.3%), and the size of the rural population of Iraq in 2030 would be (27.732%).
It's important to note that the outcomes of this investigation and the forecasting model developed and selected by ARIMA (0,2,1) are remarkably comparable to those of the study carried out in accordance with [3] and the basis of which the ARIMA (0,2,2) prediction model was constructed. The previous study was used to forecast Iraq's population census, which is progressively growing, whereas the present study focuses on the country's rural population, which is steadily declining. This disparity in results between the two studies is the result of this divergence.

5. RECOMMENDATIONS

1- The selected model can be used for forecasting the rural population of Iraq in the future.
2- We recommend doing some other studies for the same data using different methods such as multiple regression analysis, wavelet, and neural network analysis to compare the results of the current ARIMA for assessment and select the best one for forecasting the rural data of Iraq.
3- It is hoped that the current study can provide an important theoretical reference for the planning issues and adjustment of relevant policies in the rural population of Iraq.

REFERENCES

[1] J. Dai and S. Chen, "The application of ARIMA model in forecasting population data," *J. Phys. Conf. Ser.*, vol. 1324, no. 1, 2019, doi: 10.1088/1742-6596/1324/1/012100.

[2] M. Zakria and F. Muhammad, "Forecasting the Population of Pakistan using ARIMA Model," *Pak.J.Agri.Sci*, vol. 46, no. 3, pp. 214–223, 2009.

[3] Q. Abdulqader, "Time Series Forecasting Using Arima Methodology with Application on Census Data in Iraq," *Sci. J. Univ. Zakho*, vol. 4, no. 2, pp. 258–268, 2016, doi: 10.25271/2016.4.2.116.

[4] S. Ruixia and Z. Na, "Prediction And Analysis of Rural Population in Shaanxi Province Based on Population Development Equation Model," *Eur. J. Res. Reflect. Manag. Sci.*, vol. 5, no. 1, pp. 60–69, 2017.

[5] A. Guo, X. Ding, F. Zhong, Q. Cheng, and C. Huang, "Predicting the future Chinese population using shared socioeconomic pathways, the sixth national population census, and a PDE model," *Sustain.*, vol. 11, no. 13, pp. 1–17, 2019, doi: 10.3390/su11133686.

[6] W. Li, Z. Su, and P. Guo, "A prediction model for population change using ARIMA model based on feature extraction," *J. Phys. Conf. Ser.*, vol. 1324, no. 1, 2019, doi: 10.1088/1742-6596/1324/1/012083.

[7] A. N. M. R. Karim, M. N. Uddin, M. Rana, M. U. Khandaker, M. R. I. Faruque, and S. M. Parvez, "Modeling on population growth and its adaptation: A comparative analysis between bangladesh and india," *J. Appl. Nat. Sci.*, vol. 12, no. 4, pp. 688–701, 2020, doi: 10.31018/jans.v12i4.2396.

[8] A.-J. Ramya, "A Comparison of the Markov Chains, Artificial Neural Networks, and ARIMA for Forecasting of Iraq's Population," *J. Al-Rafidain Univ. Coll. Sci.*, no. 46, pp. 25–49, 2020.

[9] C. Y. Wang, S. J. Lee, "Reginal Population Forecast and Analysis Based on Macine Learning Strategy", *Entropy*, Vol.

23, no. 6, pp. 1-12,2021, doi: 10.3390/e23060656.

[10] L. N. C. Estoque, L. M. D. Fuente, R. C. Mabborang, and M. G. Molina," Forecasting Urban Population Growth in the Philippines Using Autoregressive Integrated Moving Average (ARIMA) Model", *EPRA* International Journal of Multidisciplinary Research, Vol. 8, Issue 7,pp. 132-153, 2022, doi: 10.36713/epra10819.

[11] L. Chen, T. MU, X. Li, and J. Dong, " Population Prediction of Chinese Prefecture- Level Cities Based on Multiple Models", *Sustainability*, Vol. 14, Issue 8, pp. 1-23, 2022, doi: 10.3390/su14084844.

[12] I. M. Chakravarti, G. E. P. Box, and G. M. Jenkins, "Time Series Analysis Forecasting and Control.," *Journal of the American Statistical Association*, vol. 68, no. 342. p. 493, 1973, doi: 10.2307/2284112.

[13] R. A. Yaffee and M. McGee, *Introduction to Time Series Analysis and Forecasting with Applications of SAS and SPSS*, vol. 17, no. 2. 2001.

[14] G. E. Box and D. R. Cox, "An analysis of transformations revisited, rebutted," *J. Am. Stat. Assoc.*, vol. 77, no. 377, pp. 209–210, 1982, doi: 10.1080/01621459.1982.10477788.

[15] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting - Second Edition*. 2002.

[16] T. H. D. Ngo, "The Box-Jenkins Methodology for Time Series Models," *Proc. SAS Glob. Forum 2013 Conf.*, vol. 6, pp. 1–11, 2013, [Online]. Available: http://support.sas.com/resources/papers/proceedings13/454-2013.pdf.

[17] Robert H. Shumway and D. S. Stoffe, *Time Series Analysis and Its Applications With R Examples*, Third. Springer Science+Business Media, LLC, 2016.

[18] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating time series forecasting models: an empirical study on performance estimation methods," *Mach. Learn.*, vol. 109, no. 11, pp. 1997–2028, 2020, doi: 10.1007/s10994-020-05910-7.

[19] S. A. Salie Ayalew, "Comparison of New Approach Criteria for Estimating the Order of Autoregressive Process," *IOSR J. Math.*, vol. 1, no. 3, pp. 10–20, 2012, doi: 10.9790/5728-0131020.

[20] H. R. Makridakis S, Wheelwright SC, *Forecasting methods and applications*, Third. Jhon Wiley and Sons,INC., 1997.

[21] R. S. Tsay, *Analysis of financial time series*, Second. John Wiley & Sons, Inc., 2010.

[22] G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *J. Am. Stat. Assoc.*, vol. 65, no. 332, pp. 1509–1526, 1970, doi: 10.1080/01621459.1970.10481180.

[23] C. Brooks and S. Tsolacos, Real Estate Modelling and Forecasting, First. New York, USA: Cambridge University Press, 2010.

[24] https://www.macrotrends.net/countries/IRQ/iraq/rural-population