Forecasting Electricity Generation in Kurdistan Region Using BOX-Jenkins Model

Wasfi T. Saalih Kahwachi[1] & Samyia Khalid Hasan[2]

[1] Director of Research Center, Tishk International University, Erbil, Iraq
[2] Statistics and Informatics Department, Salahaddin University-Admin.& Economics, Erbil, Iraq
Correspondence: Wasfi T. Saalih Kahwachi, Director of Research Center, Tishk International University, Erbil, Iraq
Email: wasfi.kahwachi@tiu.edu.iq

Doi: 10.23918/eajse.v9i1p142

Abstract: The objective of this research is to identify the best and most relevant statistical model for projecting electrical power generation in the KRI. Data was collected for this purpose throughout a 168-year period (2006-2019). The Box-Jenkins technique was used, and it was discovered that the series is unstable and not random after analyzing it. The essential transformations, namely the square root and the first difference, were used to achieve stability and randomization. The necessary transformations, such as the square root and the first difference, were used to achieve stability and randomness. the analysis showed that ARIMA (2,1,2) is the most appropriate model among the proposed models using some statistical criteria like (AIC, BIC, MSE, MAPE, and RMSE) were used to obtain the model that can be utilized in the prediction. A simulation was conducted in favor to the selected model.

Keywords: Electricity Generation, Time Series, Box-Jenkins, Forecasting, Simulation

1. Introduction

Time series is regarded as an important means to create plans and strategies to study a given event across time, it allows researchers to predict what the phenomenon's future values will be.

The behavior of the series, developing the model, and forecasting are the most distinguishing features of time series compared to other statistical methods. When compared to other statistical methods, the most significant feature of time series is the applied behavior and model construction, followed by analysis and future prediction.

This research intends to analyze the behavior of electricity generation to derive the best appropriate ARIMA model using some comparison criteria, including the (MSE), (MAPE), (RMSE), (AIC), and (BIC).

2. The Theoretical Aspect

2.1 Basic Definitions

a. Time series: For different phenomena, it is defined as a set of correlated data recorded for a given occurrence over time.
b. Stability: A time series is stable if it is statistically balanced around the mean and the variance.

Received: March 10, 2022
Accepted: May 20, 2022
Kahwachi, W.T.S., & Hasan, S.K. (2023). Forecasting Electricity Generation in Kurdistan Region Using BOX-Jenkins Model. *Eurasian Journal of Science and Engineering*, *9*(1),142-151.

c. Instability: The significance of the ACF can be observed if it continues beyond the first and second displacement of a large value for a number of displacements, so it is concluded that the time series is not stable around the mean and the variance, the difference of degree (d) can be adjusted, represented by the following equation:

$$\nabla^d Z_t = (1-B)^d Z_t \quad [1]$$

Where: B is the posterior displacement factor

d. The Autocorrelation Function (ACF): It is a measure of the relation between the values of a phenomenon across time. The general formula for calculating autocorrelation for a stable chain is:

$$\rho_z(k) = \frac{\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{\sum_{t=1}^{n}(Z_t - \bar{Z})^2} = \frac{\gamma_z(k)}{\gamma_z(0)} \quad [2]$$

Where:

$\rho_z(k)$: The autocorrelation of Z-values with a k-offset.

$\gamma_z(k)$: Auto-covariance, with a displacement of k.

$\gamma_z(0) = \sigma_z^2$: Variance of Z-values.

The sum of the correlation coefficients for all displacement values is the ACF. Furthermore, any points of ACF are symmetric about $(k=0)$ that is $\rho_z(k) = \rho_z(-k)$.

e. The Partial Autocorrelation Function (PACF): This function is used to compute the degree of correlation between z_t and z_{t-k} when the influence of other values is existing, i.e., displacement values for y, and is calculated regressively using $(z_{t-1,\ldots,}z_{t-p})$. The PACF value k is the estimated value of k in multiple regression, and the equation is termed the p autoregressive equation

$$\phi_{kk} = \left|\frac{R_k^*}{R_k}\right| \quad [3]$$

Where:

R_k : Rank autocorrelation matrix.

R_k^* : Is the matrix after substituting the last vector $(p_1, p_2, \ldots, p_k')$.

3. Box -Jenkins Models

3.1 Autoregressive Model AR (p)

It is one of the most common models for time series, it produces a model of order p, denoted by the sign AR (p):

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t \quad [4]$$

This model always meets the stability criterion, and the p-order self-regression equation is a regression equation, but it varies from a conventional regression equation in that the explanatory variables reflect the prior values of the response variable z_t i.e. the variable $z_t S$ creeping values $z_t S$. As a result, this formula is regarded as autoregressive, with AR denoting the link between prior data. And current $(\phi_1, \phi_2, ..., \phi_p)$: the autoregressive parameters that characterize the change in the prior values of the time series, which are a series $(Z_{t-1}, Z_{t-2}, ..., Z_{t-p})$ of random errors with a mean of zero and a variance of magnitude (σ_a^2).

3.2 Moving Average Model (MA)

This model is of the q order, represented by MA (q), and it is written as follows:

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} ... - \theta_q a_{t-q} \quad [5]$$

After order q, the ACF is equal to 0, indicating that it breaks off after (q). The PACF reduces exponentially with time.

3.3 Autoregressive - Moving Average

If the model fulfills the requirement of invariability, which can be expressed using the MA formula or the AR formula, we can find stability. The problem in describing these models is that they require a large number of parameters. It is difficult to represent these models for high ranks in the MA or AR models, thus it is required a more generic model, which offers sufficient estimate for a large number of parameters as well as for high ranks, and this is the following ARMA regression model:

$$Z_t = \phi Z_{t-1} + ... + \phi Z_{t-p} + a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q} \quad [6]$$

It makes a stable series out of the converted differences that must be considered in order to transform it into a stable series. The degree of integration is termed an integrated series of degree d. The (ARMA) model is derived from the (ARIMA) Autoregressive Integrated Moving Model (p, d, q).

3.4 Model Diagnostic Checking

Before applying the model, it must be validated that it is true and efficient in representing the time series, which may be done by:

1. For the estimated errors (residuals), use the autocorrelation coefficients.

$$r_k(\hat{a}) = \frac{\sum_{t=1}^{n-k} \hat{a}_t \hat{a}_{t+k}}{\sum_{t=1}^{n} \hat{a}_t^2} \quad [7]$$

The $r_k(a)$ is normally distributed with a mean of zero and variance 1/n then $\phi = n\sum_{k=1}^{m} r_k^2(\hat{a})$

n: the number of observations.

We also calculate the autocorrelation of the remainder. If it falls within the confidence limits with a probability of 95%, then the model is appropriate.

$$-1.96\frac{1}{\sqrt{n}} \le r_k(a) \le +1.96\frac{1}{\sqrt{n}}$$

2. Mean Squared Error (MSE): Variance of forecast errors is measured.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Z_t - F_t)^2$$

3. Mean Absolute Percentage Error (MAPE): Absolute error is calculated as a percentage of the forecast.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Z_t - F_t}{Z_t}\right| * 100$$

4. Root Mean Square Error:

$$RMSE = \frac{\sqrt{\sum_{t=1}^{n}(a_t)^2}}{n} = \sqrt{MSE}$$

5. Akaike Information criteria (AIC):

$$AIC = nlin\hat{\sigma}^2 + 2m \qquad \hat{\sigma}^2 : \text{estimator of var(at)}$$

$$\text{m : No. of parameters}$$

6. Bayesian Information Criterion (BIC):

$$BIC = nlin\hat{\sigma}^2 + 2lin(m)$$

4 The Application Part

4.1 Introduction

This element includes a practical investigation into the development and testing of an acceptable random model for the study's data, as well as the use of that model to forecast and control the problem in research. The monthly data for the years 2004 to 2009 were collected, constructing a time series of size (168), taken from (General Directorate of Control &Communication-Kurdistan Dispatch Control Center). The real data shown in appendix (1).

4.2 Building the Model According to the Box-Jenkins Method

The first stage in studying time series is to draw the series in order to learn some of its original features throughout time, because it elucidates the nature of the oscillations and whether or not they follow a general pattern. When looking at the time series for electrical power generation in Figure No. (1), it can be noticed that the form and variations of the time series change, indicating that it is unstable.

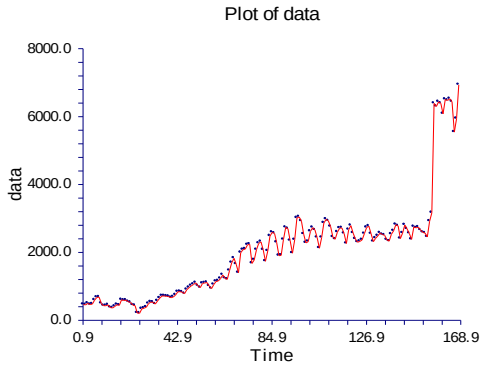

Figure 1: The Chain of Electricity Generation

4.3 Detection of Time Series Stability

Because simply glancing at a time series' graph is insufficient to establish the chain's instability, we've drawn the two functions as illustrated in Figure No (2), where it can see that the ACF values of all the series does not fall within the confidence limits. $-0.154765 \le r_k(a) \le +0.154765$, that is a chain with a high level of confidence (95%) is unstable.

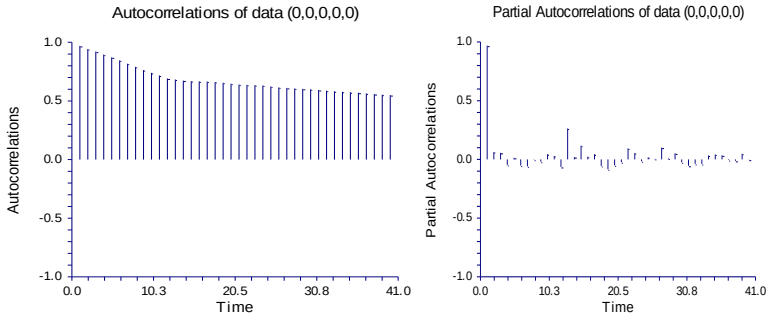

Figure 2: ACF and PACF for the Electric Power Generation

Table No. (1) shows that the series' Box-Jenkins test P-Value for randomness value is less (0.05), showing that the null hypothesis is rejected and the series is unstable.

Table 1: The P-value of the Time Series and the Assumptions

Hypothesis	P-Value
(H_0): The series is random	0.000
(H_1): The series is not random	

In order to obtain the stability of the time series, the square root, and the first difference was taken, the diagram of ACF and PACF as shown in Figure No. (4.3), Table (4.2), and the Table No. (4.3). It is clear that all the values fall within the limits of confidence at the level of confidence (95 %) $-0.154303 \le r_k(a) \le +0.154303$. In terms of the model's randomness with a P-Value of 0.32, indicating that the model was random.

Table 2: Autocorrelations of ARIMA Model (2, 1, 2)

Lag	Correlation	Lag	Correlation	Lag	Correlation	Lag	Correlation
1	0.123198	11	0.162375	21	-0.171910	31	0.062146
2	-0.151042	12	0.343156	22	-0.161371	32	-0.069786
3	-0.244879	13	0.123933	23	0.060035	33	-0.110358
4	-0.142686	14	-0.157153	24	0.326363	34	-0.138162
5	0.120247	15	-0.228779	25	0.060169	35	0.011618
6	0.188333	16	-0.057477	26	-0.204150	36	0.200711
7	0.071954	17	0.127748	27	-0.205972	37	0.075740
8	-0.127508	18	0.167834	28	-0.082516	38	-0.101275
9	-0.237406	19	0.095094	29	0.134058	39	-0.191031
10	-0.175685	20	-0.097191	30	0.196712	40	-0.064658

Table 3: Partial Autocorrelations of ARIMA model (2, 1, 2)

Lag	Correlation	Lag	Correlation	Lag	Correlation	Lag	Correlation
1	0.123198	11	0.106028	21	0.012563	31	-0.006196
2	-0.168782	12	0.216700	22	-0.091381	32	0.011828
3	-0.211479	13	0.046752	23	-0.046529	33	0.067743
4	-0.121454	14	-0.076132	24	0.166603	34	-0.042318
5	0.087789	15	-0.049485	25	-0.044464	35	-0.118461
6	0.091745	16	0.028078	26	-0.121985	36	0.000679
7	0.021837	17	0.023259	27	-0.023651	37	0.019631
8	-0.084310	18	0.000233	28	-0.022179	38	0.020186
9	-0.147408	19	0.048401	29	0.015115	39	-0.030705
10	-0.150664	20	0.005703	30	0.027826	40	-0.010760

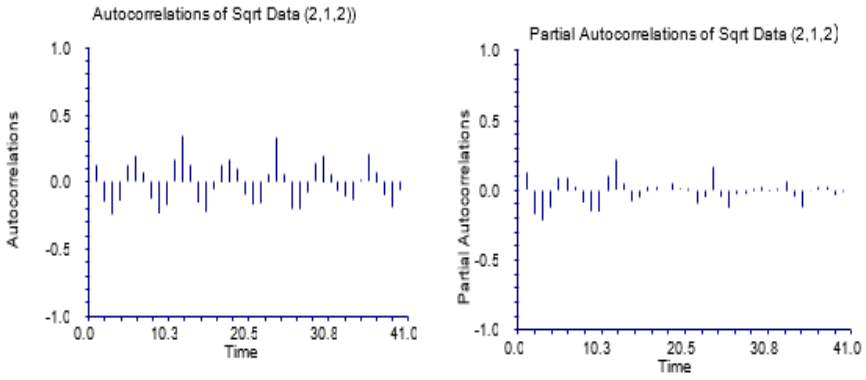

Figure 3: ACF and PACF of ARIMA Model

4.5 Choosing the Best Model for the Time Series

The ACF and PACF behavior are analyzed after acquiring a stable time series in order to distinguish the right model and its degree. However, these parameters do not always show a specific model, so a

set of appropriate models was constructed. The (AIC), (BIC), (MSE), (MAPE), and (RMSE) were obtained, as illustrated in Table No. (4). The suitable model is the ARIMA (2, 1, 2).

Table 4: The Parameter of Best Model

Measure / Model	MAPE	MSE	RMSE	AIC	BIC
ARIMA(1,1,2)	8.4919	7.856957	2.803026	354.315	366.787
ARIMA(2,1,1)	8.4720	7.71961	2.778419	349.352	363.482
ARIMA(2,1,2)	8.4636	6.575505	2.564275	326.403	336.899
ARIMA(0,1,2)	8.6028	8.030604	2.833832	355.989	370.484

4.6 Forecasting

After determining the model's appropriateness for the time series through the steps of diagnostic and evaluation, as well as assessing the model's suitability, it is used to forecast future values for the years (2020-2023) by months, as shown in Table No. (5) and Figure No. (4) respectively. The Portmanteau Test P-Value greater than 5% indicates that the forecasting was good, and the residuals has no models remaining in it.

Table 5: Shows the Suggested ARIMA (2, 1, 2) Future Values by Months.

Year / Month	2020	2021	2022	20203
1	372.4724	417.473	416.5135	683.7078
2	451.5913	351.0762	198.908	682.2776
3	530.1559	384.8478	222.839	721.2151
4	608.4308	471.0355	387.4996	759.8628
5	501.8905	513.0205	390.0129	727.9912
6	439.9122	439.9619	382.3406	712.5862
7	525.954	565.3284	461.4066	722.7201
8	649.2201	551.3273	515.2394	803.5828
9	738.1378	607.9559	551.0703	840.1553
10	574.6663	608.6122	528.2548	871.5348
11	483.7629	584.405	625.5176	847.6713
12	431.929	467.7819	665.2469	948.7057

The estimates of parameters and the standard error of the time series ARIMA Model (2, 1, 2), as well as the p-value, are shown in Table No. (6).

Table 6: the ARIMA Model's Parameters (2, 1, 2)

Parameter	Estimate	Stnd. Error	T-Value	P-Value
θ_1	1.003263	1.396581E-02	71.8371	0.000
θ_2	-0.998049	1.317045E-02	-75.7794	0.000
φ_1	0.9905713	3.603645E-02	27.4880	0.000
φ_2	-0.9274067	3.828268E-02	-24.2252	0.000

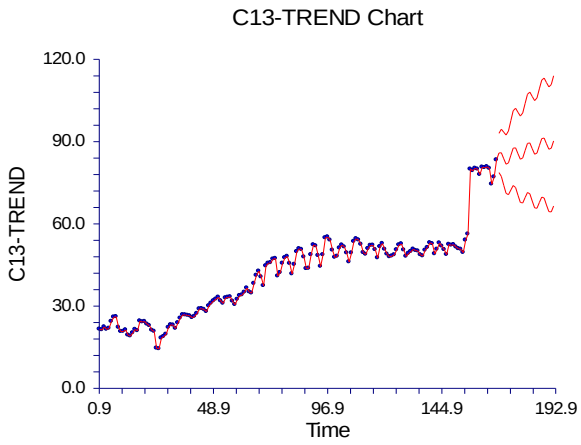

Figure 4: The Model after Transformed

5. Simulations of Data

The efficacy of the previous procedures and methods in predicting the optimal model is investigated in this work through experiments using data obtained. Given that, simulation study to evaluate the significance of the recommended model was identified, the simulation approach was utilized to the (ARIMA) model to determine the degree of the proposed model's performance for time series. A computer program has been written that generates data based on six sample sizes n=25, 50, 100, 168, 500, and 1000. The selected prognostic quality measures (MSE, MAPE, RMSE, MAE, AIC, and BIC) were computed as follows:

Table 7: The Simulation Values of the Best ARIMA Model

Measure / Sample size	MAPE	MSE	RMSE	AIC	BIC
25	129.526	298.913	17.2891	152.504	155.379
50	129.011	320.109	17.8916	298.433	294.8708
75	129.428	262.856	16.2128	427.871	424.3083
100	134.511	264.533	16.2645	567.796	564.2342
168	214.349	286.889	16.9378	960.727	564.2342
500	238.355	259.86	16.1202	2790.07	2786.509
1000	234.875	245.646	15.6731	5513.89	5510.328

Table (7) shows the results of fitting the ARIMA model (2,1,2). The sample size of 1000, has the smallest MSE and RMSE.

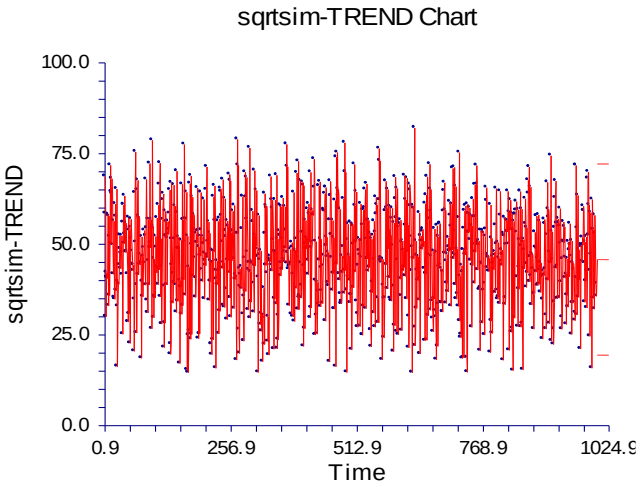

Figure 5: The Simulation of Electric Power Generation of the ARIMA (2, 1, 2)

Conclusions

1. It was found that the time series of electric power generation are not stable, so the square root around the variance and the initial difference of the series was taken to achieve stability.
2. Based on the lowest value of the criteria used (MSE, MAPE, RMSE, AIC and BIC), the ARIMA (2, 1,2) is the best forecasting model.
3. The forecasts were all good based on the values of criteria above in (2) besides the results of Portmantua test P-Value greater than 5% indicating best forecasts.
4. The results of simulation were supporting the same model (ARIMA) model (2,1,2) for the six sample sizes n=25, 50, 100, 168, 500, and 1000.
5. It was found that the probability value of the randomness tests is more than (0.05) by plotting the ACF of the residuals generated by this model, that is, those within confidence limits. This shows that this model is very good at predicting future values.

References

Adler, R.J. (1990). "An Introduction to Continuity, Extra-ma, and Related Topic for General Gaussian Processes", Lecture Notes- Monograph Series 12, Institute of Mathematical Statistics, Hayward. CA.

Atiya, A. F.; El-Shoura, S. M. Shaheen, S. I. and El-Sherif, M. S. (1999). "A Comparison between Neural-Network Forecasting Techniques—Case Study: River Flow Forecasting", *IEEE Transactions on Neural Networks*, 10(2), March.

Brown, R.G. (1963). "Smoothing Forecasting and Prediction of Disc-rete Time Series". Englewood Cliffs, Nj: Prentice-Hall.

Box, G.P. and Jenkins, G.M. (1976). "Time Series Analysis Forecasting and Control", Revised Edition Holden-Day Inc. San Fran- cisco.

Chatfield, C. (1980). "The Analysis of Time Series: An Introduction", Bath University, 2nd ed., UK.

Gershenson, Carlos. (1998). “Artificial Neural Networks for Beginners”, Sussex Academy, UK.

Handcock, M.S.and Stein, M.L. (1993)."A Bayesian Analysis of Kriging". *Technimetrics*, 35(4).

Hamilton, J. D. (1994). "Time Series Analysis", Princeton University Press, New Jersey.

Lin, Feng; Yu, Xing Huo; Gregor, Shirely and Irons, Richard. (1995). "Time Series Forecasting with Neural Networks", *Complexity International*, 02, ISSN 1320-0682, Australia.

Appendix 1: Data

https://docs.google.com/spreadsheets/d/1qyEqv60g71sigPw7CdFhHKVS5WS6HgHG/edit?usp=sharing&ouid=102631437083163869165&rtpof=true&sd=true