Graphic communication in detecting outlier cases in time column control diagram

E Enjang*, M Aliyudin, F S Nurdin, M W Laksana and S Sarbini
UIN Sunan Gunung Djati Bandung, Bandung, Indonesia

*enjang_as@uinsgd.ac.id

Abstract. Control diagrams are graphical communication media that can be used as tools in the process of controlling statistics statistically to assist in monitoring and improving a quality process by helping to separate specific cause and general cause variables. Basically the control diagram plots the area of acceptance of a hypothesis test. This study aims to detect disruptive causes in the control diagram in time series cases and to justify the statistical model in accordance with the procedure. A process is said to be controlled if the observational data from that process behaves like random iid variables (independent identically distributed), and if it is not controlled, it is out of control. The study uses 25 days’ time series observational data in PT. Guccitex Cimahi. In this time frame analysis, research that is influenced by specific causes can be treated as outliers of the time series model and can be detected by outlier detection methods. The results in this study explain that a traditional control diagram is very dependent on the assumption of iid (independent identically distributed). Application in real life faces observations that are not iid (independent identically distributed), which show the effects of trends, seasonality, and other influences. The control diagram is in principle the same, using time series analysis to overcome these effects by plotting the area of acceptance of a hypothesis test, but if the process is not controlled, the control diagram cannot provide any guidance, so it is necessary to do a data analysis using analysis time series.

1. Introduction

Control chart is commonly used in the process of controlling statistically, to help with supervision and recognize disturbances that greatly affect the statistical control process. In recognizing these disorders, special causative variables can be separated from common causative variables. With the variance of general interference intended as a normal variation in a control process, while a specific disturbance is referred to as a disturbance that affects a statistical control process.

The statistical control process has two aspects [1], namely the traditional approach and the automatic control process. These two points are interconnected and time series analysis is important for both. Shewhart's central concept is called "A statistically controlled situation". A process is said to be controlled, if the observational data from a process behaves like random iid variables (independent identically distributed). If this is not the case, it is said to be out of control [1]. Making control charts is basically the same as making a plot of the Liu and Tang hypothesis test [2]. And can also use other approaches that can be found in Montgomery [3]. In the control diagram traditionally using control limits that have been plotted, namely the upper control limit is $\mu + 3\sigma$ and the lower control limit is $\mu - 3\sigma$. 
Within the framework of this time series analysis, studies that are affected by specific disturbances can be treated as outliers of a coherent time model that can be detected by outlier detection methods. To create a more effective control chart, a standard control chart called a pattern analysis that can be used in time series work will be used.

In this regard, this study aims to detect disruptive causes in the control diagrams in time series as well as to justify the statistical model in accordance with the procedure.

2. Methodology

2.1. Outlier analysis

Outlier analysis takes the Fox approach by classifying outliers in a time series into innovative and additive outliers. Let $z_t$ be an outlier-free time series [1]. It is assumed that $z_t$ follows the Gaussian Auto Regressive Integrated Moving Average (ARIMA) model.

\[
\Phi(B)(1-B)^dZ_t = \Theta(B)\alpha_t \text{ or } z_t = \frac{\Theta(B)}{\Phi(B)(1-B)^d}\alpha_t \quad (1)
\]

Where $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\Theta(B) = 1 - \theta_1 B - \cdots \theta_q B^q$ is the many term in B with degrees p and q; B is a backshift operator with $B^j z_t = z_{t-j}$, d is a non-negative integer and $\{\alpha_t\}$ is a sequence of Gaussian random variables with $\mu = 0 \text{ dan } \sigma^2, \Phi(B) \text{ dan } \Theta(B)$ do not have the same factors and satisfy the usual conditions at stationary and invertibility.

An outlier additive (AO) at time index $t_0$ can be written as:

\[
Y_t = z_t + w.I_{t_0}^t \quad (2)
\]

Where $z_t$ is the outlier-free series, $w$ is a constant that shows the size of the additive outlier, and $I_{t_0}^t$ is an indicator variable for the time index $t_0$, $I_{t_0}^t = \{1 \text{ if } t = t_0 \}$ . From equation 2 an additive outlier represents a separate disturbance in $z_t$. That only affects the time series at the time index $t_0$. An innovative outlier (IO) in the time index $t_0$ can be written as:

\[
Y_t = z_t + \frac{\Theta(B)}{\Phi(B)(1-B)^d} w.I_{t_0}^t = \frac{\Theta(B)}{\Phi(B)(1-B)^d}\{\alpha_t + w.I_{t_0}^t\} \quad (3)
\]

Where $w$ is the size of the outlier and $I_{t_0}^t$ is an indicator variable at the time index $t_0$. From equation 3, an IO is a disturbance in the $\{\alpha_t\}$ sequence at the time index $t_0$ and has a dynamic effect on $z_t$ for $t \geq t_0$, such that $Y_t = z_t + w.\psi_{t-t_0}$ where $\psi_0 = 1$.

2.2. Outlier estimates

It is assumed that if given a time index $t_0$ and an ARIMA model with time series parameters $\Phi_j, \Theta_j, d$ and $\sigma^2$ are known. These parameters can be obtained by estimating from the data and we can use those estimates as a substitute for the actual parameters.

Let $Y_t = \frac{\Phi(B)(1-B)^d}{\Theta(B)}$, $Y_t$ be the residual of the observed $Y_t$ process and the weighted function $\pi$ of the ARIMA model in equation (1) with:

\[
\pi(B) = \frac{\Phi(B)(1-B)^d}{\Theta(B)} = 1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - \cdots
\]

By using filter $\frac{\Phi(B)(1-B)^d}{\Theta(B)}$ for equation (2) and using the model in equation (4.1). We use for an outlier additive (AO) as follows:
\[
\phi(B)(1 - B)^d \frac{Y_t}{\theta(B)} = \phi(B)(1 - B)^d \frac{z_t}{\theta(B)} + \phi(B)(1 - B)^d \frac{w_t I_t^{t_0}}{\theta(B)}
\]

or the same as:

\[
y_t = a_t + w_t \pi(B) \cdot I_t^{t_0}
\]

Because \(\{a_t\}\) is an iid (independent identically distributive) series and \(I_t^{t_0}\) is an indicator variable, so equation (4) is a simple linear regression model:

\[
y_t = w_t z_t + a_t
\]

Where \(z_t = \pi(B) \cdot I_t^{t_0}\) with \(\pi_0 = 1\) and \(\pi_j = 0\) if \(j < 0\). Then it is clear that the estimated OLS (ordinary least squares) about \(w_t\) is:

\[
\hat{w}_a = \frac{\sum_{t=0}^{n} y_t z_t}{\sum_{t=0}^{n} z_t^2}
\]

The estimated variance of \(\hat{w}_a\) is:

\[
\sigma_a^2 = \frac{\sigma^2}{\sum_{t=0}^{n} \pi_t^2 - t_0}
\]

Where \(n\) is the sample size and subscript \(a\) is used to mark the outlier additive. Using the results (6) and (7), we get a t-ratio:

\[
T_a^{t_0} = \frac{\hat{w}_a}{\sigma_a}
\]

Which can be used to test the significance of an additive outlier in the time index \(t_0\), for sample sizes where \(n\) and \(t_0\) are fixed, \(T_a^{t_0}\) is the normal standard approximation. In the same way for an innovational outlier (IO) of equations (1) and (3) obtained:

\[
\|y_t = w_t I_t^{t_0} + a_t \text{ at } au \ y_t = w_t z_t + a_t
\]

Where \(z_t = I_t^{t_0}\), and the OLS estimation of the outlier parameter is:

\[
\hat{w}_t = y_t \cdot t_0
\]

The estimated variance is \(\sigma_i^2 = \sigma^2\), where subscript \(i\) means an innovational outlier. Then the t-ratio for IO testing at the time index \(t_0\) is:

\[
T_i^{t_0} = \frac{y_{t_0}}{\sigma_i}
\]

For large samples \(T_i^{t_0}\) follows the standard normal distribution asymptotically.

2.3. Outlier detection

Two t-ratios of equations (8) and (11) can be used to detect the presence of an outlier (AO or IO) in the time index \(t_0\), the time index of an unknown priori. Following the approach [1], we begin with the case of a single outlier. Determine statistics:

\[
T_a = \max \{|T_a^{t_0}|_{t=1}^n\}
\]
\[ T_i = \max\{|T^c_i|\}_{t=1}^n \]  
\[ T = \max\{T_a, T_l\} \]  

In other words \(T_a\) and \(T_l\) respectively are the maximums in the test statistic modulus \(T^c_a\) and \(T^c_i\) from the data, and \(T\) is the global maximum of all test statistics. An AO is in the time index \(k\) if \(T = T_a = |T^c_a| > C\) where \(C\) is a positive constant. Also for an IO is on the time index \(h\) if \(T = T_i = |T^c_i| > C\). Constants \(C\) can be classified in two ways [1], using some Monte Carlo results to suggest values of 2.5, 3.0 and 3.5 for general use in time series analysis to calculate the critical value of the asymptotic distribution \(|T_a|\). It places that \(C = 3.5\) are slightly greater than the critical value of 5% for AR (1) models with an outlier additive [4].

**3. Results and discussion**

The data processed is in the form of data obtained from PT. Guccitex Cimahi for 25 effective working days consists and is taken in certain times in a row. The structure of the data analyzed is as follows:

**Table 1.** Working time 25 employees are taken from a certain time.

| Day | 06.00 | 10.00 | 14.00 | 18.00 | 22.00 |
|-----|-------|-------|-------|-------|-------|
| 1   | 13.2  | 13.3  | 12.7  | 13.0  | 12.1  |
|     | -     | -     | -     | -     | -     |
|     | -     | -     | -     | -     | -     |
|     | -     | -     | -     | -     | -     |
| 25  | 13.2  | 13.3  | 12.7  | 13.0  | 12.1  |

**Figure 1.** Plot x-bar chart results.
From Figures 1 and 2 there appears to be no mean subgroups and ranges that fall outside the control limits. However, one problem is clearly seen that these points tend to push the central line. This happens due to the seasonal influence of the data, with seasons spaced equal to 5 times a day during the observation period. So this control chart will be analyzed using time series analysis.

The time series plot in Figure 3 shows that the mean time series is not constant, which shows non stationary symptoms. Then the first or next sequence of diffusion is carried out until a stationary time series is obtained as shown in Figure 4 below:
The time series data above shows the treatment of seasonal influences, so that the time constraints of the day from 22.00 to 06.00 can be ignored. In figure 5 the autocorrelation function (ACF) shows a clear seasonal pattern of period 5.

![Graph of autocorrelation functions.](image)

**Figure 5.** Graph of autocorrelation functions.

From the graph of the autocorrelation function in Figure 6, it appears that the graph continues to decrease after lag 15. While the partial autocorrelation function is shown in the following figure:

![Partial autocorrelation function.](image)

**Figure 6.** Partial autocorrelation function.

From Figure 6 it appears that the partial autocorrelation function (PACF) will continue to decrease with or without changing mark after lag 15. We can notice that lag 15 is half of the period of each week 30 times, which gives a further impression of the effect of daily work. The detection of this outlier gives results for c = 3.0 and IO at the 91st observation and an A) at 37.41, 92. With one of the times series approaches, the outlier detection procedure in the time series shows that at the 91th and 92th time indices something special could happen on this particular day.
4. Conclusion

Control charts are used in statistical control processes to monitor the appearance of a process. However, the usefulness of a traditional control diagram is very dependent on the assumption of IID (Independent Identically Distributed). Many real-world applications face non-IID observations, which show the influence of trends, seasonality, and other influences. An appropriate way to overcome these effects by using time series analysis. Control diagram is in principle the same as plotting the area of acceptance of a hypothesis test, but if the process is not controlled, the control diagram cannot provide any guidance, so it is necessary to do a data analysis using time series analysis.

References

[1] Roberts H V and Tsay R S 1996 Making control charts more effective by time series analysis: three illustrative applications Commun. Stat. Methods 25 2767–96
[2] Liu R Y and Tang J 1996 Control charts for dependent and independent measurements based on bootstrap methods J. Am. Stat. Assoc. 91 1694–700
[3] Montgomery D C 2007 Introduction to statistical quality control (John Wiley & Sons)
[4] Berman S M 1964 Limit theorems for the maximum term in stationary sequences Ann. Math. Stat. 502–16