Efficiency of Analysis of Residuals for Selecting Time Series Models (Simulation Study)

Mohammed M Mekaeil
Faculty of Science,
University of Benghazi, Benghazi-Libya
Mohmek39@gmail.com

Suaad Ben-Farag
Faculty of Science
University of Benghazi, Benghazi-Libya
suaad.benfarag@uob.edu.ly

Abstract:

Background: The vital part of selecting the best ARIMA model that describes the historical pattern in the data is directly related to whether modelling diagnostic checks are performed well. In diagnosis checking, if the model fits well, the fundamental assumptions of the model are satisfied; the residuals should be independent, homogeneous, and normally distributed. Only if the estimated model passes all diagnostic checks should it be used for predicting and interpreting the future. Methods: When evaluating the precision of fit for an ARIMA model, statistical tests of the residuals are frequently performed to assess the residual assumption of the fitted model. As a result, tests for independence, homogeneity, and normalcy should be undertaken during diagnostic assessments. The purpose of this study is to investigate the effectiveness of diagnostic checking in selecting an optimal ARIMA model from a set of candidate models. A simulation examination was carried out in particular to analyze the likelihood of the analysis of residuals picking up the real model using statistical tests. Conclusion: Our simulation findings showed that the parsimony model was chosen among the candidate models that met all of the diagnostic checks while taking into account the significance of the model's coefficients and the minimum value of the Bayesian information criterion BIC.

Keywords: Time Series Analysis, ARIMA models; Diagnostic checks; Residual Analysis; Simulation

الملخص:

يرتبط الجزء الحيوي من اختيار أفضل نموذج ARIMA و الذي يصف النمط التاريخي في البيانات، ارتباطًا مباشرًا بما إذا كان الفحوصات التشخيصية للنموذج تكون منجزة بشكل جيد. عند الفحص التشخيصي، إذا كان النموذج مناسبًا للبيانات، يتم استيفاء الافتراضات الأساسية للنموذج, البواقي ينبغي أن تكون مستقلة ومتجانسة وموزعة توزيعا طبيعيا. النموذج المقدر

والذي يحقق كل الفحوصات التشخيصية حول البواقي سوف يستخدم للتنبؤ بالمستقبل وتفسيره. عند تقييم دقة الملاءمة النموذج المقدر لـ ARIMA ، يتم إجراء الاختبارات الإحصائية للبواقي بشكل متكرر لتقييم الافتراضات للنموذج المقدر. ونتيجة لذلك، ينبغي إجراء اختبارات (الاستقلالية والتجانس ومعرفة فيما أذا كانت البواقي تتبع التوزيع الطبيعي) أثناء عمليات التقييم التشخيصية. الغرض من هذه الدراسة هو التحقيق من فعالية الفحص التشخيصي في اختيار نموذج الأمثل لـ ARIMAمن مجموعة من النماذج المرشحة للبيانات. وأجري فحص محاكاة على وجه الخصوص لتحليل احتمالية تحليل البواقي في اختيار النموذج الحقيقي باستخدام الاختبارات الإحصائية. أظهرت نتائج المحاكاة التي توصلنا إليها أنه تم اختيار نموذج واحد فقط من بين النماذج المرشحة التي استوفت جميع الاختبارات التشخيصية مع الأخذ في الاعتبار أهمية معاملات النموذج قيمة.BIC

الكلمات الافتتاحية : تحليل السلاسل الزمنية، نمادج ARIMA ، الفحص ، تحليل البواقي ، المحاكاة.

1. Introduction

In statistics science, and in particular in the construction of statistical models for statistical inference, model validity is critically important to ensure accurate and unbiased statistical inferences. Having identified the functional relationship (model) among the variables under study, the next stages are to estimate the parameters included in the model and evaluate the adequacy of the estimated model (diagnosis checking). In diagnosis checking, if the model fits well, the fundamental assumptions of error of a statistical model are satisfied: the error should be uncorrelated (independent) with a constant variance (homogeneity) and also be normally distributed. If the estimated model does not fulfill at least one of the assumptions, a new model for the data must be specified, and the estimated and diagnosis checking cycle must be repeated. Only if the estimated model passes all the diagnostic checks, it should be used for interpretation and prediction purposes. In time series analysis, the Box-Jenkins approach is one of the most methods that are widelyused for building a model time series data, commonly known as autoregressive integrated moving average ARIMA (p,d,q) model (Box et al., 2016), (Brocwell et al., 2016) and (Chatfield et al., 2019). Many researchers have used this approach in many various scientific fields (for example (Zhang 2003) & (Hipel et al., 1994). These models are generally derived from three basic time series models: autoregressive AR (p), moving average MA (q) and autoregressiveand moving average ARMA (p,q). In practical matters, the time series required in models AR, MA, and ARMA are stationary processes. This means that the mean and the variance of thetime series data do not change with time. Therefore the ARIMA model fits the time seriesdata generally can be decomposed into two parts : The first component consists an integrated (I) component (d) which represents the order of differencing to be performed on the series totransform the nonstationary data into a stationary data by using a linear difference equation. While the second component consists of an ARMA models for the stationary time series.

To build an ARIMA model for a given time series there are three stages: model identification, parameter estimation and diagnostic checking (Box et al., 2016). The identification stage is the first stage of the construction model, which determines the appropriate orders for both AR and MA terms, followed by the estimation of the unknown parameters included in the model. Diagnostic checking is the last stage of model building that consists of evaluating the adequacy of the estimated model to fit historical data. If the fitted model is appropriate, thenthe residuals estimated from this model should resemble that of a white noise process.

There are two typical methods for checking the model’s assumptions. Using graphical approaches is the simplest method. Even while graphical approaches can be a useful tool and are more flexible in terms of evaluating assumptions, they are still difficult to interpret and donot provide clear evidence that the assumptions are hold. As results, to support the graphical methods, more formal methods which are the numerical methods (statistical tests) shouldbe performed before making any conclusion about the model’s assumptions. Both types to check the assumptions are typically based on an analysis of residuals which is a powerful andan effective tool for detecting model misspecification, including assumption violation. There isa lot of available literature on diagnostic checks for ARMA models (Jenkins). Diagnostic checks based on residual autocorrelation plots were proposed by (Mcleod et al., 1983). If data is normally distributed, the graph of the cumulative distribution for the data should appear as a straight line when plotted on normal probability paper, according to (Chow et al., 1988). (Ben-farage 2004) who using graphical methods to detecting the wrong model in time series data. (Ljung et al., 1978) who examined the properties of Portmanteau statistic tests for testing non significance of Autocorrelations between the residuals.

The main objective of the present paper is an attempt to answer the following question: Is the model diagnostic check (analysis of residuals) decisive factor in selection an appropriate model or it should be supplemented by other criteria. In other word, the objective is to evaluate the performance and efficiently of analysis of residuals in selecting the true time series model based on a simulation. Hence, this work focuses on using the various models (ARMA models) including the true model and compare these models to obtain the model that best fit to thedata by examining the residuals. This task may be helpful to known the reliability of analysisof residuals in judging the aptness of the fitted statistical model to time series data.

The organization of the rest of this paper is as follows Section 2 gives a brief introduction to the basic concepts of time series modeling. In addition, discusses the properties of the diagnostic statistics of the residuals that can be used to examine the goodness the estimated model to the time series data. Section 3 presents the results of the simulation study showing the efficiencyof the residual analysis in terms of their ability to identify the true model among various the competing time series models. Finally, the conclusions of the study are summarized in Section4.

2. METHODOLOGY OF RESEARCH

This section will provide the main concepts of time series analysis, which will be an important tool for the analysis in this paper.

2.1 Time Series Model:

The Mixed Autoregressive and Moving Average Model (ARMA) is a combination of AR and MA models, in which the current value X_t of the time series is expressed linearly in terms ofits previous values as well as current and a moving average of the current and past white noise stochastic error terms ε_t. The notation ARMA(p,q) refers to model with p order AR terms andq order MA terms. Thus an ARMA(p,q) model is written as:

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \qquad (1)$$

The error terms ε_t are generally assumed to be independent identically distributed random variables (white noise) from a normal distribution with mean zero and constant variance σ^2 . $X_{t-1}, X_{t-2}, \cdots, X_{t-p}$, are past series values (lags) and $\varphi_1, \varphi_2, \cdots, \varphi_p$ and $\theta_1, \theta_2, \cdots, \theta_p$ are the corresponding parameters which are estimated using maximum likelihood and conditional Least Squares approaches.

Once the model of time series data and parameters have been estimated, the next issue to our concern is how to select an adequate model which describes the historical pattern in the time series data in order to be used for accurate forecasting. Diagnostic checking is the most important stage of time series model building. In examining the adequacy of the estimated model, an analysis of the residuals is often performed. Time series analysis as in all other fields of statistics the objective is to covert the series of data into a series of white noise (residuals), i.e.a sequence of independent and identically distributed (i.i.d) random variables with zero meanand a constant variance. If the estimation is determined to be inadequate for the data, the methodology of Box Jenkins adopts returning to the model identification stage to re-examinethe appropriateness of the fitted model that requires examining the residuals.

2.2 Diagnostic checking

The appropriateness of the model is verified by employing white noise assumption test to check whether residuals are independent, homogeneity and normally distributed. To be able to see whether residual ε_t is white noise or not, it can be done by performing several tests. In this section we present some of the commonly applied tests to diagnostic checks for univariate and linear time series model extensively for this purpose.

Confirmation of independently Assumption. The essential assumption of the residuals of an time series model are that they white noise. A white noise is a serially uncorrelated variables. If a series has a white noise it indicates uncorrelated random variable with a zero mean and a constant variance. In order to determine whether residuals are independent (white noise), the residuals autocorrelation function (RACF) is examined. If the RACF is significantly different from zero, this implies that there is dependence between residuals. Although, RACF is a powerful complementary tool for testing independence, there are several statistical tests used for diagnostic checking of independence and randomness. In this study, the Ljung-Box Q statistic and Runs tests are used. The null hypothesis for each of these tests is that the residuals are independently distributed against the alternative hypothesis the residuals are not independently distributed; they exhibit serial correlation.

The Ljung-Box test or Q(r) statistic suggested by (Ljung et al., 1978), is a portmanteaulack of fit test for checking the independently assumption. The Q(r) statistic is calculated bythe following equation:

$$Q(r) = n(n+2)\frac{\sum_{k=1}^{h} r_k^2(a)}{n-k} \qquad (2)$$

where n is the sample size, $r_k^2(a)$ is the residual autocorrelation of order k and h is the number of autocorrelation lags being tested. Under the null hypothesis the Q(r) statistic asymptotically follows a chi-square distribution with h degrees of freedom. The Q(r) is compared to critical values from chi-square distribution with h degree of freedom. If the model correctly specified, the residuals should be uncorrelated and Q(r) should be small and corresponding the probability should be large.

The Runs test, also known as the Wald-Wolfowitz test is an non-parametric test (Siegel et al., 1988). Thistest is based on the order in which the residuals occur. A run is a set of sequential values of residuals that are either all above or below the median. To simplify computations, the residuals are first centered about their median. To carry out the test, the total number of runs is computed along with the number of positive and negative values. Let n be the number of residuals, n_1 bethe number of residuals above their median, n_2 be the number below their median and R be the observed number of runs. When n is relatively large the distribution of number of run (R) is approximately asymptotically normally distributed.

$$Z_{cal} = \frac{R - E(R)}{\sqrt{Var(R)}} \qquad (3)$$

The expected value E(R) and variance Var(R) of R are defined as:

$$E(R) = \frac{n + 2n_1n_2}{n} \qquad Var(R) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)} \qquad (4)$$

The null hypothesis is rejected if the calculated Z_{cal} value is greater than the selected criticalvalue obtained from the standard normal distribution table.

Confirmation of Normality Assumption. In statistic, to determine the white noise assumption, residuals must also meet the normal distribution. Normal residual examination canbe done using QQ-plot. This graph should appear as a straight line when it is plotted on normal probability paper. Therefore, to support the visual methods, more formal normality tests which are the numerical approaches should be performed before making any conclusion about the normality assumption. In this paper, Shapiro and Wilk (SW) (Shapiro et al., 1965) and Lilliefors Kolmogorov-Smirnov (Lilliefors, et al., 1967) tests were used as alternative approaches for diagnostic checking for normality assumption. The null and alternative hypothesis for each of these tests can be written as follows : $\mathrm{H_0}$: The residuals of the fitted ARMA model are normality distributed against $\mathrm{H_1}$: The residuals of the fitted ARMA model are not normality distributed.

Shapiro and Wilk (SW) is test for normality, developed by (Shapiro et al., 1965), has been found to be the most powerful and omnibus test in most situations. The SW statistic is calculated as follows :

$$\mathrm{SW} = \frac{1}{\mathrm{D}}\left[\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{a_i}(\mathrm{X_{(n-i+1)}} - \mathrm{X_{(n)}})\right]^2 \qquad (5)$$

where $\mathrm{m} = {}^{\mathrm{n}}/_{2}$ if n is even while $\mathrm{m} = {}^{(\mathrm{n}-1)}/_{2}$ if n is odd and $\mathrm{D} \underset{i=1}{=}$ $\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_{(i)}} - \bar{\mathrm{x}})^2$ and $\mathrm{x_{(i)}}$ represents the ith order statistic of the residuals in the sample, the constants $\mathrm{a_i}$ are given by

$$\mathrm{a} = \frac{\mathrm{m'V^{-1}}}{(\mathrm{m'V^{-1}V^{-1}m})^{1/2}} \qquad (6)$$

where $\mathrm{m} = (\mathrm{m_1}, \mathrm{m_2} \cdots, \mathrm{m_n})'$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those order statistics. The values of W lie between 0 and 1 and small values of the statistic indicate departure from normality under $\mathrm{H_0}$, thus, if the value oftest statistic W is less than the critical value, null hypothesis of normality is accepted.

Lilliefors Kolmogorov-Smirnov (LF) test is a modification of the Kolmogorov-Smimov test. Lilliefors test is based on the maximum vertical absolute difference between cumulative distribution and the Normal cumulative distribution curve (when the null hypothesis is that the cumulative distribution demonstrates normality). Given a sample of n residuals, Lilliefors statistic test is defined as (Lilliefors 1967):

$$\mathrm{D} = \max|\mathrm{F_n}(\mathrm{x}) - \mathrm{S_a}(\mathrm{x}, \mu, \sigma^2)| \qquad (7)$$

Where $\mathrm{F_n}(\cdot)$ is the cumulative distribution function based on residuals, and $\mathrm{S_a}(\mathrm{x}, \mu, \sigma^2)$ is the theoretical cumulative (normal) distribution function with μ and σ^2, where μ and σ^2 are, respectively, the mean and variance of the residuals. The null hypothesis at the level of significance α, can be rejected if the D test statistic is larger than the critical value $\mathrm{D}(\mathrm{n}, \alpha)$ obtained from the K-S Test table.

Confirmation of Homogeneity Assumption. Homogeneity of variances is often a reference to equal variances across groups. There are many statistical tests, such as the analysis of variance, assume that variances are equal across groups. In this study, Bartlett (Bartlett1937) and Levenes (Levene 1960) tests were used as alternative approaches for the diagnostic checking of residuals for homogeneity. We want to know whether variances are equal within the groups, that is to test hypothesis of variances homogeneity. Thus the null for variances equality of g groups (here, $\mathrm{g} = 4$) has the following form: $\mathrm{H_0}: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_\mathrm{g}^2$ and alternative hypothesis for each of these tests can be written as follows : $\mathrm{H_1}: \sigma_\mathrm{i}^2 = \sigma_\mathrm{j}^2$, $\mathrm{i} \neq \mathrm{j}$ where the inequality holds at least for one pair of i, j.

Bartlett test statistic test has been introduced by (Bartlett1937) to test homogeneity of variances. the null hypothesis, $\mathrm{H_0}$ that all g groups of residuals have equal variances against the alternative hypothesis $\mathrm{H_1}$ that at least the two are different. For the test, the residuals from the fitted model to the data are divided into g groups with size $\mathrm{n_i}$, sample variance of the *ith* group S_i^2 and S_p^2 is the pooled variance, then Equation (8) can be used to calculate the Bartlett test statistic :

$$\mathrm{X}_{cal} = ((n-g)\log(S_p^2) - \sum_{i=1}^{g}(n_i - 1)\, logS_i^2)(1 + \frac{1}{3(g-1)}(\sum_{i=1}^{g}\frac{1}{n_i - 1} - \frac{1}{n-g}))^{-1} \qquad (8)$$

Where

$$n = \sum_{i=1}^{g} n_i \ \ and\ \ S_p^2 = \frac{\sum_{i=1}^{g}(n_i - 1)S_i^2}{n - g} \qquad (9)$$

If the assumption of homogeneity is met; the distribution of statistic for Bartletts test follows the Chi-squared distribution with degrees of freedom $g - 1$.

Levenes Test (LV) has been proposed by (Levene 1960) to test homogeneity of variances between groups. Levene's test is an alternative to the Bartlett test. This test is less sensitive than the Bartlett test to departures from normality than the Bartlett test. The test statistic is defined as:

$$L = \frac{(n-g)\sum_{i=1}^{g} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2}{(g-1)\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2}\ , \quad Y_{ij} = |X_{ij} - \bar{X}_{i.}| \qquad (10)$$

where $\bar{Y}_{i.}$ is the mean of Y_{ij} for ith group and $\bar{Y}_{..}$ is the overall mean of the Y_{ij}. Levene's original paper only proposed using the mean as the center. (Brown et al., 1974) extended Levene's (modified Levene's) test to use either the median or the trimmed mean to substitutefor the mean. In this paper modified Levene's test is used.

Box and Jenkins (Box et al., 2016) recommended that the importance of parsimony principle in selecting the appropriate model to the time series data. In term of parsimony, they expressed the need to select the optimum model that has fulfilled all the diagnostic checks (residuals analysis) and use as few number of parameters as possible of estimated model. To evaluate the models in order to select the best model that describes the data series adequately, the statistical criteria for selecting the optimum model was used. These criteria were: kaike Information criterion (AIC) [8], Corrected Akaike Information Criterion (AICc) (Sugiura 1978) and Schwarz Information Bayesian (BIC) (Schwarz 1978). The estimated model with minimum of these criteria assumes to describethe data series adequately. A brief description about the criteria for the selection of best modelis given below :

$$AIC = -2\ln(L_{max}) + 2K \qquad (11)$$

$$AIC_C = -2\ln(L_{max}) + 2K + \frac{2K(K+1)}{n-K-1} \qquad (12)$$

$$BIC = -2\ln(L_{max}) + Kln(n) \qquad (13)$$

Where K is the number of parameters to be estimated and L_{max} is the maximized value of the likelihood function of the fitted model. The minimum value of this criterion is desirable for the adequacy of a model among all candidate models.

3. RESULTS AND DISCUSSION

Having broached the some basic concepts of time series analysis that will enable us analyze the time series data and build the appropriate model. We now display an important steps of analysis our dataset. The data were extracted from (Pristely 1981) which was generated from AR(1), it is defined as : $\mathrm{X_t} = 0.6\mathrm{X_{t-1}} + \varepsilon_\mathrm{t}$. The graphical plot of the data is given in Figure 1. It is observed that the series dose not display considerable any fluctuations over time, where the lower values display considerably the same variation as the higher values. We claim that the stationarity behavior of the series and a stationary model seem to be reasonable. In this study, we conducted simulation study to evaluate the efficiency and usefulness analysis of residuals, in order to make a comparison for selecting the appropriate model that fits the data well. The simulation study was conducted using the R packages. We are simulating different candidatemodels with different orders including the true model. These models were AR(1), AR(2), AR(3), AR(4), MA(1), MA(2), ARMA(1,1), ARMA(1,2), ARMA(2,1) and ARMA(2,2).

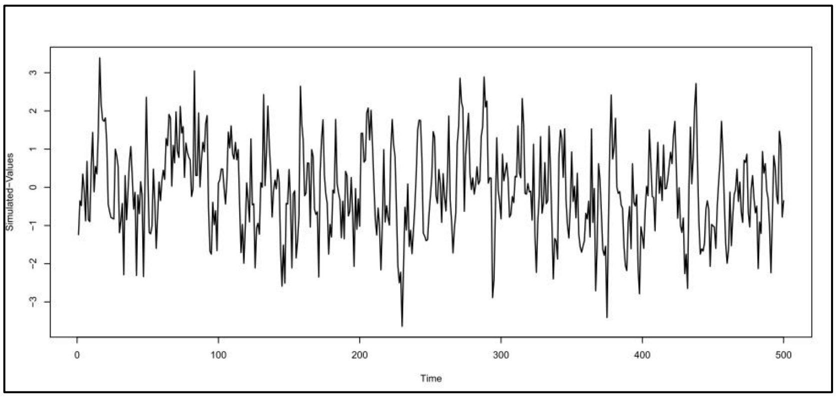


Figure 1 :The Simulated data generated from AR(1)

Table 1: Summary of the statistical parameters of the fitted models.

Model	parameters	Value	S.E	t-ratio	Sig
AR(1)	φ_1	0.5526	0.0372	14.841	0.000
AR(2)	φ_1	0.5574	0.0447	12.465	0.000
	φ_2	-0.0086	0.0447	-0.192	**0.848**
AR(3)	φ_1	0.5572	0.0447	12.463	0.000
	φ_2	0.0036	0.0513	0.070	**0.944**
	φ_3	-0.0217	0.0448	-0.485	**0.627**
MA(1)	θ_1	0.4681	0.0340	13.780	0.000
MA(2)	θ_1	0.5371	0.0424	12.655	0.000
	θ_2	0.2447	0.0420	5.825	0.000
MA(3)	θ_1	0.5522	0.0448	12.336	0.000
	θ_2	0.2802	0.0452	6.196	0.000
	θ_3	0.1127	0.0449	2.511	0.012
ARMA(1,1)	φ_1	0.5426	0.0663	8.189	0.000
	θ_1	0.0145	0.0781	0.185	**0.853**
ARMA(1,2)	φ_1	0.5075	0.1143	4.442	0.000
	θ_1	0.0489	0.1205	0.406	**0.685**
	θ_2	0.0292	0.0722	0.405	**0.686**
ARMA(2,1)	φ_1	-0.4283	0.0379	-11.298	0.000
	φ_2	0.5339	0.0379	14.072	0.000
	θ_1	1.0000	0.0061	165.216	0.000
ARMA(1,1)	φ_1	1.1519	0.3182	3.621	0.000
	φ_2	-0.4043	0.1816	-2.226	0.026
	θ_1	-0.5933	0.3179	-1.866	**0.062**
	θ_2	0.0851	0.0870	0.978	**0.328**

Figures in bold indicate to Critical values are at 5% significance level. $\varphi_1, \varphi_2, \varphi_2$ are coefficients of autoregressive models; $\theta_1, \theta_2, \theta_3$ are coefficients of moving-average models.

Table 2: Independence (randomness) test results of the residuals for each fitted model.

Model	Ljung-Box statistic		Decision	Run test		Decision
	Q(h)	P value		Z	P value	
AR(1)	29.653	0.1965	R	-0.58196	0.5606	R
AR(2)	23.5665	0.1697	R	-0.6267	0.5308	R
AR(3)	28.85	0.2258	R	-0.58196	0.5606	R
MA(1)	68.728	0.0000	NR	-1.8354	0.06644	R
MA(2)	32.97	0.1047	R	-0.4029	0.687	R
MA(3)	31.188	0.1484	R	-1.1192	0.2631	R
ARMA(1,1)	29.353	0.2071	R	-0.58196	0.5606	R
ARMA(1,2)	29.054	0.2181	R	-0.76103	0.4466	R
ARMA(2,1)	28.866	0.2252	R	-1.4773	0.1396	R
ARMA(2,2)	27.452	0.2838	R	-0.58196	0.5606	R

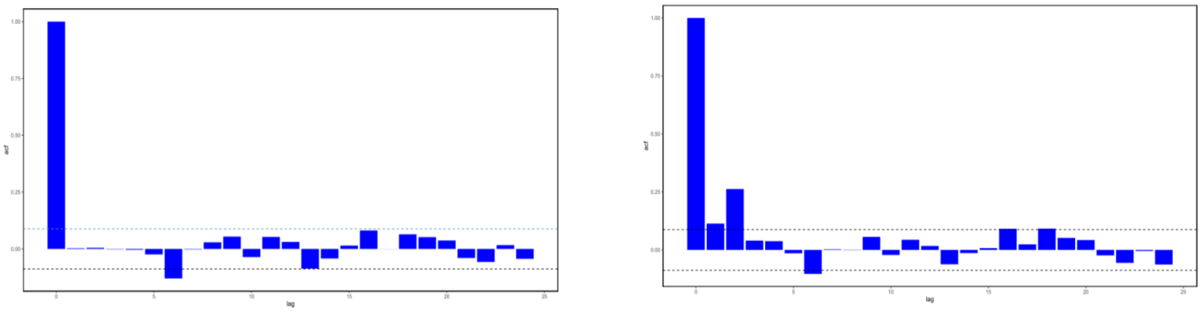
R: Residuals are randomness. NR: Residuals are not randomness.

Table 1 display summary of results for the values of the parameters concerning with the fitted models associated the standard errors (S.E), t-ratios and probabilities for the standard errors. The results shown in Tables 1 indicate that all estimated coefficients of estimated models AR(1), MA(1), MA(2) and ARMA(2,1) have P-value less than $\alpha = 0.05$. This implies that all the coefficients of these modelsare significant since the null hypothesis $H_0\colon \varphi = 0$ for (AR) or $H_0\colon \theta = 0$ for (MA) can be rejected for the significance level 5%. Other models tested resulted in coefficients with P-values higher than 0.05 and these coefficients will have little effect (over-fitting) on model description and prediction. Clearly, the models AR(2) and AR(3), ARMA(1,1), ARMA(2,2) and ARMA(2,2) are over fitted from AR(1). Although the estimated coefficients of these models are insignificant at the 5% significance level, and they should be eliminated of models to avoid over-fitting, these models will be used as the fitted model according to the purpose of study mentioned above.

In time series, the model building methodology requires examining the residuals of the modelto verify that the models are adequate. The residuals are examined to discover the residuals are white noise. Two tests, namely the Ljung-Box Q statistic and Runs tests are applied for the critical independence assumption of residuals for the best models. The results of these tests were presented in Table 2.

For testing the independence based on the Run test, the results clearly showed that the probabilities (p-values) associated of this test exceeds significance level 0.05, thus indicating that the test is not significant and residuals appear to be uncorrelated. This implies that for all the estimated models, the residual resembles random white noise. In addition, Ljung-Box-Pierce statistic is employed to check the independence of the residuals using the first 24 RACF from the fitted models. The Ljung Box Statistic values of all the estimated models, except for MA(1) model; are not significantly different from zero and associated P-value greater than 0.05, thus failing to accept the null hypothesis of white noise. While, the estimated model MA(1), the Ljung-Box Q-statistic value correspond to p-values less than 0.05 thus indicatingthat the test is significant and residuals appear to be correlated. Therefore, since Q is unduly large and the evidence contradict the hypothesis of white noise behaviour in the residuals, the model is not adequate and significantly appropriate.

Figure 2 shows the ACF of the residuals of the true model AR(1) and candidate modelMA(1) models. From Figure 2, these plots was examined, for MA(1) model, it was clear that the ACF plot of residual shows that the residuals were not within the confidence intervals at lags2 and 3 which is an indication of a misspecification of the model. All of these results emphasize that the RACF of MA(1) model was significantly different from zero. In other words, therewas



(a) MA(1)

(b) MA(1)

Figure 2: a) Autocorrelation function plot for the residual of AR(1) model and (b) Autocorrelation function plot for the residual of MA(1) model.

A significant linear dependence between residuals, while, for AR(1) model, the residuals fell within the confidence interval which is an indication of a good fit and the adequacy of the AR(1)model. Also, this was an indication that they were not significant and that the residuals were independent and thus satisfying the residual assumption.

Two approaches, namely the Lilliefors KS and Shapiro and Wilk (SW) tests, are appliedfor the normality assumption of residuals for the fitted models. The results of these tests are presented in Table (3). It is obvious from Table (3) that the probabilities of the aforementioned tests are greater than 0.05 level of significant at 95% confidence interval. These results concerning the Lilliefors KS and SW tests imply that the residuals of the fitted model are normality distributed.

Table 3: Normality test results of the residuals for each fitted model.

Model	*Lilliefors KS test*		Decision	SW test		Decision
	KS	P		AD	P value	
AR(1)	0.0320	0.2429	ND	0.9976	0.7	ND
AR(2)	0.0324	0.2288	ND	0.99764	0.7099	ND
AR(3)	0.0368	0.1036	ND	0.9976	0.6884	ND
MA(1)	0.0279	0.4463	ND	0.99761	0.6998	ND
MA(2)	0.0308	0.2960	ND	0.99796	0.8166	ND
MA(3)	0.0314	0.2723	ND	0.99749	0.6594	ND
ARMA(1,1)	0.0324	0.2305	ND	0.99764	0.7089	ND
ARMA(1,2)	0.032404	0.2286	ND	0.9976	0.698	ND
ARMA(2,1)	0.026789	0.5177	ND	0.9979	0.7967	ND
ARMA(2,2)	0.034437	0.1596	ND	0.99734	0.6038	ND

ND : Residuals are normality distributed.

For the selected best model, the results related to the homogeneity of variance of the residuals using Bartlett and Levenes test statistics are also summarized in Table (4). As the computed the probability values of the statistics test associated with these tests was greater than significance level (α = 0.05), one cannot reject the null hypothesis H_0 (the variances equality of g groups of residuals). These results concerning the Bartlett test and Levenes test statistic imply that the residual variances are constant. Thus, these approaches satisfy condition the related to homogeneity of residuals (variances are constant).

Table 4: homogeneity test results of the residuals for each fitted model.

Model	*Bartlett test*		Decision	*Levene's Test*		Decision
	B	P		L	P value	
AR(1)	2.958	0.3981	CV	0.5645	0.6387	CV
AR(2)	2.9976	0.392	CV	0.5649	0.6384	CV
AR(3)	2.9566	0.3984	CV	0.5636	0.6392	CV
MA(1)	4.0449	0.2567	CV	1.2497	0.2911	CV
MA(2)	2.9209	0.404	CV	0.5496	0.6486	CV
MA(3)	3.3119	0.346	CV	0.6981	0.5535	CV
ARMA(1,1)	2.995	0.3924	CV	0.5649	0.6384	CV
ARMA(1,2)	2.9541	0.3988	CV	0.5621	0.6403	CV
ARMA(2,1)	3.2937	0.3485	CV	0.5342	0.659	CV
ARMA(2,2)	3.2024	0.3615	CV	0.651	0.5826	CV

CV: Residuals have constant variances

Based on the diagnostic checks described in section 2.2, the selection of a best model fit to data isdirectly related to whether residual analysis is performed well. From results above, the analysis of residual chose several models as the appropriateness models that fit the data well best. These models were AR(1), AR(2), AR(3), MA(2), MA(3), ARMA(1,1), ARMA(1,2), ARMA(2,1) and ARMA(2,2), since they fulfilled the assumptions of being the residuals of the models were independent, homogeneity and normally distributed. In contract to the MA(1) model did not fulfill at least one of the diagnostic checks, so it was eliminated of analysis. Noticeably, from theresults, the residual analysis failed to pick up the true model AR(1) correctly.

Table 5: Selection of Best fitted Model.

Model	LogLik	ME	σ^2(MSR)	RMSE	AIC	AICc	BIC
AR(1)	-704.005	-0.040	0.979	0.989	1412.010	1412.034	**1420.439**
AR(2)	-703.987	-0.040	0.982	0.991	1413.973	1414.022	1426.617
AR(3)	-703.869	-0.041	0.983	0.992	1415.738	1415.818	1432.596
MA(1)	-724.197	-0.061	1.062	1.031	1452.395	1452.419	1460.824
MA(2)	-708.511	-0.050	0.999	0.999	1423.023	1423.071	1435.666
MA(3)	-705.357	-0.046	0.989	0.994	1418.714	1418.795	1435.572
ARMA(1,1)	-703.988	-0.040	0.982	0.991	1413.976	1414.024	1426.620
ARMA(1,2)	-703.905	-0.041	0.983	0.992	1415.811	1415.892	1432.669
ARMA(2,1)	**-701.822**	-0.040	**0.971**	**0.986**	**1411.643**	**1411.724**	1428.50
ARMA(2,2)	-703.560	-0.046	0.984	0.992	1417.121	1217.243	1438.195

Although analysis of residuals can be used to compare the overall goodness and adequacies offit of candidate models, it was unable to identify the true model AR(1) correctly, assess the need for additional statistical criteria and selects the best model among all the candidate models. In statistics, the models performance efficiency is evaluated using five criteria. These criteria; namely, the Mean Square Error (MSE), Root Mean Square Error (RMSE), the maximized value of the log likelihood function of the estimated models (Loglik), the Akaike Information criterion (AIC), Corrected Akaike Information Criterion (AICc) and the Normalized Bayesian information criterion (BIC), were taken into account for obtaining a parsimonious model among these candidate models that fulfilling all the diagnostic checks. The procedure for choosing between these candidate models relies on choosing the model with the maximum value LogLik and minimum values of MSE, RMSE, AIC, AICc and BIC. Comparison of the candidate models and their corresponding values of these criteria are illustrated in Table (5).

The results presented in Table (5) indicate that AR(2,1) was chosen as the appropriate model that fits the data well based on LogLiklood, MSE, RMSE, AIC, AICc, despite Bayesian information criterion (BIC) chose the model AR(1) as suitable model. Thus, the results showed that the AR(1) model is superior to the other candidate models; that fulfilled all the diagnostic checks, having the least BIC value. Overall, BIC outperform other criteria having low value, thus it did not failed to pick up the true model AR(1) correctly. Based on this model selection criteria BIC and analysis of residuals, we also conclude from Table 1, that the coefficient of the AR(1) modelis significantly different from zero at 5% significant level. Thus, the selected parsimony model for any data set among the candidate models that fulfilled the diagnostic checks considering the coefficients contained in the models is significantly and the minimum value of Bayesian information criterion BIC. Figure 3 shows the performance of different criteria for candidate models. According to Figure 3, in general, we note that the criteria BIC is the best for selecting a true time series model.

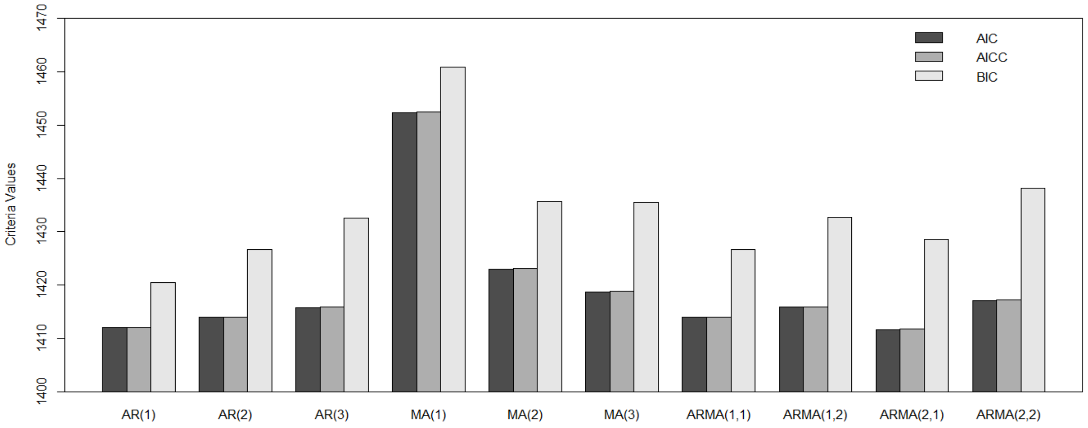


Figure 3 the performance of different criteria for fitted time series models

4. Conclusion

This paper concerns the efficiency of the analysis of residuals in selecting an appropriate statistical model that describes the historical pattern in the data in order to be used for accurate forecasting of future data. In time series analysis, diagnostic checking (analysis of residuals) in building models consists of evaluating the adequacy of the estimated model. The null hypothesis assumes that the fitted model is an appropriate model and the residuals behave like white noise series. It was found that the analysis of residuals should not be taken as a final judgment about the adequacy of the estimated statistical model for the time series data. This study reveals that when analysis of residuals was used as the main criterion of diagnostic checking of the model in picking up the true model, many alternative models were proposed for the data, which means that the residuals analysis alone is not enough as a test for the goodness of the adequacy of the estimated model. In addition to that, it can be concluded that the analysis of residuals alone was not able to detect the wrong model when the order of the true model was lower than the order of the fitted model.

The study also found that when different class of models from the true class was fitted to the data, analysis of residuals was not sharp enough to detect the wrong model. It is obvious that when the fitted model was ARMA with fitted order is lower or higher than the true order the analysis of residuals did not clearly to indicate wrong model. The main findings from this work can be summarised as follows: the analysis of residuals should not be the final word in examining the validity of the fitted model, but it should be supplemented by significant parameters and other criteria such as Bayesian information criteria.

5. References

- Akaike,. H,1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19(6): 716 – 723.
- Bartlett, M. S.1937. Properties of sufficiency and statistical tests. Proceedings of the Royal Statistical Society, Series A 160, 268–282.
- Ben-farag, O.S.2004. The reliability of the analysis of residuals in detecting the wrong time seriesmodel. M.Sc Thesis. University of Benghazi.
- Box, G.E.P. & Jenkins, G.M.2016. Time Series Analysis Forecasting and Control, Fifth edition, Wiley, Hoboken, NJ.
- Brockwell, P.J. & Davis, R. A.2016. Introduction to time series and forecasting, Third edition, Springer, Switzerland.
- Brown, M.B. & Forsythe, A.1974. Robust Tests for Equality of Variances. Journal of the American Statistical Association. 69(346) : 364-367.
- Chatfield, C. & Xing, H.2019. The analysis of time series : an introduction. Seventh edition, CRC Press, Boca Raton, Florida.
- Chow, V., Maidment. D. & Mays, L.1988. Applied Hydrology. International Edition, McGraw-Hill Book Company, New York.
- Hipel, K.W. & McLeod, A.I.1994. Time Series Modelling of Water Resources and Environmental Systems. First Edition, Elsevier, Amsterdam.
- Levene, H.1960. Robust testes for equality of variances. In Contributions to Probability and Statistics (I. Olkin, ed.) 278–292. Stanford Univ. Press, Palo Alto, CA.
- Lilliefors, H. W.1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association. 62(318): 399–402.
- Ljung, G.M. & Box, G. E. P.1978. On a Measure of Lack of Fit in Time Series Models. Biometrika, Oxford University Press, 65(2) : 297-303.
- Mcleod, A.I. & Li, W.K.1983. Diagnostic Checking ARMA Time Series Models Using Squared-Residual Autocorrelation. Journal of Time Series Analysis 4(4) : 269-273.
- Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics, 6, 461 - 464.
- Shapiro, S.S. & Wilk, M.B.1965. An Analysis of Variance Test for Normality (Complete Samples). Biometrika. Oxford University Press. 52(3/4) : 591-611.
- Siegel, S. & Castellan, N.J.1988. Nonparametric Statistics for the Behavioral Sciences. Second Edition, New York: McGraw-Hill.
- Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections, *Communications in Statistics - Theory and Methods*, 7: 13–26.
- Pristely, M.B.1981. Spectral Analysis and Time Series. Acadamic Press.
- Zhang, G.P.2003. Time series forecasting using a hybrid arima and neural network model. Neurocomputing, (50): 159–175.