Comparison of Box-Jenkins models predicting Iraq's population growth rate

Dr. Meshal Harbi Odah
College of administration and Economics - AL-Muthanna University
m.algelidh@gmail.com
07807609085

Abstract

The prediction of the urban population growth rate it gives estimate the expected change in the rate of growth in the future Based on socio-economic development and geography of the population, The use of time series analysis is one of the most important statistical methods used in the study and analysis of annual, monthly and daily data To prediction the values of future random phenomena based on what happened in the past which helps in making future plans for economic development. The goal of most of the statistical population studies is to provide an approximate forecast future population. In this paper, the Box-Jenkins models were compared to predict the rate of population growth in urban areas until the year (2033) based on the criterion of the mean absolute percentage error (MAPE) to choose the best prediction model.

Keyword: Box-Jenkins, Forecasting, Time series & MAPE
1. Introduction

The prediction of the urban population growth rate is an approximation in a later period and gives the expected change in the future population. Based on socio-economic development and geography of the population. The goal of most population statistical studies is to provide an approximate future forecast of the population. The study of population data is an important field of interest to governments and international organizations; this is because of the importance of the human element in the planning and implementation of economic and social development.

Developing countries are less aware of their population and their needs. Developed countries are concerned and more aware of the current and future population distribution. It is fact no country can be on the correct scientific path and technological development in the world, there should be a population census in the right method. There is no doubt that the future population studies by predicting through the use of statistical methods that help in providing a set of data and numbers on population, which leads to the face of economic and social development as well as health problems. As many of these problems can be addressed and preparedness after the development of rational economic policies to contain the lowest possible negative effects in the future.

The Box-Jenkins methodology is one of the most important methods used to predict time series, this method does not assume any particular pattern of historical data for the series we predict, and where is selecting the appropriate model, the distributions of the self-correlation coefficients of the time series are compared with the theoretical distributions of the different models the model selection is good if the differences (residuals) between the estimated values of small historical data are distributed normally, and independent of each other.

There are many researchers addressing the Box-Jenkins methodology from them (Zakria & Muhammad, 2009) autoregressive moving average (ARIMA) model is used to model the time series data in different fields.

(Makridakis & Hibo, 1997) determine the Box Jenkins Methodology in empirical tests it is forecasting the accuracy than time series methods.
The main focus of this paper is to forecast the urban population growth rate for the next period (2033) on the basis of previous trends for model fitting and forecasting.

2. The methodology of Box – Jenkins [1].

Box-Jenkins models that methodology applied by both George Box & Gwilym Jenkins on the 1970 time series this methodology is distinct from other methods Its ability to modelling and predict random phenomena without assuming any prior model and provide comprehensive solutions for all stages of analysis of time series starting from the initial selection of the appropriate model stages of the methodology by estimating the parameters of the model, diagnosing it, and ending with predicting future observations this methodology is accompanied by some statistical tests that enable us to identify the appropriate data model and the inability to use the appropriate statistical tests to validate the model that is being built.

Among the most prominent methods of prediction Autoregressive Integrated Moving Average Models (ARIMA), this methodology is based on a consolidating between autoregressive model and the moving average model.

2.1 Autoregressive model: AR (p)

The current value of the time series \(Y_t\) it is written as a weighted sum of the previous values of the time series \(Y_{t-1}, Y_{t-2}, \ldots\) plus the current error value \(b_t\).

The formula model as the following (p):

\[
Y_t = \alpha + \eta_1 Y_{t-1} + \eta_2 Y_{t-2} + \cdots + \eta_p Y_{t-p} + b_t \quad \ldots \quad (1)
\]

Where:

\(\eta_1, \ldots, \eta_p\): Parameters of the model

\(t\): Time

\(b_t\): Random error

2.2 moving average model :MA(q)

The current value of the time series \(Y_t\) in terms of weighted sum of the previous values of the random error \(b_t, b_{t-1}, \ldots\).
The formula model as the following (q):

\[ Y_t = \alpha + b_t - \delta_1 b_{t-1} - \delta_2 b_{t-2} - \cdots - \delta_q b_{t-q} \quad \ldots \ldots (2) \]

Where:

- \( \alpha \): Constant
- \( \delta_1, \ldots, \delta_q \): Parameters of the model
- \( t \): Time

**2.3 Mixed model**

In this step the forms are mixed above (MA (P), MR (q)) for analysis of time series data.

The formula model as the following (ARMA (p,q))

\[ Y_t = \alpha + \eta_1 Y_{t-1} + \eta_2 Y_{t-2} + \cdots + \eta_p Y_{t-p} + b_t - \delta_1 b_{t-1} - \delta_2 b_{t-2} - \cdots - \delta_q b_{t-q} \quad \ldots \ldots (3) \]

Where:

- \( \alpha \): Constant
- \( \eta_1, \ldots, \eta_p \) & \( \delta_1, \ldots, \delta_q \): Parameters of the model
- \( t \): Time

**2.4 Autoregressive Integrated Moving Average (ARIMA)**

ARIMA models are the most usually used time series models these models consist of the above steps, where AR (p), MA(q) & (d) represents the differences required by time series to be stable then the model turns into (p,d,q).

**3. standard compared**

In this paper we will based on the criterion of the mean absolute percentage error (MAPE), for comparison between models whichever is more accurate in prediction. The formula as the following:

\[ MAPE = \frac{\sum_{t=1}^{n} |b_t|/Y_t}{n} \times 100 \quad \ldots \ldots (4) \]
4. Empirical data analysis

The data was collation urban population growth (annual %) in Iraq for the period (1960-2018) from the website World Bank.

Table (1) represented urban population growth (annual %) in Iraq for the period (1960-2018)

| Years | Rate | years | Rate | years | Rate |
|-------|------|-------|------|-------|------|
| 1960  | 6.03 | 1970  | 5.32 | 1980  | 3.95 |
| 1961  | 6.02 | 1971  | 5.21 | 1981  | 3.82 |
| 1962  | 6.04 | 1972  | 5.12 | 1982  | 3.71 |
| 1963  | 6.08 | 1973  | 5.03 | 1983  | 3.58 |
| 1964  | 6.14 | 1974  | 4.97 | 1984  | 3.44 |
| 1965  | 6.19 | 1975  | 4.92 | 1985  | 3.3  |
| 1966  | 5.6  | 1976  | 4.88 | 1986  | 3.16 |
| 1967  | 5.42 | 1977  | 4.8  | 1987  | 3.05 |
| 1968  | 5.45 | 1978  | 4.36 | 1988  | 2.24 |
| 1969  | 5.4  | 1979  | 4.09 | 1989  | 2.01 |
| 1970  |      |       |      |       |      |
| 1971  |      |       |      |       |      |
| 1972  |      |       |      |       |      |
| 1973  |      |       |      |       |      |
| 1974  |      |       |      |       |      |
| 1975  |      |       |      |       |      |
| 1976  |      |       |      |       |      |
| 1977  |      |       |      |       |      |
| 1978  |      |       |      |       |      |
| 1979  |      |       |      |       |      |
| 1980  |      |       |      |       |      |
| 1981  |      |       |      |       |      |
| 1982  |      |       |      |       |      |
| 1983  |      |       |      |       |      |
| 1984  |      |       |      |       |      |
| 1985  |      |       |      |       |      |
| 1986  |      |       |      |       |      |
| 1987  |      |       |      |       |      |
| 1988  |      |       |      |       |      |
| 1989  |      |       |      |       |      |
| 1990  | 2.2  | 2000  | 3.08 | 2010  | 2.79 |
| 1991  | 2.4  | 2001  | 3.05 | 2011  | 3.49 |
| 1992  | 2.56 | 2002  | 3.02 | 2012  | 3.95 |
| 1993  | 2.69 | 2003  | 2.89 | 2013  | 4.13 |
| 1994  | 2.76 | 2004  | 2.65 | 2014  | 3.94 |
| 1995  | 2.8  | 2005  | 2.36 | 2015  | 3.54 |
| 1996  | 2.82 | 2006  | 2.01 | 2016  | 3.12 |
| 1997  | 2.83 | 2007  | 1.75 | 2017  | 2.8  |
| 1998  | 3.07 | 2008  | 1.76 | 2018  | 2.59 |
| 1999  | 3.14 | 2009  | 2.12 |       |      |

When drawing the iterative curve of the time series data for the period from (1960) to (2018), represented by Figure No. (1) It is noted that the population growth in urban areas has been affected greatly during these years, as it has decreased and instability. The chain has been fluctuating due to the reasons mentioned previously.
4.1 Data analysis

When analyzing experimental data using the SPSS statistical program and by drawing Autocorrelation Function (ACF) vs. Partial Autocorrelation Function (PACF) Figure 2 and figure 3 show that time-series data can follow one of the Autoregressive Integrated Moving Average (ARIMA).
Figure (3) curve of the Partial Autocorrelation Function (PACF)

After drawing Autocorrelation Function (ACF) vs. Partial Autocorrelation Function (PACF). We start by selecting the best model from a set of proposed models for ARIMA models, based on the mean absolute percentage error.

| The Model    | MAPE |
|--------------|------|
| ARIMA(0,1,1) | 4.329|
| ARIMA(1,1,0) | 3.95 |
| ARIMA(0,1,2) | 4.029|

Through the comparison between the proposed models, it was found that the model (ARIMA (1, 1, 0)) is preferable because the criteria for differentiation for it were lower than the rest for other models. Note that the rest of the models were not mentioned and were neglected because their parameters were not significant.
4.2 ARIMA Model Parameters

Table (3) ARIMA Model Parameters

| Model  | Rate | Rate No. | Constant | AR Lag 1 | Difference |
|--------|------|----------|----------|----------|------------|
| Model_1 |      |          | 2.283    | .638     | 1          |
|        |      |          | .104     | .057     |            |
|        |      |          | .293     | 6.150    |            |
|        |      |          | .000     | .000     |            |

Table (3) Shows estimates of model parameters, and through the value of (sig = 0.000), which is less than (0.05), it turns out that the Equation model as the following

\[ Y_t = 2.283 + .638Y_{t-1} \quad \ldots \ldots (5) \]

4.3 Model Fit Statistics

Table (4) Model Fit Statistics & Ljung-Box Q

| Model       | Model Fit statistics | Ljung–Box Q(18) |
|-------------|----------------------|-----------------|
|             | RMSE | MAPE | MAE | Statistics | DF | Sig. |
| rate-Model_1 | .198 | 3.950 | .119 | 17.006 | 17 | .374 |

Table (4) shows the necessary measures for predictive accuracy tests of the model, which represent the first sector (Model Fit Statistics), and the lower the values of these measures, the more accurate the model used in the prediction, while the second sector represents a statistic (Ljung-Box Q), which tests the correlation between the sample values according to the following hypothesis:

\[ H_0: \rho_1 = \rho_2 = \cdots = \rho_K = 0 \]

If this hypothesis is significant, this means that the sample values are independent and the correlation between them is zero and this leads to the stability of the time series.
4.4 Residual analysis

![Figure (4) curve of the Residual Autocorrelation Function (ACF) vs. Residual Partial Autocorrelation Function (PACF)](image)

Drawing Residual Autocorrelation Function (ACF) vs. Residual Partial Autocorrelation Function (PACF) of the (prediction errors). The residues follow the pattern of the white noise chain (White Noise) and this means that they are independent and distribute a natural distribution with an arithmetic mean of (0) and a variance of its value ($\sigma^2$).

4.5 Forecast

| Years | Forecast population growth |
|-------|----------------------------|
| 2019  | 2.45                       |
| 2020  | 2.34                       |
Table (5) shows the predicted values of population growth in urban areas in Iraq, which starts in 2019 and ends in 2033.

| Year | Value |
|------|-------|
| 2021 | 2.27  |
| 2022 | 2.22  |
| 2023 | 2.17  |
| 2024 | 2.14  |
| 2025 | 2.11  |
| 2026 | 2.08  |
| 2027 | 2.06  |
| 2028 | 2.03  |
| 2029 | 2.02  |
| 2030 | 2     |
| 2031 | 1.98  |
| 2032 | 1.97  |
| 2033 | 1.95  |

**Figure (6)** curve of the original and estimated values

Figure (6) represents the graph of the iterative curve of the original and estimated values and the prediction values.
5. Conclusions

A standard model was built to predict the urban population growth rate in Iraq through 2033. The rate of urban population growth in Iraq during the studied period is an unstable time series because the charting of the correlation coefficients function showed that as a result of the significant decrease in the rate of population growth the best model among the proposed models is to predict the rate of urban population growth ARIMA (1, 1, 0). A decrease in the rate of urban population growth is expected in the coming years, through the proposed model.

REFERENCES

1. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
2. Jakaša, T., Andročec, I., & Sprčić, P. (2011, May). Electricity price forecasting—ARIMA model approach. In 2011 8th International Conference on the European Energy Market (EEM) (pp. 222-225). IEEE.
3. Makridakis, S., & Hibon, M. (1997). ARMA models and the Box–Jenkins methodology. Journal of Forecasting, 16(3), 147-163.
4. Odah, M. H. (2018). Asymptotic Least Squares Estimation of Tobit Regression Model. An Application in Remittances of Iraqi Immigrants in Romania.
5. Zakria, M., & Muhammad, F. (2009). Forecasting the population of Pakistan using ARIMA models. Pakistan Journal of Agricultural Sciences, 46(3), 214-223.
6. https://data.worldbank.org/country/IQ