

## MODEL BUILDING FOR AUTOCORRELATED PROCESS CONTROL: AN INDUSTRIAL EXPERIENCE

Djauhari, M.A., S.L. Lee and Z. Ismail

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Johor, Malaysia

Received 2014-01-12; Revised 2014-01-15; Accepted 2014-03-29

### ABSTRACT

We show that many time series data are governed by Geometric Brownian Motion (GBM) law. This motivates us to propose a procedure of time series model building for autocorrelated process control that might consist of two steps. First, we test whether the process data are governed by GBM law. If it is affirmative, the appropriate model is directly given by the properties of that law. Otherwise, we go to the standard practice at the second step where the best model is constructed by using ARIMA method. An industrial example will be reported to demonstrate the advantages of that procedure. In that example, a comparison study with ARIMA method will be reported to illustrate the effectiveness and efficiency of the GBM-based model building.

**Keywords:** Box-Jenkins Method, Control Charts, Log Normal Distribution, Statistical Process Control, Stochastic Process

### 1. INTRODUCTION

Statistical Process Control (SPC) is a powerful method to deliver high quality of products by monitoring and controlling the production process. If “monitor” reflects the awareness of the state of a process, “control” means setting standards, measuring actual performance and taking corrective action. Among the SPC basic tools, control chart is widely used to detect changes in that process. Since it was introduced in the early twentieth, its application can be found in wider and wider areas of process-based scientific investigations (Tsung and Wang, 2010). As remarked in Montgomery (2012), this is due to the fact that it is simple to implement and provides an effective means to understand the history of the process and detect process changes.

In classical SPC, the basic assumption is that the observations within and between samples are independent and identically distributed (i.i.d.). However, in practice, it is hard to achieve that assumption. On the other hand, if that assumption is violated, then the implementation of the control chart might be misleading since the presence of autocorrelation has a significant effect on the performance of control chart.

**Corresponding Author:** Djauhari, M.A., Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Johor, Malaysia

This remark can be found in the literature. See, for example, (Haiyu, 2010; Costa and Castagliola, 2011; Tasdemir, 2012; Goswami and Dutta, 2013; Singh and Prajapati, 2013).

In light of the latent detrimental effects of autocorrelation on the control chart, (Snoussi, 2011; Karaoglan and Bayhan, 2012; Kim *et al.*, 2012; Areepong, 2013) and many others suggested to fit a time series model to the process data and then apply the control chart to the residuals. They have remarked that this method is appropriate. From the literature, we learn that the standard procedure in time series modeling is by using ARIMA also known as Box-Jenkins method. However, this method might be laborious especially in the case when the time series data are governed by a particular mathematical law. Based on ARIMA, it is very often in practice that even a satisfactory model is not easy to obtain. This is not the case if we know that the time series data are governed by a mathematical law.

Under such law, the appropriate model can directly be obtained from the properties of that law and, therefore, neither model identification process nor model verification process is necessary. These results motivate us to develop a procedure of time series model

building for autocorrelated process monitoring which consists of two steps where ARIMA is used only if the process is not a GBM process.

The rest of the paper is organized as follows. In section 2, we present a real problem in a cocoa manufacturing industry that motivates this study. Later, we recall the notion of GBM process and present its practical guidance of the proposed control charting procedure in section 3. In section 4, the industrial problem presented in the second section is revisited to illustrate the advantages of the proposed procedure. Promising results issued from a comparison study with ARIMA method will also be reported. Finally, concluding remarks will be delivered in the last section.

## 2. MOTIVATION-AN INDUSTRIAL PROBLEM

A Malaysian manufacturing company produces cocoa powder to fulfill local as well as export demands. The name of the company is kept undeliverable due to its confidentiality. An important characteristic which determines the quality of cocoa powder is the content of fat. It needs to be controlled since it has an important functional impact on the end-products in which cocoa powder is used.

During a period of production process control, we found that fat content data in cocoa powder is time dependent. Therefore, the classical control charting procedure cannot directly be used to control the production process of cocoa powder. It should be used on the residuals after having fitted an appropriate or at least a satisfactory time series model.

In the search for the best model, before going directly to use ARIMA, in this study we propose to test first whether the time series data of an autocorrelated process under study are governed by GBM law. If it is a GBM process, then the appropriate model is determined by the properties of GBM. Otherwise, we use ARIMA method of model building. This proposal will be elaborated in the next sections.

## 3. IS THE PROCESS GOVERNED BY GBM LAW?

### 3.1. Recall on GBM Process

A GBM process  $X_t$  is a stochastic process satisfying the following stochastic differential equations:

$$dX_t = \mu X_t dt + \sigma X_t dW_t \quad (1)$$

where,  $W_t$  is a Wiener process and  $\mu$  and  $\sigma$  are the drift and volatility, respectively. The solution of Equation 1 is well very known and can be found easily in the literature. Here we briefly recall that solution.

Let  $X_0$  be an initial value satisfying Equation 1. Then, the solution of Equation 1, Sheldon (2011), is given by:

$$\ln X_t - \ln X_0 = (\mu - \sigma^2 / 2)t + \sigma W_t$$

or, equivalently:

$$X_t = X_0 \exp((\mu - \sigma^2 / 2)t + \sigma W_t)$$

As a corollary, since  $W_t$  is a Wiener process, we have:

$$X_t / X_{t-1} = \exp((\mu - \sigma^2 / 2) + \sigma Z_t)$$

where,  $Z_t = W_t - W_{t-1}$  is standard normal random variable. Consequently, the logarithmic return  $R_t = \ln(X_t/X_{t-1})$  is simply:

$$R_t = (\mu - \sigma^2 / 2) + \sigma Z_t$$

This corollary implies that, since  $Z_t = W_t - W_{t-1}$  and  $W_t$  is a Weiner process,  $R_t$ 's are i.i.d and normally distributed. Therefore,  $X_t/X_{t-1}$  is a lognormal random variable. Furthermore, Sheldon (2011),

$$R_t = \theta R_{t-1} + \varepsilon_t \quad (2)$$

where, the error terms  $\varepsilon_t$ 's are i.i.d. normal random variables with mean zero and constant variance. In other words,  $R_t$  is an autoregressive process of order one; AR(1).

### 3.2. Proposed Procedure

The idea to use GBM-based model building for an autocorrelated process is basically inspired by the works of economists to model the behavior of economic commodities' prices. Its history started with the work of Bachelier in early nineteenth century and was popularized by Paul Samuelson, a Noble laureate in economics in 1970 s (Djauhari and Gan, 2013). Since then, there is a great number of applications of GBM process in different areas such as, for example, strategic and planning decisions in supply chain (Wattanarat, 2010), energy prices (Esunge and Snyder-Beattie, 2011), mid-term planning for thermal electricity

production system (Kovacevic and Paraschiv, 2013) and many others. Motivated by those significant works, in what follows, we propose a new procedure of model building that might consist of two steps. First, we start by testing whether the time series data are governed by GBM law. If it is affirmative, as shown in Equation 2, the logarithmic returns  $R_t$  is an AR(1) process. Thus, the fitted model for each time  $t$  is:

$$\hat{X}_t = X_{t-1} (X_{t-1} / X_{t-2})^{\hat{\theta}} \quad (3)$$

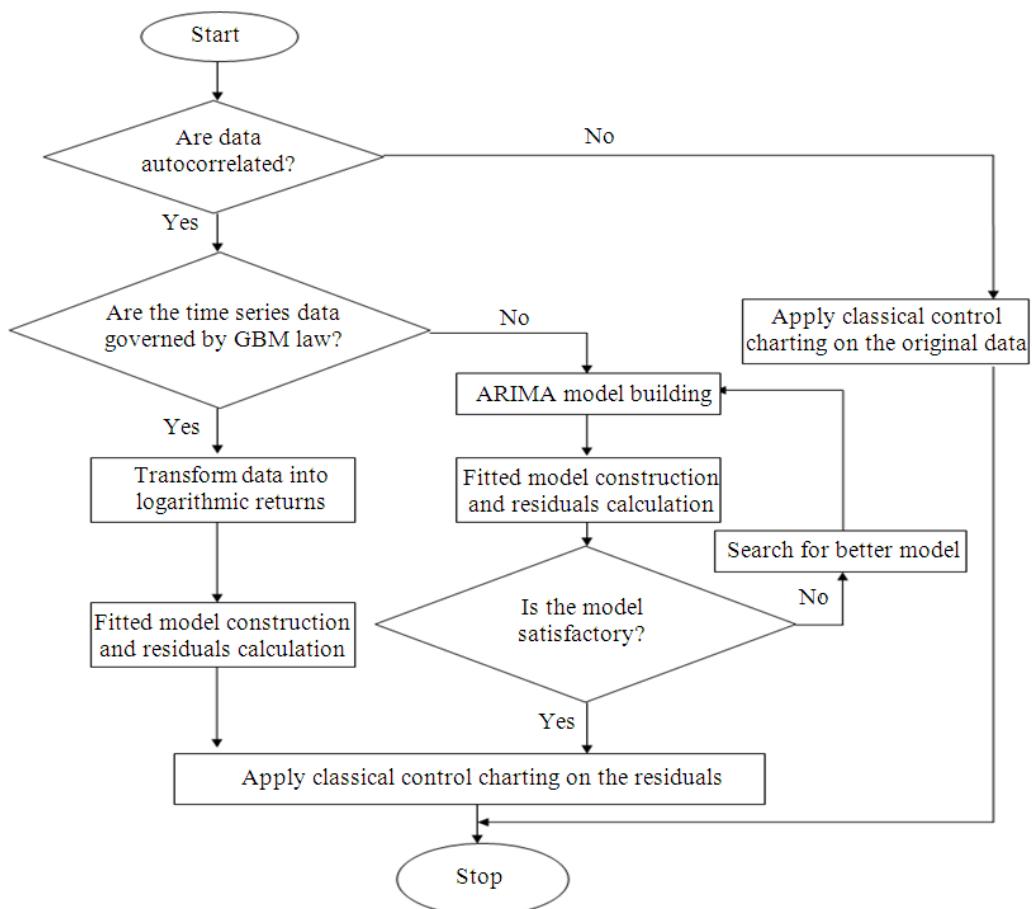
where,  $\hat{\theta}$  is a least square estimate of  $\Theta$  in Equation 2. Otherwise, we go to the second step where the fitted model is constructed by using ARIMA. In both cases, classical control charting is conducted on the residuals.

Diagrammatically the proposed control charting of an autocorrelated process is summarized in **Fig. 1**. It

shows that, if the process is a GBM, the method of model building is simpler than ARIMA. Neither model identification nor model verification is necessary as usually encountered in the latter and the appropriate model is given by Equation 3. These advantages of GBM process will be exploited in the next section when we deal with fat content process.

#### 4. INDUSTRIAL EXAMPLE

We return to the industrial problem of fat content process quality control presented in Motivation section. In **Fig. 2a** we present the run chart of fat content data which shows that the autocorrelation is seemingly presence. In the next paragraph a further analysis to confirm the presence of autocorrelation is conducted before the control charting procedure proposed in **Fig. 1** is considered.



**Fig. 1.** Flow chart of the proposed procedure of control charting

## 4.1. Does Fat Content Data Follow GBM Process?

### 4.1.1. GBM Verification

First, we verify the presence of autocorrelation. Since this is the only concern, instead of using ACF and PACF, here we use the simpler method suggested in NIST/SEMATECH (2012) by drawing the lag-1 scatter plot followed by a confirmatory analysis based on Durbin-Watson test. These two statistical tools are to visualize and confirm the presence of autocorrelation, respectively. The lag-1 scatter plot in **Fig. 2b** strongly indicates that the autocorrelation is present in fat content data; the correlation between  $X_t$  and  $X_{t-1}$  cannot be ignored. To confirm that indication, Durbin-Watson test D is used. From the data we obtain  $D = 0.0004$ . On the other hand, at 5% significance level, the critical points are  $D_L = 1.70166$  and  $D_U = 173194$ . Since  $D < D_L$ , we conclude that the autocorrelation is present.

Second, after having confirmed that the process is autocorrelated, the original data  $X_t$  is transformed into logarithmic return  $R_t$ . The run chart, lag-1 scatter plot, QQ-plot and histogram of the transformed data are presented visually in **Fig. 3**.

The four graphs in that figure strongly indicates the independency and normality of  $R_t$ . To confirm the

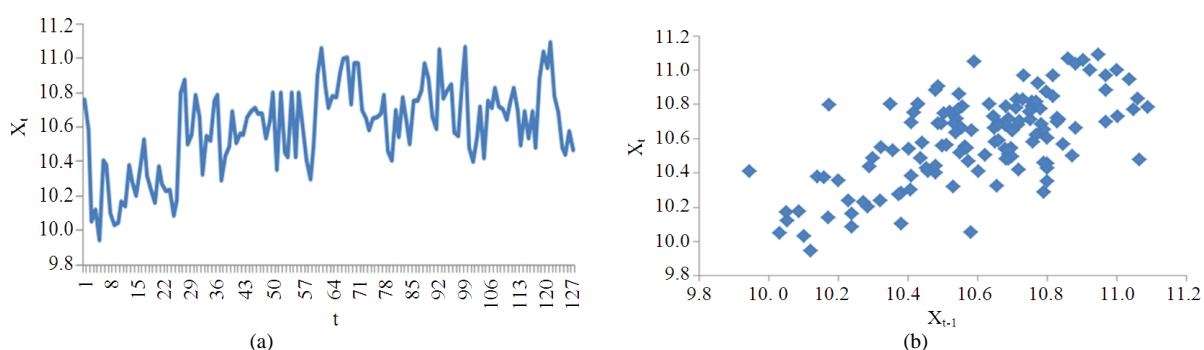
independency, we again use the Durbin-Watson test D. From the data of  $R_t$ , we obtain  $D = 2.4994$ . For 5% significance level, the critical points are  $D_L = 1.70049$  and  $D_U = 1.73100$ . Since  $D > D_U$ , then  $R_t$ 's are independent. To test the normality, as suggested in NIST/SEMATECH (2012), we use Anderson-Darling test AD. The data give  $AD = 0.205$  and p-value = 0.869. Thus, at 5% significance level, the assumption that  $R_t$  is normally distributed cannot be rejected. From the above analysis, since  $R_t$ 's are i.i.d. and normally distributed, we conclude that the time series data of fat content is a GBM process.

### 4.1.2. Fitted Model

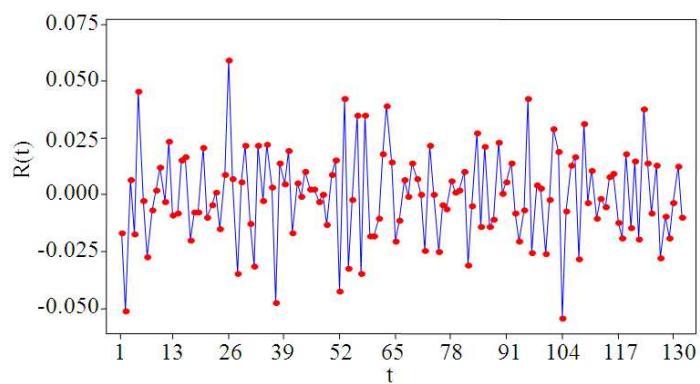
Since fat content data are governed by GBM law which implies that  $R_t$  is an AR (1) process, from Equation 3 the fitted model for fat content data is:

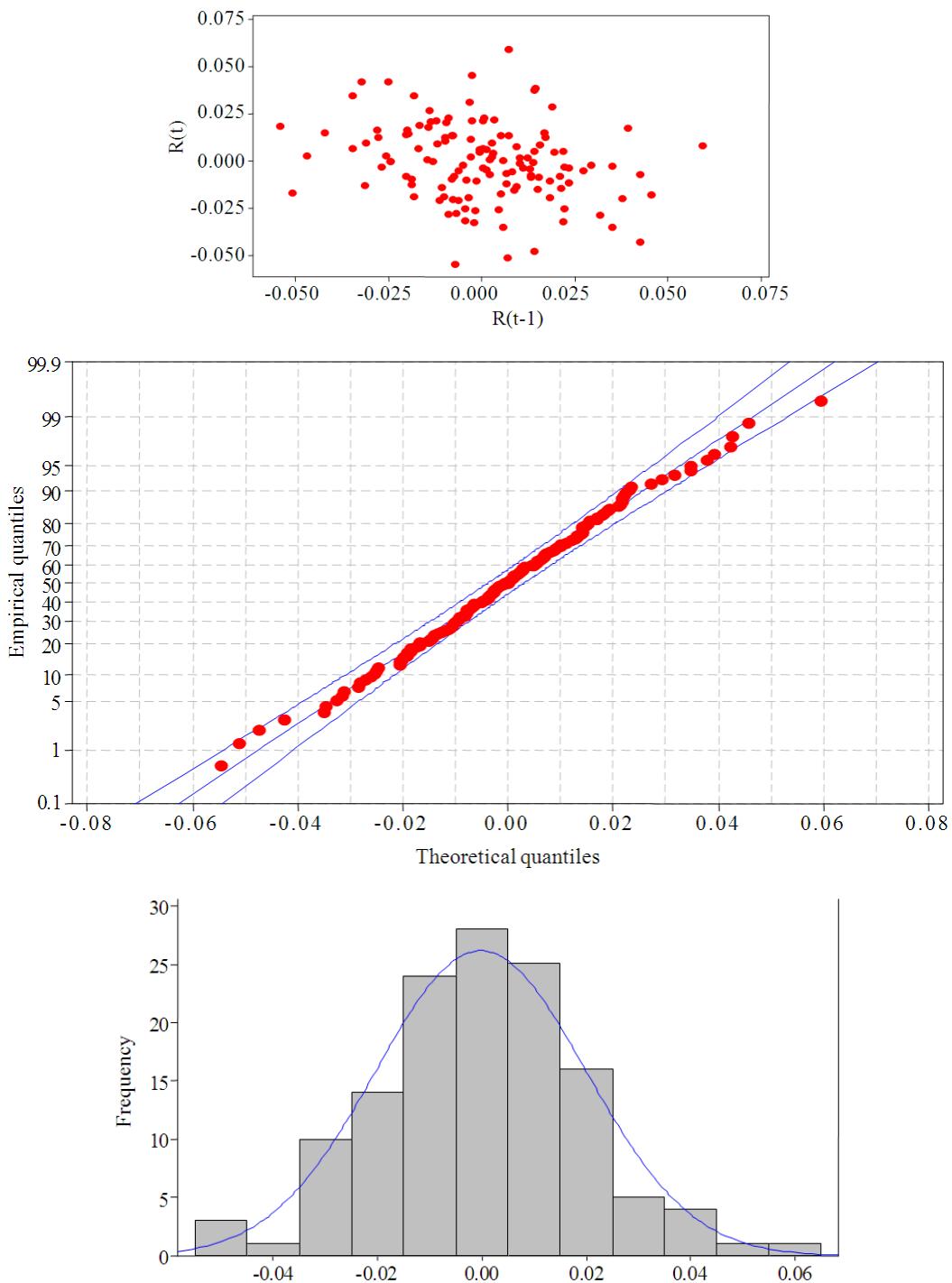
$$\hat{X}_t = X_{t-1} (X_{t-1} / X_{t-2})^{-0.23255} \quad (4)$$

The exponent is the regression coefficient when we regress  $R_t$  with respect to  $R_{t-1}$ . It is worthwhile to note that the MAPE of the fitted model in (4) is 1.54% which is far less than 10%. This means that, see Gundalia and Dholakia (2013), the model is highly accurate.



**Fig. 2.** (a) Run chart and (b) lag-1 scatter plot of fat content data





**Fig. 3.** Run chart, lag-1 scatter plot, QQ-plot and histogram of logarithmic returns

Before we proceed to the control charting procedure, we check all assumptions of the residuals

$e_t = X_t - \bar{X}_t; t = 3, 4, \dots$ . For this purpose, we repeat the procedure in previous sub-sub-section of GBM

verification and we come up with the following conclusion:

- **Figure 4** shows four diagnostic graphical tools suggested in NIST/SEMATECH (2012); run chart, lag-1 scatter plot, QQ-plot and histogram of the residuals issued from GBM. The run chart indicates the stationary of the residuals with constant mean and variance and thus needs further analysis
- The lag-1 scatter plot in the second graph strongly indicates the independency. According to Durbin-Watson test,  $D = 2.14248$  and the critical points at 5% significance level are  $D_L = 1.69931$  and  $D_U = 1.73005$ . Since  $D > D_U$ , the independency assumption is not rejected
- The last two graphs are the histogram and QQ-plot of the residuals. The QQ-plot indicates the normality of residuals while the histogram shows how close the distribution of residuals to normality. According to Anderson-Darling test which gives  $AD = 0.170$  with  $p\text{-value} = 0.932$ , at 5% significance level, normality assumption is fulfilled

#### 4.1.3. Fat Content Process Monitoring Based on GBM

The above analysis confirms that the residuals are i.i.d. and normally distributed. Therefore, fat content process can be monitored based on the GBM residuals data. In **Fig. 5** the I-MR charts are presented to monitor the fat content.

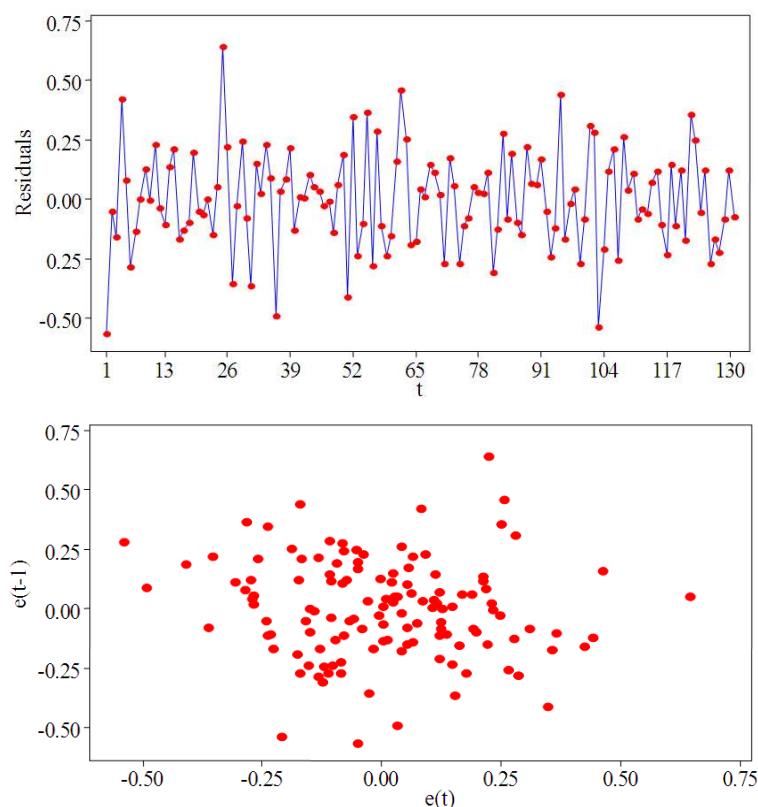
From **Fig. 5b**, we learn that an out-of-control signal occurs in MR-chart at sample 103. To illustrate the significant role of GBM model building, in the next subsection, we compare the history of process performance represented by this control chart with that issued from ARIMA model.

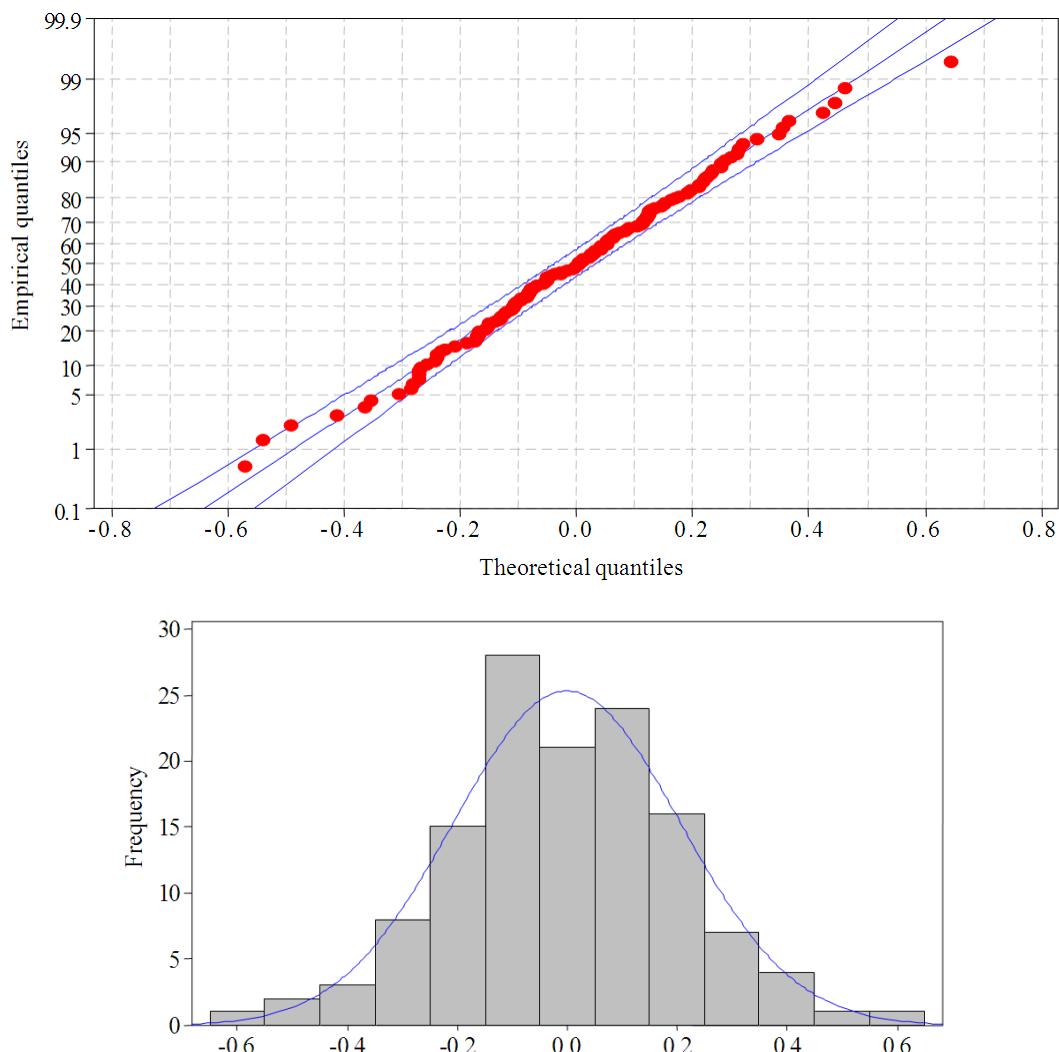
#### 4.2. ARIMA Method

##### 4.2.1. Model Building

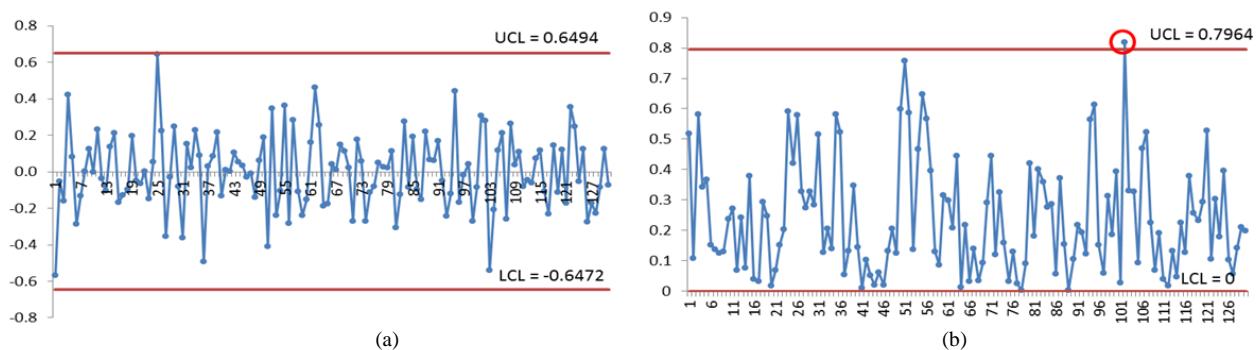
A comparison study has been done to see what if ARIMA model is used for fat content process control. After analyzing the behaviour of the ACF and PACF of those data, the best fitted model is ARIMA(2,1,2):

$$\hat{X}_t = 0.47732X_{t-1} + 0.80057X_{t-2} - 0.27788X_{t-3} + \varepsilon_t + 0.16369\varepsilon_{t-1} - 0.86999\varepsilon_{t-2} \quad (5)$$





**Fig. 4.** Run chart, lag-1 scatter plot, QQ-plot and histogram of GBM residuals.



**Fig. 5.** I-MR charts based on GBM

With MAPE equals 1.38%. This value of MAPE signifies that the model is as highly accurate as GBM model.

Similar tests as in the previous sub-section are used to determine whether all assumptions of the residuals are fulfilled. In **Fig. 6**, four diagnostic graphs indicates the fulfilment of those assumptions. Specifically, at 5% significance level, we cannot reject that the residuals are i.i.d. ( $D_L = 1.70049$ ,  $D_U = 1.73100$  and  $D = 2.20652 > D_U$ ) and normally distributed ( $AD = 0.301$  with  $p\text{-value} = 0.576$ ).

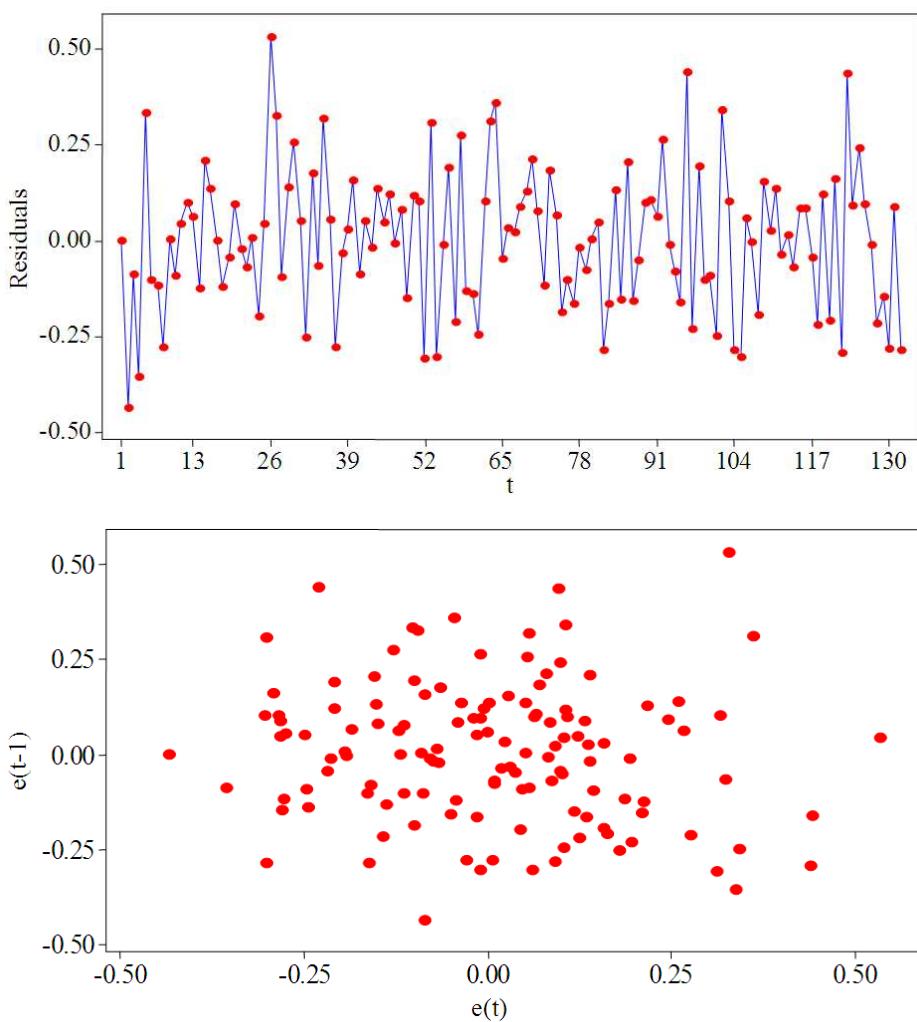
#### 4.2.2. Fat Content Process Monitoring Based on ARIMA

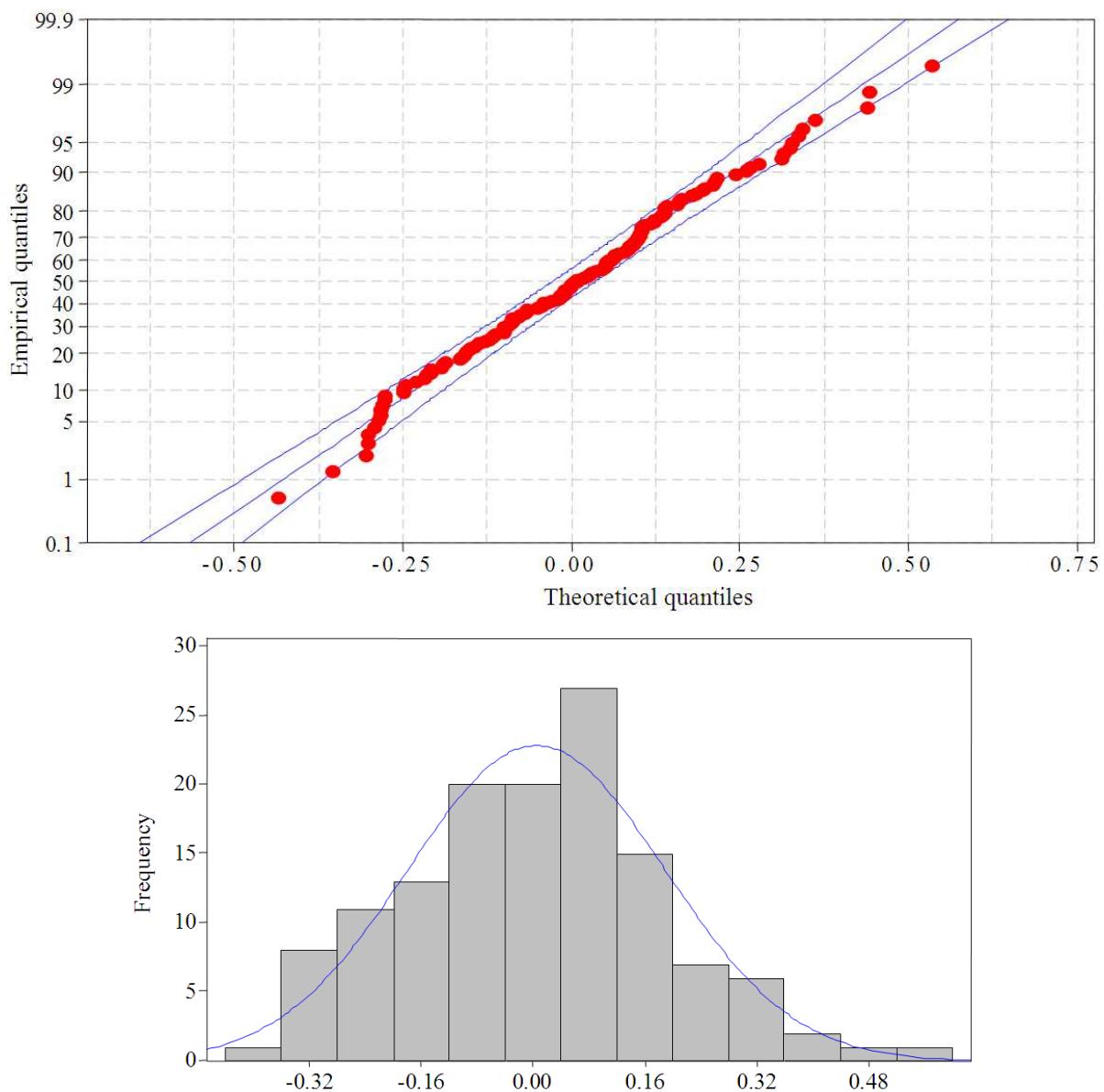
In **Fig. 7**, we present the I-MR charts on the residuals issued from ARIMA(2,1,2). In MR-chart,

see **Fig. 7b**, three out-of-control signals (samples 5, 96 and 123) occur. This result is different from that given by GBM. Since the residuals given by GBM is more preferable compared to those issued from ARIMA(2,1,2), in this study the I-MR charts constructed based on GBM process is used for further actions to improve the process.

#### 4.3. Comparison of Both Methods

The MAPE of GBM-based model building and that of ARIMA are 1.54% and 1.38%, respectively. Actually, there is no significant difference between their accuracy. Let us consider the 95% confidence interval of Mean Absolute Error (MAE). According to the residuals of ARIMA(2,1,2), that confidence interval is [0.1276, 0.1663].



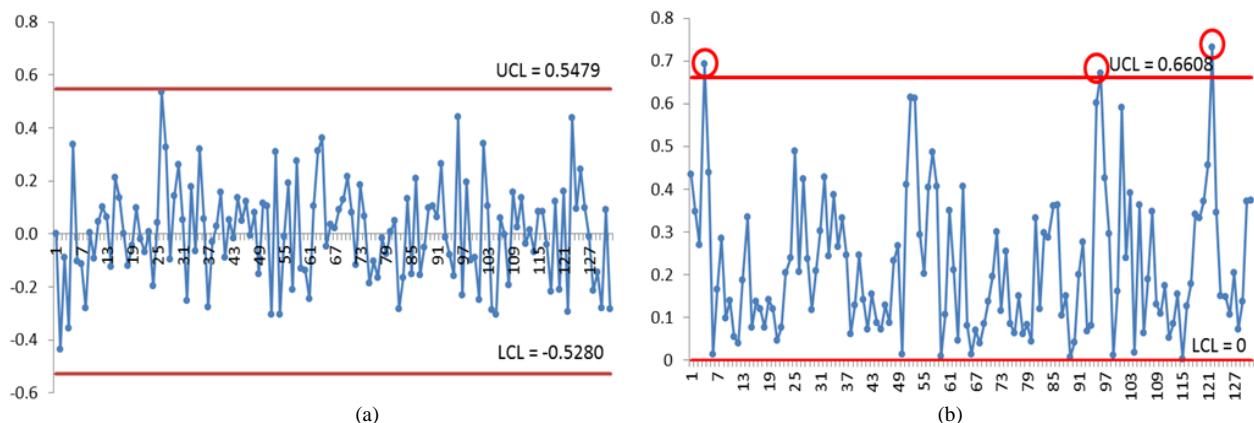


**Fig. 6.** Run chart, lag-1 scatter plot, QQ-plot and histogram of ARIMA residuals

On the other hand, the MAE of GBM model is 0.16267. Since this value is in that interval, then the GBM-based model Equation 4 is as highly accurate as the model issued from ARIMA Equation 5. However:

- Anderson-Darling test for ARIMA(2,1,2) and GBM gives the p-value 0.576 and 0.932, respectively. This means that, in terms of the degree of normality, the residuals from GBM model are higher than those issued from ARIMA(2,1,2)

- Durbin Watson test D for ARIMA(2,1,2) and GBM model are 2.20652 and 2.14248, respectively. Since GBM gives the value of D closer to 2, this means that the degree of independency of the residuals issued from GBM is higher than that from ARIMA(2,1,2)
- The histogram of residuals issued from GBM is closer to normality compared to those given by ARIMA(2,1,2)

**Fig. 7.** I-MR chart based on ARIMA

Therefore, according to these results, the use of GBM-based model Equation 4 in fat content process control is more preferable than ARIMA(2,1,2) in Equation 5; the residuals issued from the former are better than those given by the latter.

## 5. CONCLUDING REMARKS

The proposed procedure of model building for an autocorrelated process might consist of two steps. First, start by testing whether the process is governed by GBM law. If it is a GBM process, then the fitted model is given by Equation 3. Otherwise, go to the second step where the standard ARIMA model building is used.

In the first step, unlike the second, there is no need to conduct model identification nor model verification. Accordingly, there is no need to construct the ACF and PACF of the process which might not be easy to analyse. All we need is to (i) transform the original data into logarithmic returns  $R_t$ , (ii) estimate the parameter of the corresponding AR(1) process of  $R_t$  and (iii) use that estimate to get the appropriate model such as Equation 4. This is the advantages once we know that the process is a GBM process.

The experience with fat content process monitoring demonstrates the effectiveness and efficiency of the proposed model building. In that example, a comparison study with ARIMA method shows that the fitted model issued from GBM-based model building is more preferable than that given by ARIMA; the quality of residuals issued from the former is better than that given by the latter. Moreover, if ARIMA usually requires several iterations before producing the desired estimator and thus needs special statistical package, GBM-based model building does not need such package.

To close this presentation, we conclude that once we know that the process is governed by GBM law, time series model building becomes efficient and most importantly effective in monitoring autocorrelated process.

## 6. ACKNOWLEDGEMENT

Special thanks go to the Ministry of Higher Education, Government of Malaysia, for financial support under Research University Grant vote number 4F260 and to Universiti Teknologi Malaysia for providing research facilities.

## 7. REFERENCES

- Areepong, Y., 2013. A comparison of performance of residual control charts for trend stationary AR(p) processes. *Int. J. Pure. Applied. Math.*, 85: 583-592. DOI: 10.12732/ijpm.v85i3.13
- Costa, A.F.B. and P. Castagliola, 2011. Effect of measurement error and autocorrelation on the  $\bar{X}$  chart. *J. Applied. Stat.*, 38: 661-673. DOI: 10.1080/02664760903563627
- Djauhari, M.A. and S.L. Gan, 2013. Minimal spanning tree problem in stock networks analysis: An efficient algorithm. *Phys. A*, 392: 2226-2234. DOI: 10.1016/j.physa.2012.12.032
- Esunge, J.N. and A. Snyder-Beattie, 2011. Dissecting two approaches to energy prices. *J. Math. Stat.*, 7: 98-102. DOI: 10.3844/jmssp.2011.98.102
- Goswami, A. and H.N. Dutta, 2013. Time series approaches to statistical process control. *Int. J. Innovative Res. Comput. Sci. Technol.*, 1: 34-38.

- Gundalia, M.J. and M.B. Dholakia, 2013. Modelling pan evaporation using mean air temperature and mean pan evaporation relationship in middle south Saurashtra Region. *Int. J. Water Res. Environ. Eng.*, 5: 622-629. DOI: 10.5897/IJWREE2013.0426
- Haiyu, W., 2010. Statistical process control of stationary autocorrelated process. Proceedings of the International Conference on Intelligent Computation Technology and Automation, May 11-12, Changsha, China, pp: 386-389. DOI: 10.1109/ICICTA.2010.443
- Karaoglan, A.D. and G.M. Bayhan, 2012. ARL performance of residual control charts for trend AR(1) process: A case study on peroxide values of stored vegetable oil. *Sci. Res. Essays*, 7: 1405-1414. DOI: 10.5897/SRE11.1801.
- Kim, S.B., W. Jitpitaklert, S.K. Park and S.J. Hwang, 2012. Data mining model-based control charts for multivariate and autocorrelated processes. *Expert Syst. Appl.*, 39: 2073-2081. DOI: 10.1016/j.eswa.2011.08.010.
- Kovacevic, R.M. and F. Paraschiv, 2013. Medium-term planning for thermal electricity production. *OR Spectrum*. DOI: 10.1007/s00291-013-0340-9
- Montgomery, D.C., 2012. Introduction to Statistical Quality Control. 7th Edn., John Wiley and Sons, Inc., New York, USA. ISBN-10: 9781118146811, pp: 768.
- NIST/SEMATECH, 2012. Engineering statistics: E-handbook of statistical methods.
- Singh, S. and D.R. Prajapati, 2013. Performance of CUSUM and EWMA charts for serial correlation. *TQM J.*, 25: 309-324. DOI: 10.1108/17542731311307474
- Sheldon, M.R., 2011. An Elementary Introduction to Mathematical Finance. 3rd Edn., Cambridge University Press, New York, ISBN-10: 9780521192538. pp: 305.
- Snoussi, A., 2011. SPC for short-run multivariate autocorrelated processes. *J. Applied Stat.*, 38: 2303-2312. DOI: 10.1080/02664763.2010.547566
- Tasdemir, A., 2012. Effect of autocorrelation on the process control charts in monitoring of a coal washing plant. *Physicochem. Probl. Miner. Process*, 48: 495-512. DOI: 10.5277/ppmp120215
- Tsung, F. and K. Wang, 2010. Adaptive Charting Techniques: Literature Review and Extensions. In: *Frontiers in Statistical Quality Control 9*, Lenz, H.J., Wilrich, P.T. and Schmid, W. (Eds.), Physica-Verlag, Berlin, ISBN-10: 9783790823790, pp: 19-35.
- Wattanarat, V., 2010. Demand and Price Forecasting Models for Strategic and Planning Decisions in a Supply Chain. Proceeding of the School of Information and Telecommunication Engineering Tokai University, (ITE' 10), pp: 37-42.