Statistical Analysis of a Nonstationary Fatigue Data Using the ARIMA Approach

S. ABDULLAH[1], M. D. IBRAHIM, A. ZAHARIM AND Z. MOHD NOPIAH
Engineering Faculty, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor,
MALAYSIA
[1]shahrum@eng.ukm.my

Abstract: - Auto Regressive Integrated Moving Average (ARIMA) is a broad class of time series models, and it has been achieved using the statistical differencing approach. It is normally being performed using the computational method. Thus, it is useful to choose the suitable model from a possibly large selection of the available ARIMA formulations. The ARIMA approach was then analysed with the presence of stationary behaviour in a nonstationary data. For the purpose of the random data analysis, a nonstationary data that exhibiting a random behaviour was used. This random data was measured in the unit of microstrain on the lower suspension arm or a car travelling on a country road surface. With this engineering unit, hence, the data is known as a variable amplitude fatigue loading. Experimentally, the data was collected for 225 seconds at the sampling rate of 200 Hz, which gave 45,000 discrete data points. Using the computational analysis by means of statistical software package, the ARIMA parameters were estimated by the application of the data smoothing technique in order to reduce the random variation of the fatigue data. Therefore, the significant ARIMA parameters were established and being applied in the study of the variation in nonstationary data. For this paper, finally, it is suggested that the ARIMA method provided a good platform to analyse fatigue random data, especially in the scope of the durability research.

Key-Words: - ARIMA, Statistical analysis, Fatigue, Nonstationary data, Statistics.

1 Introduction

A time series typically consists of a set of observations of a variable taken at equally spaced intervals of time [1]. Today, most experimental measurements, or data samples, are performed digitally. And it is also known as a discrete time series, which is formed as a function of time. The objective of time series analysis is to determine the statistical characteristics of the original function by manipulating the series of discrete numbers. Based on the different term of a time series, a signal is a series of numbers that come from measurement, typically obtained using some recording method as a function of time. In the case of fatigue research, the signal consists of a measurement of the cyclic loads, i.e. force, strain and stress against time.

In addition to the data analysis of variable amplitude fatigue loadings, many data mining applications deal with privacy-sensitive data [2]. The best means of obtaining unpredictable random numbers is by measuring physical phenomena such as fatigue damage, radioactive decay, thermal noise in semiconductors and even digitized images of a lava lamp. However few computer users accessed to the specialized hardware that required for these sources, and must rely on other means of obtaining random data [2].

The objective of this study is to observe the capability of a technique called Auto Regressive Integrated Moving Average (ARIMA) in preserving a nonstationary behaviour of a data by underlying probabilistic properties. This study has been motivated from the development of a class of data algorithms [3,4] that were used to extract the data pattern without directly accessing the original data and guarantees that the process. A major advantage of performing this process is the ability of the modeller to select the proper model from possibly large selection of the available model formulation. This approach is used to preserve data privacy from random noise [5]. Typically, these data are the used with curve-fitting techniques to develop the average fatigue behaviour of the material over an appropriate range of stress levels.

2 Literature Background

Many signals in nature exhibit random or nondeterministic characteristics which provide a challenge in analysis [6]. A signal representing a random physical phenomenon cannot be described in a point by point manner by means of a deterministic mathematical equation. A signal representing a random phenomenon can be characterised as either stationary or nonstationary.

A stationary signal is characterized by values of the global signal statistical parameters, such as the mean, variance and root-mean-square, which are unchanged across the signal length. Stationary random processes can further be categorized as being ergodic or nonergodic. If the random process is stationary, and the mean value and the autocorrelation function do not differ when computed over different sample segments measured for the process, the random process is defined as ergodic. In the case of nonstationary signals the global signal statistical values are dependent on the time of measurement [7]. Nonstationary signals can be divided into two categories: mildly nonstationary and heavily nonstationary. A data is said does not satisfy the stationary condition is defined as non-stationary data [8]. This characteristic is common among a fairly large number of time series met in the real world. Nonstationary pattern happens when the data is not constant about mean or level (due to trend or seasonal pattern) and hence can be expressed as deterministic function for example. This situation is illustrated by a random data set as in Fig. 1, for which this data was experimentally measured for the purpose of this study on the lower suspension arm of a car, for which this car was travelling over a country road surface.

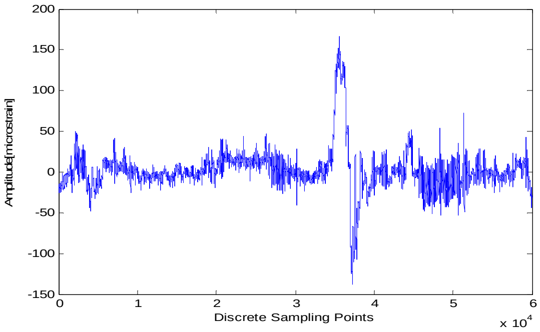

Fig. 1. A Nonstationary fatigue data which was measured on a lower suspension arm of a car

Global signal statistics are frequently used to classify random signals. The most commonly used statistical parameters are the mean value, the standard deviation value. For a signal with a number n of data points, the mean value of $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1}$$

The standard deviation (SD) is mathematically defined as

$$SD = \left\{\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2\right\}^{1/2} \tag{2}$$

for the samples more than 30 [9]. The standard deviation value measures the spread of the data about the mean value.

The nonstationarity of the random data can be easily determined from the calculation of the kurtosis value. Kurtosis, which is the signal 4th statistical moment, is the global signal statistic which is highly sensitive to the spikiness of the data. For discrete data sets the kurtosis value is defined as

$$K = \frac{1}{n(r.m.s.)^4}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^4 \tag{3}$$

where x_j is the instantaneous value, $\bar{x}$ is the mean value of a signal, $r.m.s.$ is the root-mean-square value (represents the amount of the time-domain vibrational energy of a signal) and n is the number of values in the sampled sequence. For a Gaussian distribution the kurtosis value is approximately 3.0. Higher kurtosis values indicate the presence of more extreme values in a Gaussian distribution, showing the behaviour of a nonstationary signal. The kurtosis value is used in engineering for detection of fault symptoms because of its sensitivity to high amplitude events.

Since most of the nonstationary data exhibits the random pattern (especially for the data shown in Fig. 1 with the kurtosis value is 13.745), the ARIMA method is introduced as one of the approaches for smoothing the time series data. It can be estimated by smoothing the data in order to reduce the random variation [10]. Although a range of smoothers is available from the current statistical applications, but it begin with the simplest and oldest smoother. ARIMA model is also a dependence relationship to set up among the successive error terms.

The idea of applying the ARIMA concept is related to the data observations which are likely to be closed in value. By taking an average of the points near an observation, it provides a reasonable estimation of the data. Thus, it eliminates the randomness of the data, and producing a smooth trend with respect to the original nonstationary data pattern [10].

The ARIMA method is to identify the class of models most suitable to be applied to the data set [11]. The process of determining the final or 'best' model is an iterative one as indicated in Fig. 2. It means that before final model is arrived at the process of formulating and estimating, the model has to be performed repeatedly, going back and forth, between the first two phases, each time revising and improving the model until one estimated model, which is superior to all other competing models, is found [12]. The main criterion used in the ARIMA analysis is based on the model forecasting performance [13].

The advantage of ARIMA is its robustness and having an excellent data seasonality analysis. In addition, the ARIMA method provides fast computational analysis and easily to be use with any kind of nonstationary data. The steps in ARIMA are also constituted with the important aspect of the Box-Jenkins methodology. The basis of the Box-Jenkins modelling approach consists of three main stages as listed in Table 1. Diagrammatically, the flow of the Box-Jenkins modelling can be referred in Fig. 2.

Table 1. Stages of the Box-Jenkins Modelling

Stage	Process
I	Model Identification
II	Model Estimation and Validation
III	Model Application

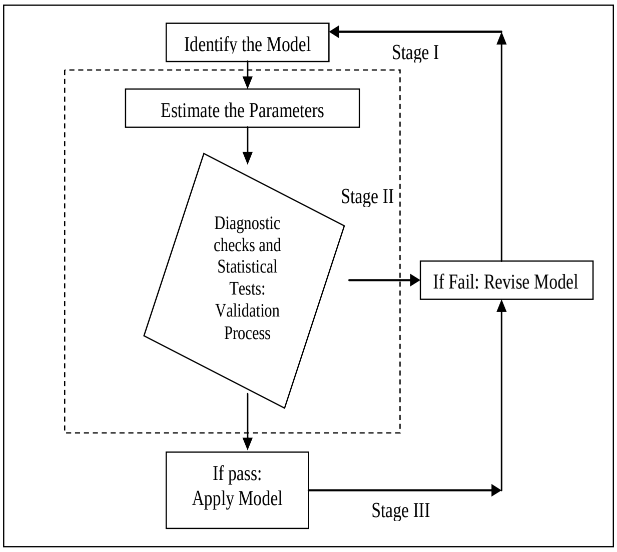

Fig. 2. Stages of the Box-Jenkins methodology which can be applied in the ARIMA computational analysis

The first step in ARIMA computational analysis is the model identification which is used to identify the class of models that is the most suitable to the data set. This approach is being performed by computing and analysing various statistics based on the data [14]. Once the particular subclass of the model has been identified, the next step is to identify the 'best model' to be fitted, such that the fitted values come as close as possible to capturing the pattern exhibit by the actual data set [15].

There are two important objectives that need to be achieved. The first objective is the fitted values should be as close as possible to the actual values [15]. The model fits the data well if it minimises the error measure in the amplitude loading data. The process is more specifically to search the estimated parameter values that minimise the actual values. The second objective is the models should require the least possible parameters consistent with a 'good' model fit [16].

The model is considered to be correctly specified if it includes the correct set of independent variables. An independent variable is considered as correct if its inclusion in the model helps to explain the phenomenon under study [15]. Thus, a model is said to be mis-specified if it fails to meet some or all of these test criteria. There are four common types of misspecification error [17]:

a. Relevant variables are omitted from the model.
b. Irrelevant variables included in the model.
c. The functional form of the model is questionable.
d. Issues related to an analysis of the residuals or errors associated with any specific regression model are not satisfactorily answered.

If all test criteria are met and that the model fitness has been confirmed, it is therefore can be used to generate the ARIMA significant parameters. Finally, the application of the Box-Jenkins methodology lies on the assumption that concerns the characteristic of the initial data [18].

3 Computational Data Analysis: The ARIMA Application Using Nonstationary Data

For the application of the nonstationary fatigue data (refer to Fig. 1 for the time series plot) with the computational analysis of this paper, the signal was measured on the front left lower suspension arm of an automobile which was travelling on the country road surface (mixture of smooth and irregular asphalt). In the data collection experiment, it was sampled at 200 Hz for 45,000 data points, and the record length of 225 seconds was obtained. Based on the simple statistical analysis, the data produced the mean and the standard deviation values of 2.337 microstrains and 25.4657 microstrains, respectively.

Using the ARIMA computational analysis with this data, a simple procedure was used to remove the presence of the nonstationary behaviour of this data. Thus, the differencing technique in ARIMA has been performed. A data that requires first difference to be stationary is said to be integrated of order one. However, there are cases in which a nonstationary process does not achieve when it fluctuates randomly around some fixed values, generally either around the mean value of the data.

On the other hand, a data is stationary if it does not show growth or decline or Y_t is stationary if these following condition are fulfilled:

a. The mean of Y_t, $E(Y_t) = E(Y_t -1) = E(Y_t -2) = E(Y_t -3) = \dots\dots = \mu$, which μ is a constant. (4)

b. The variance
$var(Y_t) = E(Y_t - \mu)^2 = \sigma^2 < \infty$ (constant). (5)

c. The covariance between Y_t and Y_{t-p} is
$\gamma_p = cov(Yt, Y_{t-p}) = E[(Y_t - \mu)(Y_{t-p} - \mu)]$. (6)

Another parameter which is used for the analysis is the Mean Squared Error (MSE), as being mathematically defined in Eq. (7).

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \quad (7)$$

where n is the number of observations in the series and e is a error terms. It is a measure of accuracy computed by squaring individual error for each item in a data set and then finding the average or mean value of the sum of those squares. The MSE value gives greater weight to the large errors than to the small errors.

In addition, the Akaike's Information Criterion (AIC) approach is also used and it is described as a measure of the goodness-of-fit of a model. The AIC approach is commonly applied with the ARIMA model in order to determine the appropriate model order. The AIC is equal to twice the number of parameters in the model minus twice the log of the likelihood function. The AIC was developed based on the entropy concept and it is mathematically formulated as the following equation

$$AIC = -2 \log L + 2m \quad (8)$$

where L denoted as the likelihood of the data, $m = p + q + P + Q$, the p and q parameters are the usual respective terms of the AR and MA part, and the P and Q parameters are the seasonality part of the ARIMA model. Most of the computer programs produced the value of σ^2 so the AIC value can be approximately found as

$$\text{AIC} \approx n\{1 + \log(2\pi)\} + n \log \sigma^2 + 2m \quad (9)$$

where σ^2 is the variance of the residuals and n is the number of observations in the series.

Another parameter used for the analysis of this paper is the Bayesian Information Criteria (BIC). BIC is used to choose the optimal number of factors when q is not fixed and the number of factor and lag length where there are AR components in the specification. The BIC statistics provides a simple but accurate approximation of two times the log Bayes factor. For this reason, the baseline for the model comparison is a saturated model that fits the data perfectly. The BIC for a linear regression model k is written as

$$BIC_k = n \log\left(1 - R_k^2\right) + p_k \log n \quad (10)$$

where $R^2{}_k$ is the R^2 from the least squares fit and p_k is the number of coefficients in the model excluding the intercept. A negative BIC value indicates superior prediction of model k in comparison to the saturated model. The relevant statistical values which are based on other model comparisons can be found by simply taking the difference of two BIC statistics, i.e.

$$2 \log B_{12} \approx BIC_2 - BIC_1 \quad (11)$$

The BIC approximation has been developed for a number of standard statistical procedures, such as linear regression, analysis of variance, logistic regression, log-linear modelling, event-history analysis and structural equation models. Bayes factor codifies rational rules for the evaluation of evidence.

Other than the AIC and BIC parameters, Autocorrelation function (ACF) is used in the analysis in order to identify the seasonality of the data. For this case, the ACF indicates the specific situations and also to determine the stationary pattern of the data. In addition, the ACF was also being used to recognize appropriate models for nonstationarity of the random data. In addition to the ACF analysis, the Partial Autocorrelation Function (PACF) approach is also needed in order to identify the extent of relationship between current values of variable with the previous values. For this situation, the same variable is used between the current and previous elements in order to retain the same statistical effects to all related constant parameters.

Both graphical ACF and PACF are computational constructed by performing the data set using the SPSS and Minitab software packages. Finally, the implementations of both ACF and PACF with the variable amplitude (having a nonstationary behaviour) fatigue data are vital due to their used to help identifying the most suitable characteristics of the ARIMA model.

4 Results and Discussion

The first step of ARIMA in the analysis of this paper is to inspect the sample by ACF, as its distribution of the given fatigue random data (refer to Fig. 1 for the data) is illustrated in Fig. 3. The term autocorrelation coefficient measure the correlation of data with itself, lagged either by 1 or 2 or 3 or more. Since the sample of the ACF values were large, therefore, the data was assumed to have the nonstationary pattern.

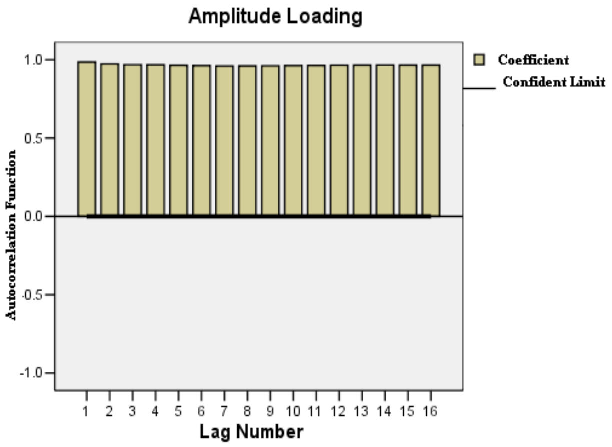

Fig. 3. The ACF distribution of the original data condition

The next step is to observe the PACF distribution. In this aspect, PACF were used to measure the degree of association between lag *t* and lag *t+q,* when the effect of other time lags 1, 2, 3, ……., up to *q-1* (*t* and *q* are defined as time and time lags, respectively). The main purpose of this analysis is to identify an appropriate ARIMA model, which is suitable for analyzing random fatigue time history. The PACF distribution is illustrated in Fig. 4 and it shows a significant larger spike followed by smaller spikes at the lag value higher than unity. Thus, it is suggested that the data can be formed stationary after performing the first ARIMA difference of random fatigue time history.

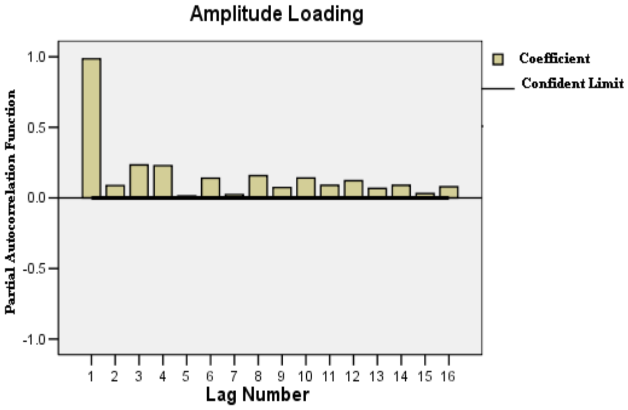

Fig. 4. The PACF distribution of the original data condition

The first ARIMA difference of the original nonstationary data (refer to Fig. 1) was then performed in order to observe stationary. The ACF and PACF were then calculated and plotted in Fig. 5 and 6, respectively. The result showed that the ACF distribution was drastically declined after the first lag. As the continuation to this situation, it is formed that the PACF has one significant spike. The results verified the earlier argument, as the results presented in Fig. 4, with respect to the nonstationarity characteristic of the data.

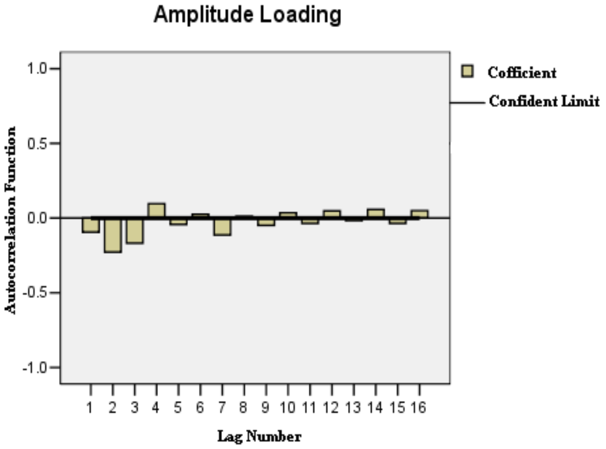

Fig. 5. The ACF distribution for the analysed data after the first difference in ARIMA

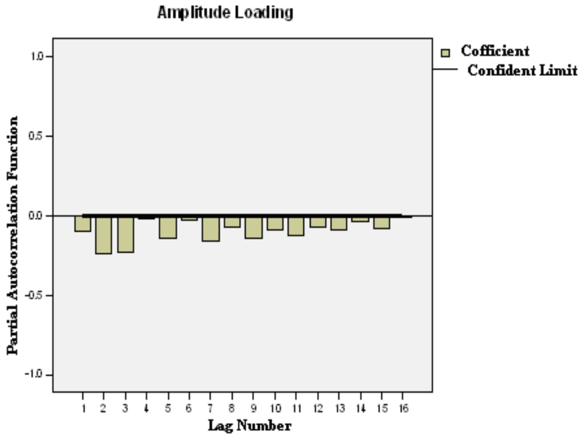

Fig. 6. The PACF distribution after the first difference in ARIMA

This situation are illustrated by a random data set as in Fig. 7, Fig. 8 and Fig. 9, for which this data was experimentally measured for the purpose of this study on the lower suspension arm of a car after applied the ARIMA model process, for which this car was travelling over a country road surface.

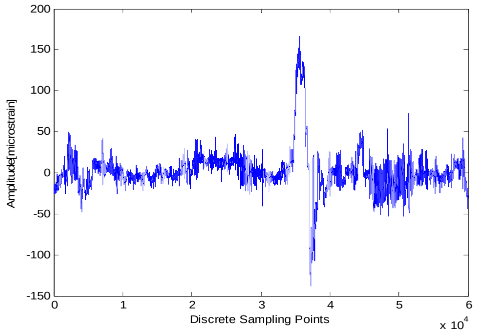

Fig. 7. A Fatigue data which was measure by Model ARIMA (0,1,0)

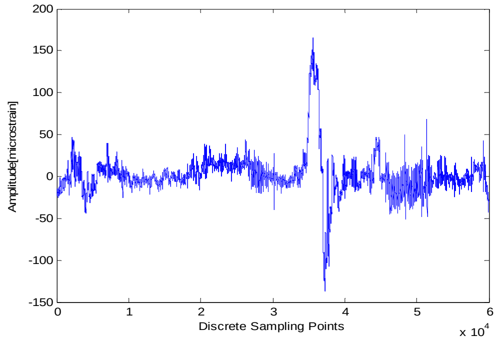

Fig. 8. A Fatigue data which was measure by Model ARIMA (0,1,1)

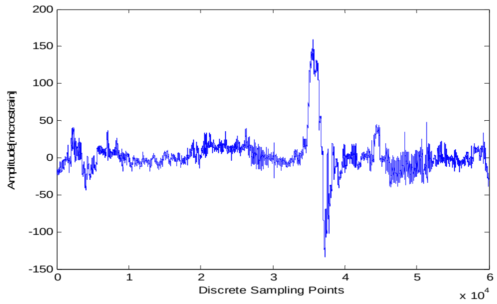

Fig. 9. A Fatigue data which was measure by Model ARIMA (1,1,1)

In order to determine the model which follows the ARIMA approach fit the best, two criteria were used in the analysis. The first criterion is the AIC value and the second is the MSE value. These values are tabulated in Table 2. The MSE and AIC formulations can be referred to Eq. (7) and (8), respectively. From this table, The ARIMA (1,1,1) is defined as AR (1), I (1) and MA (1) where one Autoregressive (AR), with only the first different and Moving Average (MA) is the first step ahead.

Table 2. Global statistical parameter for the nonstationary fatigue data

Statistical Criteria	ARIMA Model		
	ARIMA (0,1,0)	ARIMA (0,1,1)	ARIMA (1,1,1)
Kurtosis	13.75	13.93	14.26
Skewness	1.87	1.89	1.94
AIC	347294.1	346164.8	337796.3
BIC	347312.1	346164.8	337832.4

The information from this table can be used to identify the suitable ARIMA model, and it is based on the AIC and BIC values. Accordingly, it showed that ARIMA (1,1,1) is the best approach, since the lowest error value of 337796.3 microstrains was produced from the computational analysis using the Minitab software. This value is lower than the error produced by the ARIMA (0,1,1) and ARIMA (0,1,0) approaches. According to the computational analysis in Mintab using the BIC criterion, it showed that the ARIMA (1,1,1) model produced the lowest error value compared to other two models, i.e. at 337832.4. Hence, it was found that (similar to the AIC criterion) this ARIMA model is the suitable model for the nonstationary fatigue data set. Based on two statistical criteria listed in Table 2, finally, it can be concluded that the smallest MSE values was obtained from the ARIMA (1,1,1) processing for both AIC and BIC statistical criteria.

Figure 10 shows the probability distribution plot for model ARIMA (0,1,0). It is used to view and compare the shape of distribution curves and to view areas under distribution curves corresponding to either probabilities or data values. The skewness value for model ARIMA (0,1,0) was found to be 1.87. Finally the distribution exhibited skewed to the right.

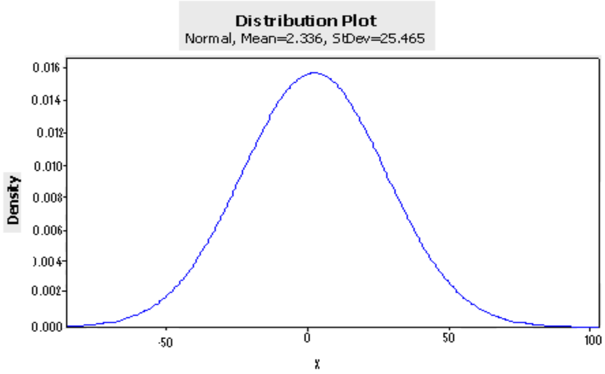

Fig. 10. Normal Plot of ARIMA (0,1,0)

Figure 11 shows the probability distribution plot for model ARIMA (0,1,1). It is used to view and compare the shape of distribution curves and to view areas under distribution curves corresponding to either probabilities or data values. The skewness value for model ARIMA (0,1,1) was found to be 1.89. Finally the distribution exhibited skewed to the right.

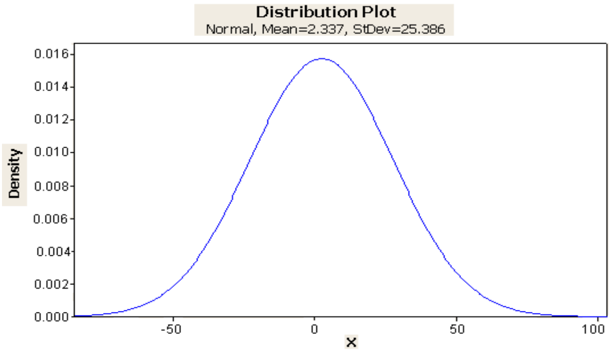

Fig. 11. Normal Plot of ARIMA (0,1,1)

Figure 12 shows the probability distribution plot for model ARIMA (1,1,1). It is used to view and compare the shape of distribution curves and to view areas under distribution curves corresponding to either probabilities or data values. The skewness value for model ARIMA (1,1,1) was found to be 1.94. Finally the distribution exhibited skewed to the right.

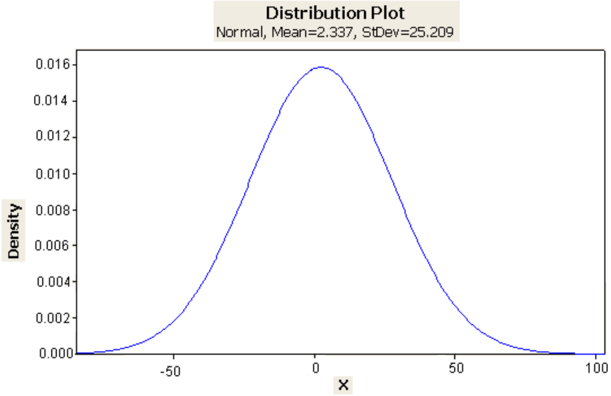

Fig. 12. Normal Plot of ARIMA (1,1,1)

5. Conclusions

The ARIMA generator has proven to be relatively portable across different systems, provide a good source of practically strong random data on most systems. Using the computational analysis of this ARIMA approaches, better and accurate results were obtained from the nonstationary fatigue loading. From the finding of this paper, it showed that ARIMA (0,1,0) was found to be the best model as it produced the lowest error value compared to the ARIMA (0,1,1) and the ARIMA (1,1,1) models. Therefore, the overall results of this study suggested that the model can give a better statistical technique, by means of the moving average approach in analysing variable amplitude fatigue loading. However, a conclusive study on this aspect should also be performed in order to know a better situation between ARIMA and fatigue damage characteristics.

References:

[1] Y. Meyer. *Wavelets: Algorithm & Applications*, SIAM, Philadelphia USA, 1993.

[2] DE. Newland, *An Introduction to Random Vibrations Spectral and Wavelet Analysis*, 3rd Edition, Longman Scientific and Technical, 1993.

[3] RG. Stockwell, L. Mansinha, RP. Lowe, Localization of the complex spectrum, *IEEE Transactions on Signal Processing*, Vol. 44, No. 4, 1996, pp. 998-1001.

[4] D. Agrawal, CC. Aggawal. On the design and quantify of privacy preserving data mining algorithms. *Proceedings of the 20th ACM SIMOD Symposium on Principles of Database Systems*, 2001, pp. 247-255.

[5] R. Agrawal, R. Srikant. Privacy-preserving data mining. *Proceeding of the ACM SIGMOD Conference on Management of Data*, 2000, pp. 439-450.

[6] B. Tacer, PJ. Loughlin. Nonstationary signal classification using the joint moments of time-frequency distributions, *Pattern Recognition*, Vol. 31, 1998, pp. 1635-1641.

[7] JS. Bendat, AG. Piersol. *Random Data: Analysis and Measurement Procedures*, 2nd Edition, Wiley-Interscience, New York, 1986.

[8] RF. Engle, CWJ. Granger. Co-integration and error correction: representation, estimation, and testing, *Econometrica*, Vol. 55, 1987, pp. 251-276.

[9] PR. Hinton. *Statistics Explained: A Guide for Social Science Students*, Routledge, London, 1995.

[10] V. Estivill-Castro, L. Brankovic. Data swapping: Balancing privacy against precision in mining for logic rules. *Proceedings of the first Conference on Data Warehousing and Knowledge Discovery (DaWaK-99)*, 1999, pp. 389-398.

[11] DA. Berry. Teaching elementary Bayesian statistics with real applications in science, *The American Statistician*, Vol. 51, 1997, pp. 247-253.

[12] S. Makridakis, SC. Wheelwright, RJ. Hyndman. *Forecasting Methods and Applications*, 3rd Edition, John Wiley & Sons,

1998.

[13] J. Albert. Teaching Bayes' rule: A data-oriented approach, *The American Statistician*, Vol. 51, 1997, pp. 47-253.

[14] MG. Kendall, A. Stuart, K. Ord. *The Advanced Theory of Statistics*, Charles Griffin, London, 1983.

[15] T. Kisinbay. *Forecasting inflation with diffusion index and non-linear models*. Working paper, York University, 2001.

[16] DS. Moore. Bayes for beginners? Some reason to hesitate, *The Americans Statistician*, Vol. 51, 1997, pp. 247-253.

[17] TH. Wonnacott, RJ. Wonnacott. *Introductory Statistics for Business and Economics*, 4th edition, John Wiley & Sons, New York, 1997.

[18] MC. Lovell. Data mining, *Review of Economics and Statistics*, Vol. 65, pp. 1-12, 1983.