Time Series Forecasting Using Arima Methodology with Application on Census Data in Iraq

Qais **M**ustafa **A**bdulqader
Dept. of Hospital Management, Zakho Technical Institute, Dohuk Polytechnic University, Zakho, Iraq.
(Accepted for publication: April 24, 2016)

Abstract:

In this paper, the methodology of Box-Jenkins of Autoregressive Integrated Moving Average (ARIMA) has been used for applying and forecasting the census in Iraq by taking (61) observations of the annually census from 1950 to 2010. Several adequate models of time series have been built and some of the performance criteria have been used for the purpose of comparison between models. Results of the analysis showed that the ARI(2,2) model is adequate to be used to forecast the annually census data of Iraq. During the period 2011 to 2020, there will be (33.58%) increase in the population, and the population of Iraq in 2020 would be (41358200) persons.

Keywords*: Box-Jenkins, ARIMA Models, Time Series Forecasting, Census.*

1. Introduction:

Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is the process of using a model to generate predictions (forecasts) for future events based on known past events. Box-Jenkins Forecasting methodology is a univariate version method and it is a self-projecting time series forecasting method. It has popularized and become widely known by George E. and Gwilym M. Jenkins in 1970 (George et. al, 2008).

Many applications have been done in this area. (Zakria and Mohammad, 2009), have used ARIMA models for forecasting the population of Pakistan. They showed that the estimated model ARI(1,2) are close to other researcher's finding as well as non-government organizations for future planning and projects. (Mutar and Ilias, 2010), have made a comparative forecasting work between ARIMA methodology and neural network method. They showed that ARIMA methodology has given more appropriate forecasts than those given by feed forward artificial neural network. (Tuama, 2012), used ARIMA methodology to forecast numbers of the patients malignant tumors in Anbar province. The results from the analysis showed that the proper and suitable model is integrated autoregressive model of order two ARI(2,1). (Sarpong, 2013), applied ARIMA models for modeling and forecasting maternal mortality. The results of the study showed that the ARIMA(1, 0, 2) model is adequate for forecasting quarterly maternal mortality ratios at the hospital. (Ghafil, 2013), used the Box-Jenkins models for forecasting the production of the electric power. The results of the analysis showed that The best model was ARIMA(1,0,2) than ARIMA(1,0,1) model and AR(1) from performance of predict methods.

Recently, the ARIMA methodology has been used in population forecasting studies. (Wan et. al., 2013), have used the ARIMA methodology for forecasting prison populations using sentencing and arrest data .The results from the analysis showed that although modeling suggests an uptrend in the remand prisoner population, this should be more than offset by a decrease in the sentenced prisoner population over the next months. (Pang and McElroy, 2014), used ARIMA methodology for forecasting fertility and mortality by race/ethnicity and gender. Results of the analysis are produced using fertility and mortality data dating from 1989 to 2009. For total rates, it is determined that a model without drift produces more tenable forecasts in comparison to the occasionally implausible results from the model with drift. (Brajesh and Shekhar, 2015), used statistical models to forecast the population of accidental mortality in India. The results of the study showed that on validation of models, ARIMA performed better than the damped trend exponential smoothing (DTES). This will be help for policy maker to control such type of incidence in future.

The underlying goal in this paper is to use ARIMA methodology so as find an appropriate formula when building a model from time series data of the population census in Iraq, so that the residuals are as small as possible and exhibit no pattern.

2. Methodology

Box-Jenkins analysis methodology refers to a systematic method of identifying, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models. The method is appropriate for time series of medium to long length (at least 50 observations). The general Box-Jenkins ARIMA (P,I,Q) model for w is written as (George et. al., 2008):

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

Where $\emptyset$ and θ are unknown parameters and the a are independent and identically distributed normal errors with zero mean, P is the number of lagged value of w_t, it represents the order of autoregressive (AR) dimensions, I is the number of times w is differed, and Q is the number of lagged values of the error terms representing the order of moving average (MA) dimension of the model. The term integrated means that to obtain a forecast for w from this model it is necessary to integrate the forecast w_t. ARIMA methodology may be involves in three phases: identification, estimation and testing, and application.

2.1. Identification the Model

It should be borne in mind that preliminary identification model commits us to nothing except tentative consideration of a class of ARIMA models that will later be efficiently fitted and checked (George et. al., 2008). The Box-Jenkins assumes that the time series is stationary. A series is said to be strictly stationary if it has a fixed mean, fixed constant variance, and a constant auto-covariance structure. If the last condition is not satisfied, then the series is said to be stationary in the weak sense, or second order (Yafee and McGee, 1999).

When preparing the data and testing for non-stationarity If the autocorrelation starts high and decline slowly, then the series is non-stationary, and the Box-Jenkins methodology recommend differencing one or more times to get stationarity. Also we should keep in mind that the variance of the errors of the underlying model must be invariant (i.e. constant). This means that the variance for each subgroup of data is the same and does not depend on the level or the point in time. If this is violated then one can solve this by stabilizing the variance through taking a suitable and kind of transformation by using Box-Cox test. We refer the reader to (Box and Cox, 1982) for more details.

The identification phase for choosing and building the appropriate (P,Q) values of the ARMA model for the stationary series w_t is carried out on the grounds of its characteristics, that is, the mean, the autocorrelation function(ACF), and the partial autocorrelation function (PACF).Table1 represents the autocorrelation patterns of ARMA processes. We refer the reader to (Brockwell and Davis, 2002) for more details.

Table (1): Autocorrelation patterns of ARMA processes

Process	ACF	PACF
AR(P)	Infinite: exponential and/or sine-cosine wave decay	Finite: cut off at lag p
MA(Q)	Finite: cut off at lag p	Infinite: exponential and/or sine-cosine wave decay
ARMA(P,Q)	Infinite: exponential and/or sine-cosine wave decay	Infinite: exponential and/or sine-cosine wave decay

2.2. Estimation and Testing the Model

Not only does the Box-Jenkins model have to be stationary, it also has to be invertible. Invertible means recent observations are more heavily weighted than more remote observations; the parameters used in the model decline from the most recent observations down to the further past observations (Ngo and Bros, 2013).There are many approaches of estimation to fitting Box-Jenkins models such as (George et. al., 2008, Shumway and Stoffer, 2011):

1- Ordinary least square method

2- Maximum likelihood method

3- Non-linear estimation method

4- Moments method

To select the best model from several adequate models we should use suitable criteria that deal with measures of accuracy and also with measures of goodness of fit of a model. In this paper we depend on some criteria such as (Ayalew et. al., 2012, Makridakis et. al., 1998, Polhemus, 2011):

1- Root Mean Square Error(RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} a_t^{\,2}}{n-c}} \qquad (2)$$

Where t is the time period, n is the total number of observations and c is the number of parameters in the model. The RMSE has the advantage of being easier to handle mathematically.

2- Mean Absolute Error(MAE)

$$\mathrm{MAE} = \frac{\sum_{t=1}^{n} |a_t|}{n} \qquad (3)$$

The selected model is the one with the smallest Mean Absolute Error. The MAE has the advantage of being more interpretable and is easier to explain to non-specialists.

3- Akaike Information Criterion(AIC)

$$\mathrm{AIC} = 2\mathrm{Ln}(\mathrm{RMSE}) + \frac{2C}{n} \qquad (4)$$

Where (RMSE) is the Root Mean Square Error during the estimated period, C is the number of estimated coefficients in the fitted model. Notice that the AIC is a function of the variance of the model residuals, penalized by the number of estimated parameters. In general, the model will be selected that minimizes the mean squared error without using too many coefficients.

4- Hannan-Quinn Criterion(HQC)

$$\mathrm{HQC} = 2\mathrm{Ln}(\mathrm{RMSE}) + \frac{2c\mathrm{Ln}(\mathrm{Ln}(n))}{n} \qquad (5)$$

This criterion uses a different penalty for the number of estimated parameters.

A fitted model must be examined carefully to check for possible model inadequacy. If the model is adequate, then the residual series should behave as a white noise (or independent when their distributions are normal). The ACF and PACF of the residuals can be used to check the closeness of a_t to a white noise. The procedure is by studying the autocorrelation plots of the residuals to see if further structure (large correlation values) can be found. If all the autocorrelations and partial autocorrelations are small, the model is considered adequate and forecasts are generated.If some of the autocorrelations are large, the values of P and/or Q are adjusted and the model is re-estimated (Tsay, 2002).

It is also possible to test the joint hypothesis that all m of the r_k correlation coefficients are simultaneously equal to zero using the Portmanteau test developed by Box and Pierce in 1970,and the formula can be represented as (Brooks and Tsolacos, 2010):

$$Q = n\sum_{k=1}^{m} \hat{r}_k^{\,2} \qquad (6)$$

Where n is a sample size, m is a maximum lag length. The correlation coefficients are squared so that the positive and negative coefficients do not cancel each other out. Since the sum of the squares of independent standard normal variates is itself a χ^2 variate with degrees of freedom equal to the number of

squares in the sum, it can be stated that the Portmanteau test is asymptotically distributed as a χ^2_m under the null hypothesis that all m autocorrelation coefficients are zero. As for any joint hypothesis test, only one autocorrelation coefficient needs to be statistically significant for the test to result in a rejection.

2.3. Application and Enforment

Once a model has been selected carefully and judiciously and its parameters estimated appropriately, then it can be used for application and to make forecasts and the users of the forecasts will be evaluating the pros and cons of the model as time progresses. A forecasting assignment is not complete when the model has been fitted to the known data. The performance of the model can only be properly evaluated after the data for the forecast period have become available (Makridakis et. al., 1998). The Box - Jenkins methodology can be represented and summarized through figure1.

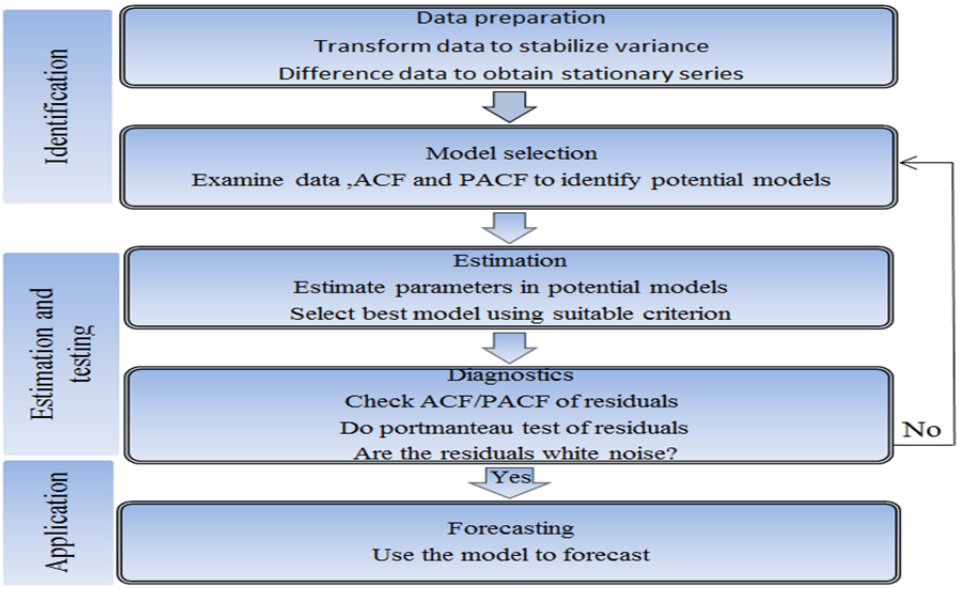

Figure (1): Schematic representation of the Box-Jenkins methodology

3. The Data

Censuses provide population numbers, household or family size and composition, and information on sex and age distribution. They often include other demographic, employment, disability, fertility, migration, education, economic and health-related topics as well. Table2 and also figure2 shows the variable used in the analysis which is the annually data of census in Iraq (in thousands) and represents a sample size (61) observations from 1950 to 2010.The first (51) observations were used for estimation and the last (10) observations were used for forecasting. The source of the data present in the web page of the United Nations Statistics Division.

Table (2): The annually data of census in Iraq (in thousands) during the period 1950-2000

Years	Population	Years	Population	Years	Population
1950	5719	1975	11685	2001	24517
1951	5902	1976	12068	2002	25238
1952	6065	1977	12461	2003	25960
1953	6216	1978	12860	3004	26674
1954	6360	1979	13258	2005	27377
1955	6503	1980	13653	2006	28064
1956	6647	1981	14045	2007	28741
1957	6795	1982	14436	2008	29430
1958	6951	1983	14823	2009	30163
1959	7116	1984	15203	2010	30962
1960	7290	1985	15576		
1961	7475	1986	15941		
1962	7674	1987	16302		
1963	7889	1988	16673		
1964	8122	1990	17074		
1965	8376	1991	17518		
1966	8651	1992	18010		
1967	8947	1993	18547		
1978	9261	1994	19124		
1970	9586	1995	19732		
1971	9918	1996	20363		
1971	10256	1997	21017		
1972	10600	1998	21694		
1973	10951	1999	22387		
1974	11312	2000	23091		

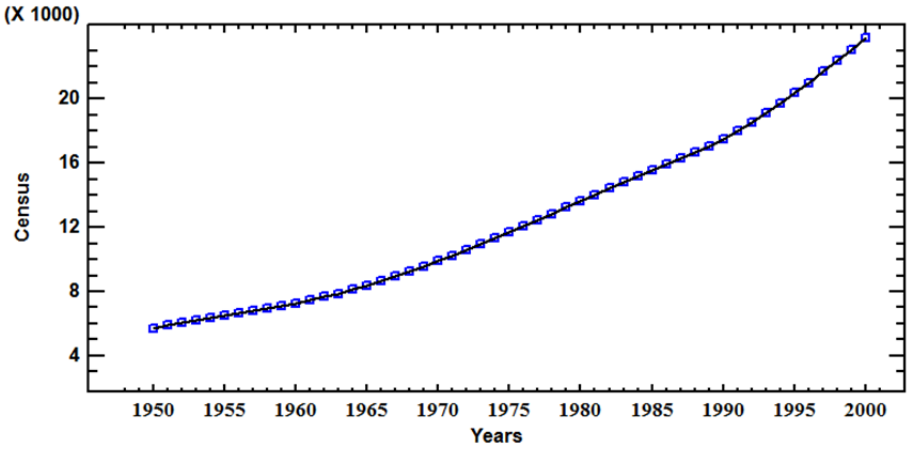

Figure (2): Census of Iraq during the period 1950-2000

Figure2 shows upward increasing trend and suggests that the given time series is non-stationary. Figure3 presents the plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) respectively.The values of the ACF are gradually declining from a first - order autocorrelation coefficient to the end. The computed Portmanteau test of Box-Pierce with seventeen lags takes a value of 261.468 (p-value = 0.00), which is highly significant, confirming the autocorrelation pattern. The partial autocorrelation function shows a large peak at lag 1 with a rapid decline thereafter, which is indicative of a highly persistent autoregressive structure in the series.

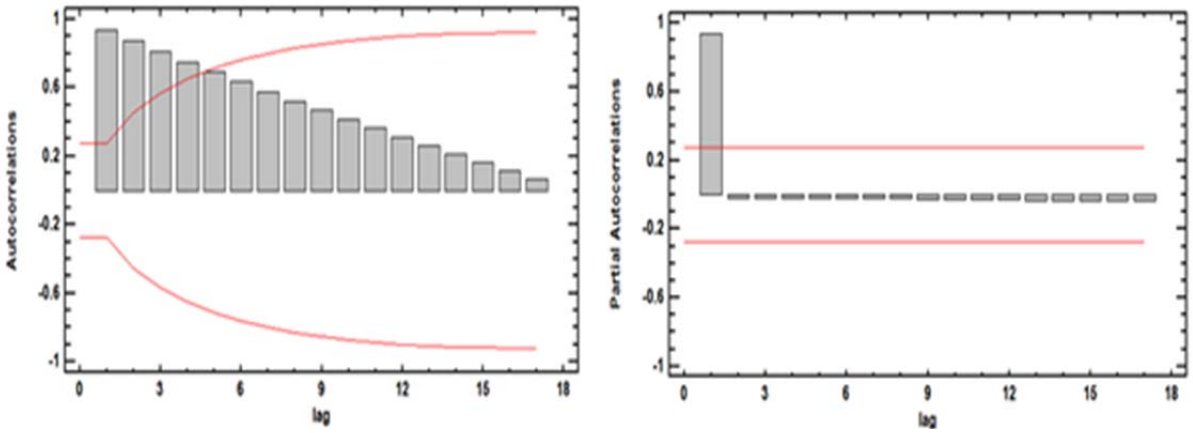

Figure (3): ACF and PACF of census of Iraq during the period 1950-2000 from left to right.

Box-Cox transformation gave the values of = 0.0 and its interval was (-0.745, 0.741) which contains the value zero. This recommended that the log transformation is appropriate choice to make our series stationary in variance before to take the difference of the series. After applying the log transformation on the original series and checking again the ACFs and PACFs, we concluded that the series need to be difference twice so as to be stationary in the mean. Figures4 and 5 represents the transformed series after 2nd differencing, ACF and PACF after 2nd differencing respectively.

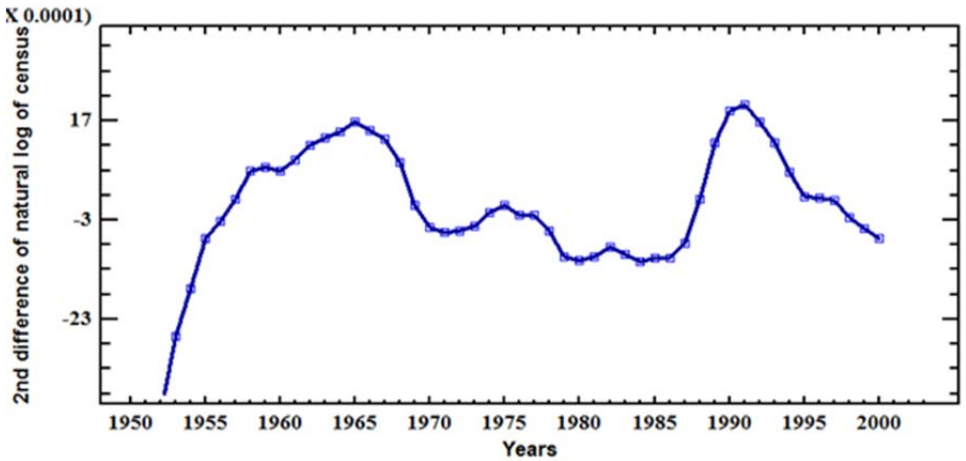

Figure (4): Log of census in Iraq after 2nd differencing 1950-2000

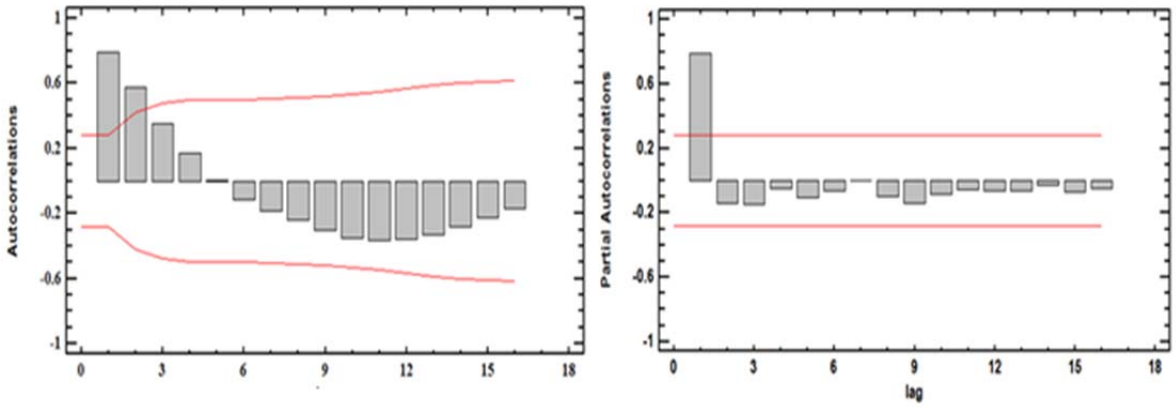

Figure (5): ACF and PACF of log of census in Iraq after 2nd differencing 1950-2000 from left to right.

After getting stationarity, we proceed to fit an ARMA model to the log of the second difference of the census of Iraq series. We apply the two measures of accuracy: RMSE and MAE with the two measures of the goodness of fit of a model AIC and HQC mentioned in theoretical part to select the appropriate model order. Table2 shows different combinations of ARIMA specifications and the estimated criteria values.

Table (3): ARIMA model comparison using criteria values

Models	RMSE	MAE	AIC	HQC
ARI (2,2)	**3.34155**	2.50926	**2.49130**	**2.52025**
ARIMA(2,2,1)	3.38335	2.51019	2.55538	2.59880
ARIMA(2,2,2)	3.40277	**2.49847**	2.60604	2.66394
ARIMA(1,2,2)	3.85454	2.71061	2.81615	2.85958
ARIMA(1,2,1)	4.46132	3.18464	3.06932	3.09827

From table 3 interestingly, both measures of the goodness of fit AIC and HQC selected ARI(2,2) having the smallest values comparing with the others. Despite the fact that AIC often tends to select higher order ARMAs, in this paper there is a consensus across the two criteria. Also, the accuracy measure RMSE selected the same model and the measure MAE selected the ARIMA(2,2,2), but its value was very close to the ARI(2,2). In general and depending on measures of accuracy, and measures of the goodness of fit of a model, the suitable and appropriate model is ARI(2,2). The estimated ARI(2,2) is presented in table4.

Table (4): Estimation of ARI(2,2)

Parameters	Estimates	Standard Error	t-ratio	P-value
AR(1)	1.703	0.067	25.557	0.000
AR(2)	-0.795	0.062	-12.735	0.000

After estimating the ARI (2, 2) model, we have to check for randomness. Figure6 shows the ACF and PACF of residuals using ARI (2, 2) on log of census of Iraq.

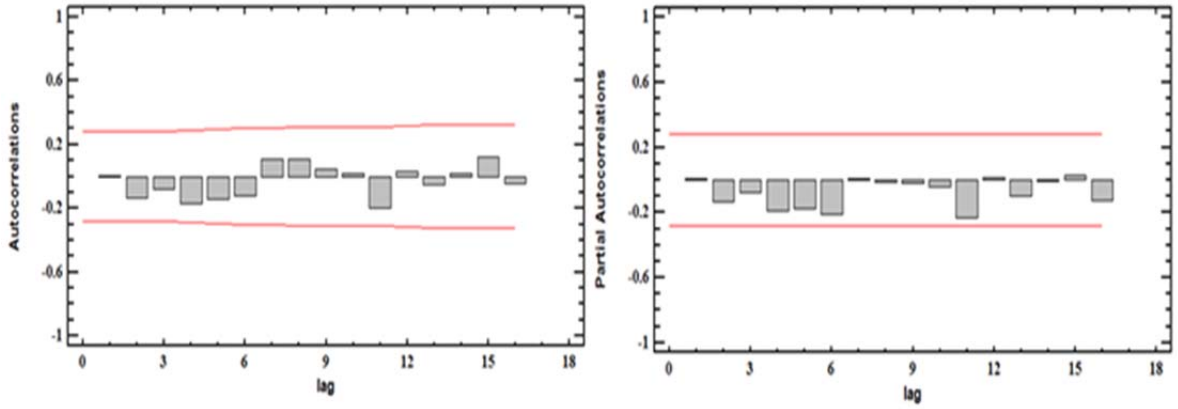

Figure (6): ACF and PACF of residuals using ARI(2,2) on log of census of Iraq from left to right.

Through looking at the figure6, we conclude that none of the autocorrelations coefficients of ACF and PACF are statistically significant, implying that the time series may well be completely random (white noise). Also, we did a test for randomness of residuals using a Portmanteau test (or Box-Pierce test), which has been mentioned in the theoretical part using the equation (6). The value of the test statistics was equal to (8.52774) and the P-value was (0.860063). Since the P-value for this test is greater than or equal to 0.05, we cannot reject the hypothesis that the series is random (white noise) at the 95% or higher confidence level.

After we checked and selected the best model which represented an ARI(2,2) of log of the annually census data of Iraq, it's time to use the model for application and make forecasts, because the main purpose of modeling a time series is to make forecasts which are then are used directly for making decisions. We validate the forecast by splitting the data in two parts: one part of the data (i;e, the first 51 observations) we used it for modeling and the other part of the data (i;e, the last10 observations) is used for forecasting. Table5 shows the forecasting for the next twenty years of the annually census (thousands) of Iraq.

Table (5): Forecast values of the annually census (thousands) of Iraq using ARI(2,2) model

Year	True value	Forecast	Lower limit 95%	Upper limit 95%
2001	24517	24513.9	24500.9	24527.0
2002	25238	25228.7	25177.3	25280.2
2003	25960	25946.0	25817.6	26075.0
2004	26674	26668.1	26412.3	26926.3
2005	27377	27398.3	26955.6	27848.2
2006	28064	28140.7	27445.5	28853.4

2007	28741	28899.5	27883.4	29952.6
2008	29430	29679.0	28273.4	31154.5
2009	30163	30482.9	28621.2	32465.7
2010	30962	31314.3	28933.8	33890.6
2011		32175.5	29218.0	35432.4
2012		33068.2	29480.3	37092.8
2013		33993.3	29726.2	38872.9
2014		34951.1	29960.0	40773.6
2015		35941.2	30184.5	42795.9
2016		36963.3	30401.4	44941.5
2017		38016.6	30611.3	47213.3
2018		39100.5	30813.9	49615.5
2019		40214.4	31008.2	52153.9
2020		41358.2	31192.8	54836.2

Looking at the table5 and using the model obtained ARI (2,2), we forecast the annually census data of Iraq from 2001 to 2020 and compared it to the first ten observed and real values from 2001 to 2010, with the statistical software Statgraphics Centurion XVI. We can see that in 2001, the predicted value (24513.9) is very close to the true value (24517) recorded and published by the United Nations Statistics Division. Also, this observed value fall inside the confidence interval, and so on for the remain values. There is something else to be mentioned, the increase of population of Iraq depending on true values during the period 2001 to 2010 was (30.08%), and depending on forecasting values was (31.57%) and the two ratios are close to each other. Hence, we can say that, ARI(2,2) model is adequate to be used to forecast annually census data of Iraq, and during the period 2011 to 2020, there will be (33.58%) increase in the population, and the population of Iraq in 2020 would be (41358200) persons. Figure7 presents the forecasts for the log of the annually census data of Iraq from the period 2001 to 2020 using ARI (2,2) model.

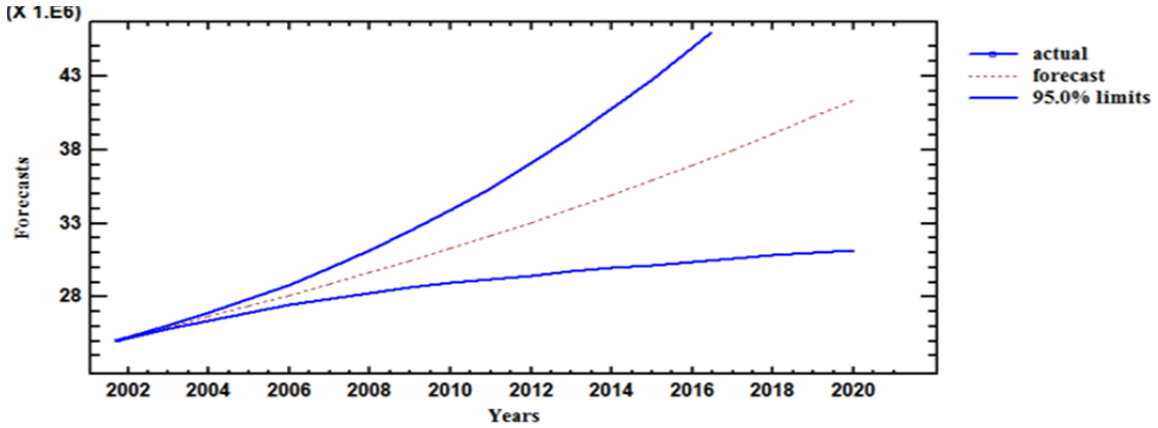

Figure (7): Forecasts for the log of the annually census of Iraq from 2001 to 2020 using ARI (2,2) model

4. Conclusions

In this paper, we have built and used a systematic methodology of Box-Jenkins ARIMA forecasting for annually census data of Iraq. Indeed, we concluded that

1- The suitable model for forecasting the census of Iraq is ARI(2,2).

2- The ratio of increase of population of Iraq of true values and forecasted values during the period 2001 to 2010 was close to each other.

3- During the period 2011 to 2020, there will be (33.58%) increase in the population, and the population of Iraq in 2020 would be (41358200) persons.

5. Recommendations

1- We may use this model for forecasting the census of Iraq for future.

2- We recommend comparing the results of ARIMA methodology for forecasting the census of Iraq with other methods like wavelet transforms or neural network method to see the differences.

References:

Ayalew S., Babu M. C. and Raw L.K. (2012). Comparison of New Approach Criteria for Estimating the Order of Autoregressive Process. *Journal of Mathematics*, 1, 10-20.

Box G.E.P. and Cox D.R. (1982). An analysis of transformations, revisited, rebutted. *Journal of American Statistical Association*, 77, 209-210.

Brajesh and Shekhar C. (2015). Accidental mortality in India: Statistical models for forecasting. *International Journal of Humanities and Social Science Invention,* 4, 35-45.

Brockwell P. J. and R. A. Davis R. A. (2002). Introduction to time series and forecasting. (2nd ed.). Springer.

Brooks C. and Tsolacos S. (2010). Real Estate Modeling and Forecasting. Cambridge university press.

George E., Jenkins G. M. and Reinsel G. C. (2008). Time series analysis: Forecasting and control, (4th ed.). John Wiley & Sons, INC.

Ghafil A. A. (2013). Using Box-Jenkins (ARIMA) models for forecasting the production of electric power. *Journal of Karbala University*, 11, 196-207.

Makridakis S., Wheelwright S. C. and R. J. Hyndman R. J. (1998). Forecasting: Methods and applications. (3rd ed.). John Wiley & Sons, INC.

Mutar D. R. and I. I. Ilias I. I. (2010). Analysis and modeling time series of water flow into Mosul city: A comparative study. Iraqi Journal of Statistical Sciences.10, 1-32.

Ngo T. H. D. and Bros W. (2013). The Box-Jenkins Methodology for Time Series Models. *Journal of statistics and data analysis*, 454, 1-13.

Pang O. and McElroy T. (2014). Forecasting Fertility and Mortality by Race/Ethnicity and Gender. *Center for Statistical Research & Methodology,* 3, 1-55.

Polhemus N. W. (2011).Time Series Analysis Using Statgraphics Centurion. StatPoint Technologies, INC.

Sarpong S. A. (2013). Modeling and Forecasting Maternal Mortality; an Application of ARIMA Models. *International Journal of Applied Science and Technology,* 3, 19-28.

Shumway R. H. and Stoffer D. S. (2011). Time series analysis and its applications. (3rd ed.). Springer.

Tsay R. S. (2002). Analysis of Financial Time Series. John Wiley & Sons, INC.

Tuama S. A.(2012). Using analysis of time series to forecast numbers of the patients Malignant Tumors in Anbar province. *Al-Anbar University Journal of Economics and Administration Sciences*, 4, 371-393.

Wan W-Y, Moffatt S., Xie Z., Corben S. and Weatherburn D. (2013). Forecasting prison populations using sentencing and arrest data. *Crime and Justice Bulletin*, 174, 1-12.

Yaffee R. and McGee M. (1999). Introduction to time series analysis and forecasting: With applications of SAS and SPSS. Academic Press.

Zakria M. and Muhammad F. (2009). Forecasting the population of Pakistan using ARIMA models. *Pakistan Journal of Agricultural Science*, 46, 214-223.

پێشبینیکرنا زنجیرێن کاتی ب کارئینانا کارناما ARIMA
دگەل جێبەجێکرن ل سەر داتایێن سەرژمێرا دانیشتوێن عیراقێ

کورتیا لێکولینێ:

د ڤێ ڤەکولینێ دا کارناما بوکس– جینکنز یا نشیفبوونا خوێەکی و ناڤبەرێن لڤلڤوك ئەوێن تەواوکار ARIMA هاتە ب کارئینان ب مەرەما جێبەجێکرن وپێشبینی کرن ب سەرژمێرا دانیشتوێن عیراقێ ئەو ژی ب وەرگرتنا (61) دانا ژ سەرژمێرا دانیشتوێن عیراقێ د ماوێ 2010-1950 دا . هندەك مودێلێن زنجیرێن کاتی ئەوێن گونجای و جوراو جور هاتنە ئاڤاکرن و هندەك پیڤەرێن بچ ئینانێ هاتنە ب کارئینان ب مەرەما بەراوردکرنێ دناڤ بەینا مودێلا دا.

ئەنجامێن شلوڤەکرنێ دیارکر کو مودێلا ARI(2,2) یا تێرا هەیە و گونجایە ژ بو ب کارئینانێ وپێشبینی کرنا سەرژمێرا سالانە یێن دانیشتوانی عیراقێ. د ماوێ 2020-2011 دا دێ زێدەبوون پەیدا بت د ژمارا کەساندا ب ریژا (33.58%) و ژمارا دانیشتوێن عیراقێ ل سالا (2020) ێ دێ بت (41358200) کەس.

التنبؤ بالسلاسل الزمنية باستخدام منهجية ARIMA
مع التطبيق على بيانات التعداد السكاني في العراق

الخلاصة :

في هذا البحث تم استخدام منهجية بوكس – جينكنز للإنحدار الذاتي والأوساط المتحركة التكاملية ARIMA من اجل التطبيق والتنبؤ بالتعداد السكاني في العراق وذلك باخذ (61) مشاهدة من التعداد السكاني السنوي خلال الفترة الزمنية 2010 – 1950. تم بناء نماذج كافية ومختلفة من السلاسل الزمنية وتم استخدام بعض مقاييس الأداء لغرض المقارنة بين النماذج .

اظهرت نتائج التحليل بان النموذج ARI(2,2) هو كافي وملائم لإستخدامه في التنبؤ بالتعداد السكاني السنوي في العراق.خلال الفترة الزمنية 2020-2011 سيكون هناك زيادة في عدد السكان بنسبة قدرها (33.58%) وعدد سكان العراق في عام (2020) سيصبح (41358200) شخص.