International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 7, Issue 5 Ver. I (May 2018), PP. 08-15

Statistical Process Control Methods for Detecting Outliers in Gps Time Series Data

Ha Trung Hieu[1] Tien Yin Chou[2], Yao Min Fang[2], Thanh Van Hoang[2]
[1]Ph.D program of Civil and Hydraulic Engineering, College of Construction and Development, Feng Chia University, Taiwan
[2]GIS Research Center, Feng Chia University, Taichung, Taiwan
Corresponding Author: Ha Trung Hieu

ABSTRACT: During system operation, for some reasons GPS time series data include some values that deviate from the rest of the dataset which is called *outliers*. To improve the quality of data in used for data analysis, it is very important to detect these outliers. Statistical Process Control methods are studied in this paper to solve this issue. The result shows that Shewhart control chart with the usage in pairs of chart can present both data trend and the dispersion of data. This control chart is suitable and highly feasible for detecting outliers in GPS time series data. In which, the mean – standard deviation ($\bar{x} - s$) chart is more effective to detect small value outliers in dataset, whereas individual – moving range (I-MR) chart and mean – range ($\bar{x} - R$) are suitable for detecting outliers with larger values. The result also shown that the EWMA control chart with chosen parameters is not suitable in GPS time series outliers detection.

KEY WORDS: Statistical process control, detect outlier, Shewhart control chart, GPS time series data

Date of Submission: 19-05-2018 Date of acceptance: 04-06-2018

I INTRODUCTION

Real Time Kinematic technique of Global Positioning System (GPS RTK) technology is increasingly used in different fields thanks to its high accuracy and effectiveness. In recent days, GPS RTK data is used to assess and evaluate the operation of a structure such as bridge, skyscraper or for disaster monitoring, etc. The applications of GPS in monitoring has many advantages in many ways. For instance, GPS monitoring for the earth's crust movement in the region can be considered as an informative method for evaluating its tension deformation conditions and can be used for seismic zoning and assessment of seismic risk [1]. In bridge monitoring, GPS is a common method used to assess and evaluate the different behaviors of structures under various load cases using real-time observations [2]. The article [3] presented that the advantage of GPS monitoring is to directly provide displacements without any further integration. GPS was also developed to continuously monitor slope stability in landslide hazards monitoring, or for seismological applications. Also, GPS is used in providing real-time information to user and manager in disaster event management, potential warning before the disaster, during disaster and after the disaster occurs. [4]. However, GPS measurement error from differential causes limit the application of this system. The main disadvantages of GPS monitoring systems for structural health monitoring are discussed in [5], [6].

During system operation, for some reasons GPS time series data include some values that deviate from the rest of dataset and it is called outlying observations, or *outliers*, that do not follow the statistical distribution of the bulk of the data, which leads to erroneous results in statistical analysis [7]. There are wide ranges of potential reasons which lead to outliers arise in dataset, because of internal operating system, external factors or even from weather conditions such as wind speed, temperature changes and so on [8]. Because of this, detect and remove the outliers in GPS time series data can improve the quality of data used for data analysis.

This paper focuses on detecting outliers in GPS time series data by using Statistical Process Control method.

II STATISTICAL PROCESS CONTROL METHOD

2.1. Statistical Process Control overview

Statistical methods play an important role in every aspect of science. In statistical field, Statistical Process Control (SPC) is a tool that measures and attains quality control. Generally, SPC is an analytical tool using control charts which allow us to see when a process is statistically “in-control” or “out-of-control”. A process is considered as “in statistical control” if the plotted data in control charts lies within control ranges, otherwise the “out of statistical control” situation happens when the plotted data in control charts fall outsides control limits [9].

The control charts can be separated into various categories based on different characteristics. According to Koutras (2007) et al., we may classify control charts into three types of chart: Firstly, Shewhart control chart which was first introduced in 1924 by W.A. Shewhart, secondly, Cumulative Sum (CUSUM) control chart based on sequential testing theory was developed by E.S. Page in 1954 and finally Exponentially Weighted Moving Average (EWMA) control charts which uses the weight putting into the recent observations were introduced in 1959 by S.W. Roberts [10], [11], [12]. Besides, various other charts including run charts, median charts, moving average, zone charts are also studied and used as part of the total quality management and SPC.

As mentioned above, this paper aims to detect outliers in GPS time series data by using some of SPC methods.

2.2. Shewhart control charts

In practical literature, there are two different phases of control chart: Phase I for testing historical data to specify whether they were sampled from an in-control or out-of-control process, while Phase II aim to identify the future data staying in-control or the out-of-control situation has occurred. Walter Andrew Shewhart (1891-1967) classified the variation of a process into two categories which called “assignable cause” and “chance cause” variation, or in other words, they are respectively called “special cause” and “common cause” variation of process [13].

A Shewhart control chart is a graphic illustration of the process measurements that have been observed versus the sample number or time line. The Shewhart chart contains the CL, UCL, LCL stands for center line, upper control limit and lower control limit, respectively.

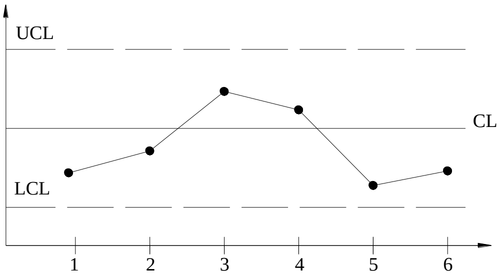

Figure 1. Shewhart control chart

These control chart components are popularly chosen as:

$$UCL = M(Y) + 3\times\sqrt{\mathrm{var}(Y)}, CL = M(Y), LCL = M(Y) - 3\times\sqrt{\mathrm{var}(Y)} \quad (1)$$

Whereas Y = f(x) is a statistical function of the process observations, in which *x* can be used to estimate the process mean, M(Y) is the mean value of Y and var(Y) is the variance of Y.

There are six basic types of Shewhart control charts, two for attribute data (p-chart and c-chart) and four for continuous data including individual (I) chart, moving range (MR) chart, mean ($\bar{x}$) chart, range (R) chart and standard deviations (s) chart. This paper contains the charts for continuous measure.

The I-chart presents the individual value x_1, x_2,…, x_n of observations, from the individual or k samples of data we may calculate the overall average:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \text{ or } \bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_k}{k} \quad (2)$$

The MR chart illustrates the difference between data point x_i and its predecessor x_{i-1}:

$$MR_i = |x_i - x_{i-1}| \text{ (3) and the mean is calculated as } \overline{MR} = \frac{\sum_{i=2}^{m} MR_i}{m-1} \quad (4)$$

The R chart demonstrates the difference between the smallest and the largest value in a sample. This range reflects the process variability instead of the tendency toward a mean value.

$R = x_{\max} - x_{\min}$ (5) and the mean of k samples is calculated as $\overline{R} = \frac{\sum_{i=1}^{k} R_i}{k}$ (6)

The s chart reflects the change of standard deviation in k observed samples, s is calculated by using the historical data (Phase I):

$s_i = \sqrt{\frac{\sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2}{n-1}}$ (7) and the mean is $\overline{s} = \frac{\sum_{i=1}^{i=k} s_i}{k}$ (8)

The most popular uses in control charts for continuous data are used in pairs to monitor both process location and dispersion: I-MR for sample size = 1, $\overline{x}$ - R for sample size from 1 to 10 and $\overline{x}$ -s for sample size ≥ 10. whereas the UCL, CL and LCL are calculated as follow [14]:

Table 1. Shewhart control limits

Control charts	**Control limits**	
	CL	**UCL and LCL**
MR	$\overline{MR}$	$3.267\,\overline{MR}\,;0$ (9)
I	$\overline{x}$	$\overline{x} + 2.66\overline{MR}\,;\ \overline{x} - 2.66\overline{MR}$ (10)
R	$\overline{R}$	$\overline{R} + D_4\overline{R}\,;\overline{R} + D_3\overline{R}$ (11)
$\overline{x}_R$	$\overline{\overline{x}}$	$\overline{\overline{x}} + A_2\overline{R};\,;\overline{\overline{x}} - A_2\overline{R}$ (12)
s	$\overline{s}$	$\overline{s} + B_4\overline{s}\,;\overline{s} + B_3\overline{s}$ (13)
$\overline{x}_s$	$\overline{\overline{x}}$	$\overline{\overline{x}} + A_3\overline{s};\ \overline{\overline{x}} - A_3\overline{s}$ (14)

Values of constants A_2, A_3, B_3, B_4, D_3, D_4 were developed specifically for determining the control limits for Shewhart control charts [14].

2.3. Exponentially Weighted Moving Average charts

Exponentially Weighted Moving Average (EWMA) control chart is a statistic for monitoring the process by the choice of factor λ to the prior observations. The EWMA control chart consists control limits UCL, LCL and based on the statistic:

$$Z_i = \lambda X_i + (1-\lambda)Z_{i-1};\ 0 < \lambda \le 1 \qquad (15)$$

Whereas Z_i is the EWMA values and Z_0 is set to be the target value, X_i is the sequentially observations can be the individual observations or sample averages. The average of preliminary data is sometimes used as the starting of EWMA, $Z_0 = \mu_0$. The parameter λ determines the rate at which previous observations are used to calculate the EWMA. The larger value of λ, the more weight to recent data and the less weight to past data. The process is considered as "out of control" situation when Z_i falls outside the control limits [17].
The upper and lower EWMA control limits are given as follow:

$$UCL = Z_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-\left(1-\lambda\right)^{2i}]}, LCL = Z_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-\left(1-\lambda\right)^{2i}]}; CL = Z_0 \qquad (16)$$

Where μ_0 and σ are respectively the sample mean and sample standard deviation of the process, estimated from preliminary data; The CL of EWMA chart is set as target value, $CL = Z_0$, sometimes calculated as the mean value of observations and L is suitable in control width limit. This chart reduces to Shewhart control chart for λ =1 and L = 3. The combinations of these two parameters are chosen by using an ARL (Average Run Length), which is the number of points will be used to plot in a control chart prior to an out-of-control situation is occurred. Lucas and Saccucci (1990) provided tables which contains optimal parameters to design an EWMA chart.

Table 2. Average run lengths of EWMA scheme (Lucas and Saccucci, 1990)

ARL$_0$ = 500							
Shift	**0.5**	**0.75**	**1**	**1.5**	**2**	**2.5**	**3**
λ = 0.75, L = 3.087	140	42.4	30.5	9.86	4.52	2.67	1.87
λ = 0.5, L = 3.071	88.4	35.7	17.3	6.44	3.58	2.47	1.91
λ = 0.1, L = 2.814	30.6	15.5	10.1	5.99	4.31	3.41	2.85
λ = 0.04, L = 2.477	28.0	16	11.2	7.03	5.18	4.14	3.48

The following procedure is recommended to design an EWMA control chart: Firstly, specify the in control ARL and the shift in the process. Secondly, choose the optimal parameters in the table given by Lucas and Saccucci (1990). Finally, evaluate the ARL for this EWMA to determine whether it provides sufficient protection against the other shifts [18].

III STATISTICAL PROCESS CONTROL METHOD IN DETECTING OUTLIER IN GPS TIME SERIES DATA

3.1. GPS RTK data from bridge health monitoring system

The data used in this paper is the Bridge Health Monitoring (BHM) data. It is collected from BHM system in Cantho bridge which is located in the South of Vietnam, it is the longest cable stayed bridge in South East Asia. The BHM system has been installed in 2010, which includes many sensors such as GPS sensors, accelerometers, temperature sensors, anemometers.

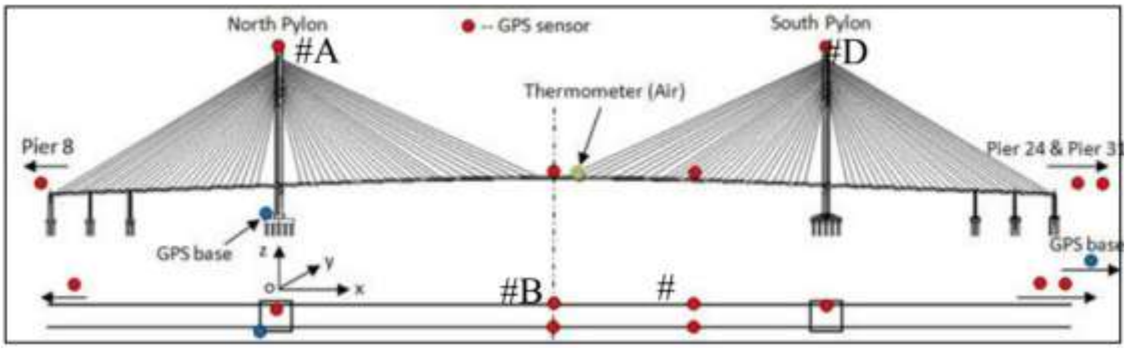

Figure 2. Arrangement of GPS sensors in Cantho bridge

The GPS system in Cantho bridge consists 9 sensors as rover stations and 2 base stations. In which, the GPS signal at each rover station was acquired in 20 Hz, and the acquired data are the 10-minutes-averaged values. The acquired data from each sensor, which includes three-direction coordinates that are x, y and z stand for longitudinal, lateral and vertical directions respectively. This paper calculates with x direction of the rover station locating at the center of the bridge deck, extract in 5 days from January 1st to January 5th, 2017 (720 observations).

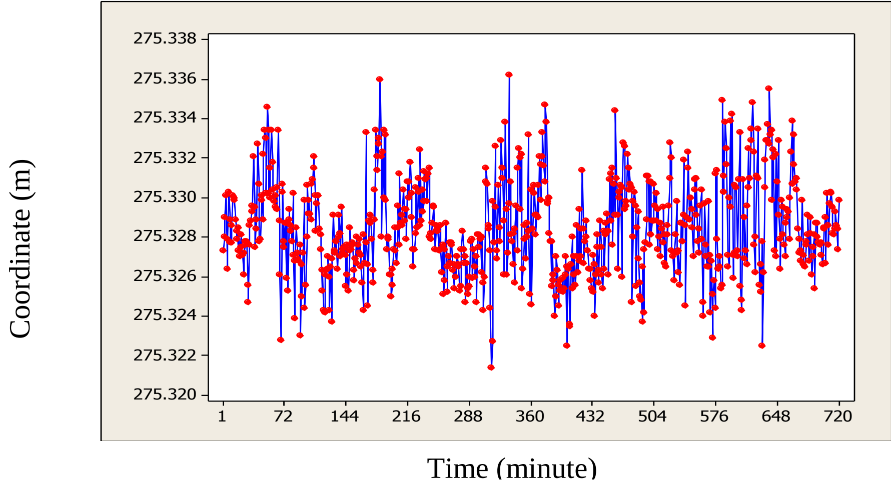

Figure 3. Acquired GPS data (x-direction)

This paper will use SPC methods to detect outliers in this acquired data.

3.2. Statistical Process Control in detect outlier in GPS time series data

Based on the theory of SPC method that mentioned in 2.2 and 2.3, calculate and detect outlier in GPS data showing in 3.1.

3.2.1. Shewhart control charts

For Shewhart control charts, we use in pairs: I-MR, $\overline{x}-R$ for sample size = 2 and $\overline{x}-s$ for sample size = 10. The mean and standard deviation are estimated by the observations. Values of constants are chosen by Wheeler and Chambers (1992) as A_2 =1.880, A_3 = 2.659, B_3 = 0.284, B_4 = 1.716, D_3 = 0, D_4 = 3.267. The related parameters of these control charts are calculated by using the equations from (2) to (14). The Shewhart control charts are schemed as follow:

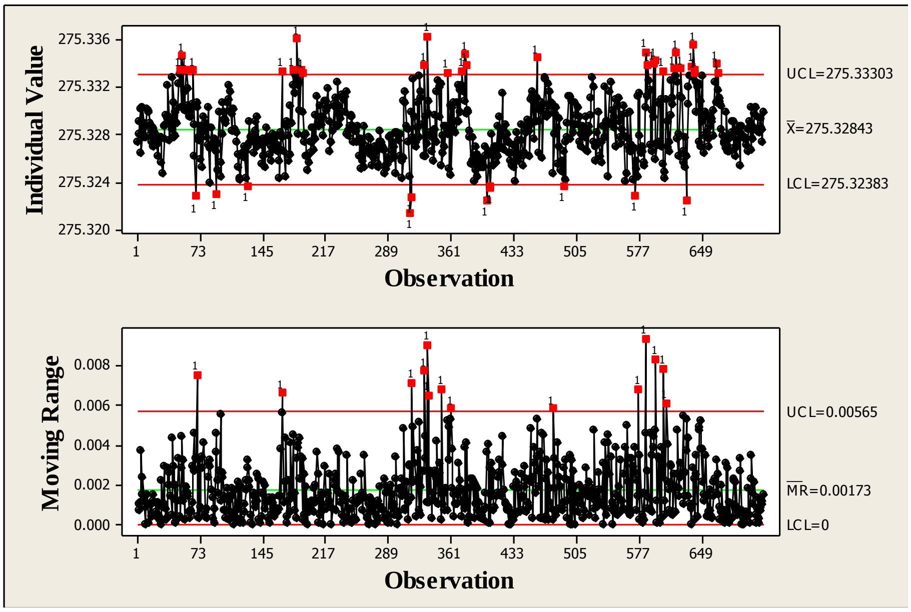

Figure 4. The I-MR of Shewhart control chart

As mentioned above, the I chart refer to the location and the MR chart reflects the dispersion of the data. The result shows that control charts can detect the outliers in GPS RTK data: 14 observations fall outside the control limits of MR chart and 42 observations fall outside control limits of I chart. The detected observations are anomalous from the bulk of the dataset. By combining these two of the charts, it can be seen that there are two main bulks of data which more dispersed than the others and it can be assumed that there some “special causes” in this time of monitoring.

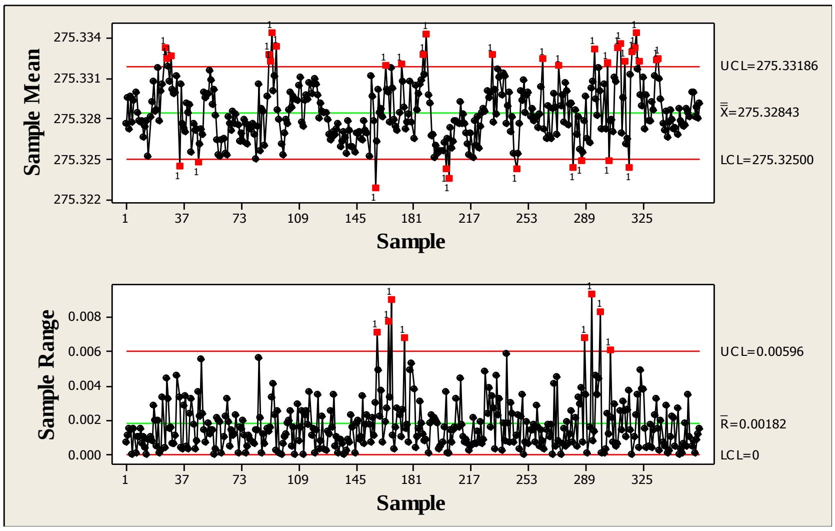

Figure 5. The $\overline{x}-R$ of Shewhart control chart

By grouping the data into subgroups equal to 2, there are 8 subgroups and 35 subgroups fall outside the control ranges of R chart and $\bar{x}$ chart, respectively. The result also shows that there are two bulks of data that are more dispersed than those in other subgroups. Besides, the control ranges of this control chart are narrower than those in I – MR control chart.

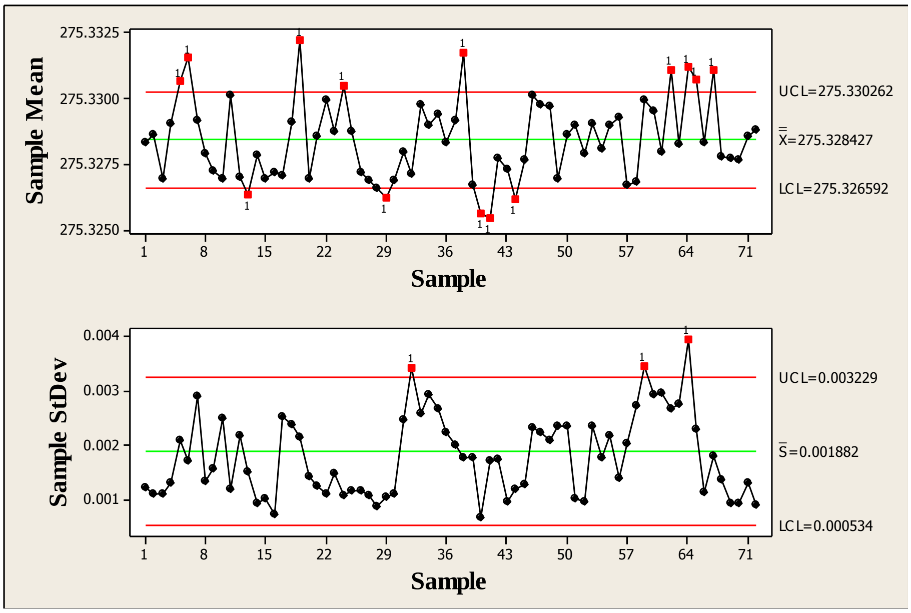

Figure 6. The $\bar{x}-s$ of Shewhart control chart

The result illustrates that the pair of control chart using sample mean and sample standard deviation ($\bar{x}-s$) to scheme with the subgroup equal to 10 can detect the outliers which are enormous in dataset: 3 subgroups of standard deviations and 14 subgroups of mean values lie outside the control limits. The control limits of this pair of chart are also smaller than those in I-MR chart and $\bar{x}-R$. Besides, the standard deviation values in each of the subgroups are calculated by determining the square of each data point distance to the mean of dataset, this means that this pair of chart are more effective to detect outliers with the smaller values.

3.2.2. EWMA control chart

Because the data used in this paper is the average of ten-minute monitoring data, we assume that the first 500 observations monitored in more than 3 days are in statistical control, by that the ARL$_0$ is chosen as 500. Lucas and Saccucci (1990) has found that a thumb rule is that the smaller values of λ, the smaller shifts of the process are detected. Based on the characteristic of GPS RTK data used in this paper, the ARL has to be "*long enough*" before the out-of-control situation occurs, the values of λ are chosen equal to 0.75, 0.5, 0.1 and 0.04 to scheme control charts and the shift of the process is chosen as 0.5 (the first column of Table 2).

The control limits of EWMA chart are calculated as equation (16). The EWMA control charts are schemed as follow:

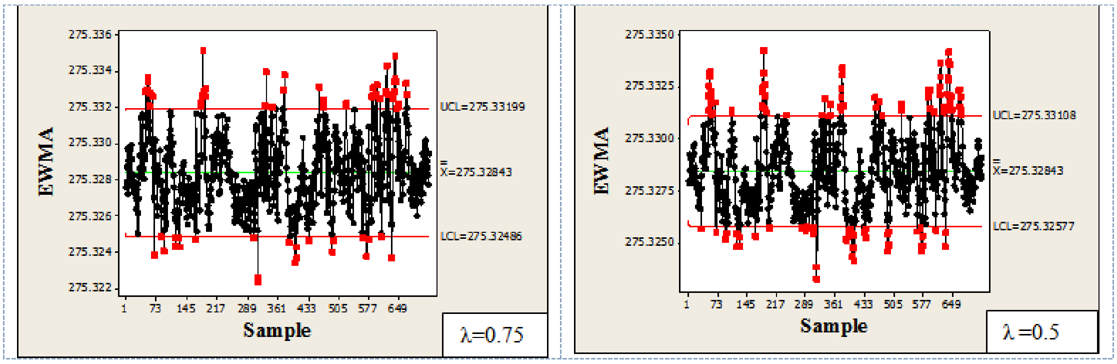

Figure 7. EWMA control chart, λ=0.75, 0.5

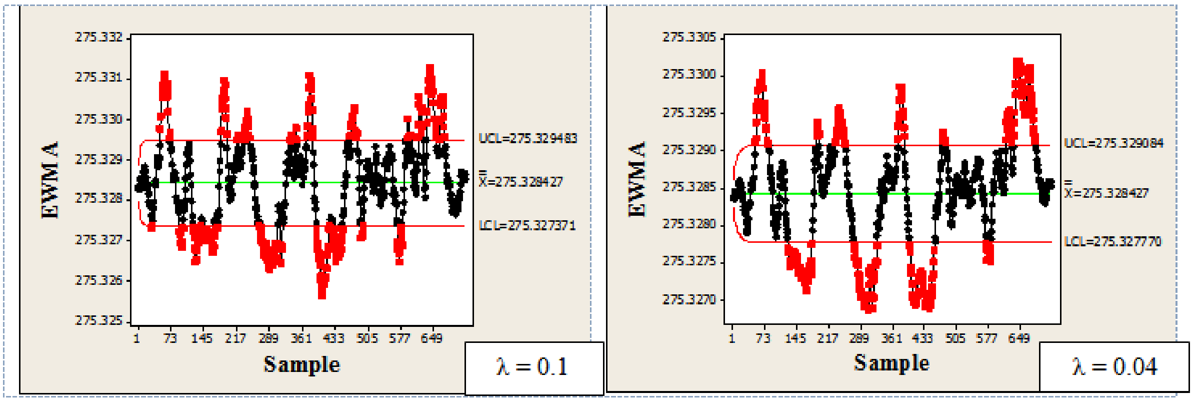

Figure 8. EWMA control chart λ=0.1, 0.04

The results show that: 74, 123, 297 and 312 are the numbers of outliers detected by EWMA control charts with respect to the values of λ, 0.75, 0.5, 0.1 and 0.04 respectively.

It can be seen that, if the values of λ are changed to the smaller ones, there are more points fall outside the control limits.

IV RESULTS AND DISCUSSION

- Using these pairs of Shewhart control chart to detect outliers in GPS time-series data, which are I-MR, $\bar{x}-R$ and $\bar{x}-s$, it can be assessed data trend (by plotting I, $\bar{x}$ chart) and the dispersion of data (by plotting MR, R, s chart) at the same time. It can be seen that the I, $\bar{x}$ chart are not only describes the data changes (upward, downward) but also can detect the outliers which are abnormal to the rest of the dataset, while the MR, R and s control chart can clearly reflect the differences between adjacent individual or subgroup observations.

- The results also show that the larger size of subgroups, the smaller control ranges of control charts. In comparison to the results of 3 kinds of Shewhart control charts, the mean – standard deviation control chart is more sensitive to small shifts in the process.

- By choosing the smaller values of λ to plot EWMA control chart (the more weights are put into past data), the result shows that the UCL and LCL calculated in Equation (16) are also getting smaller, this means that more and more plotted GPS time series data lies outside control ranges. As a result, some of the normal data are also considered as *outliers*. Otherwise, if the values of λ are chosen to be larger, the less weights are put into past data. Meanwhile, the GPS time series data used in this paper are the ten-minute-averaged observations, the more past data are used, the more working conditions of monitored structures are reflected. By this, the EWMA control chart with chosen parameters is not suitable for detecting outliers in GPS time series data.

V CONCLUSION

This paper studies the application of Statistical Process Control method in detection outliers in GPS time series data. Specifically, this paper focuses on Shewhart control chart and Exponentially Weight Moving Average control chart to detect abnormal data. The results can be summarized as below:

- Shewhart control chart with the usage in pairs of chart can present both data trend and the dispersion of the data. This control chart is suitable and highly feasible for detecting outliers in GPS time series data. In which, the mean – standard deviation ($\bar{x}-s$) chart is more effective to detect small value outliers in dataset, whereas individual – moving range (I-MR) chart and mean – range ($\bar{x}-R$) are suitable for detecting outliers with larger values.
- It is also shown that with the chosen parameters, the EWMA control chart is not suitable to detect outliers in GPS time series data.

REFFERENCES

[1]. A.V. Vilayev, Zh.Sh. Zhantayev, A.Zh. Bibosinov (2017), *Monitoring crustal movements in northern Tianshan Mountain based on GPS technology*, JSC "National Center of Space Research and Technology", DTOO "Institute of Ionosphere", Almaty, 050020, Kazakhstan;

[2]. Mosbeh R. Kaloop, Jong Wan Hu and Emad Elbeltagi (2016*), Adjustment and Assessment of the Measurements of Low and High Sampling Frequencies of GPS Real-Time Monitoring of Structural Movement*, ISPRS International Journal of Geo-Information (ISSN 2220-9964);3

[3]. [G. Esteban Vazquez B., J. Ramon Gaxiola-Camacho, Rick Bennett, G. Michel Guzman-Acevedo, Ivan E. Gaxiola-Camacho (2017), *Structural evaluation of dynamic and semi-static displacements of the Juarez Bridge using GPS technology*, Measurement Journal, Volume 110, Pages 146-153, Publisher Elsevier;

[4]. Kamil Muhammad Kafi, Mohamed Barakat A. Gibril (2016), *GPS Application in Disaster Management: A Review*, Asian Journal of Applied Sciences (ISSN: 2321 – 0893) Volume 04 – Issue 01;

[5]. Seok Been Im, Stefan Hurlebaus, M.ASCE, and Young Jong Kang, M.ASCE (2013), *Summary Review of GPS Technology for Structural Health Monitoring*, JOURNAL OF STRUCTURAL ENGINEERING, American Society of Civil Engineers;

[6]. Ting-Hua Yi, Hong-Nan Li and Ming Gu (2013), *Recent research and applications of GPS-based monitoring technology for high-rise structures*, Struct. Control Health Monit., Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/stc.1501;

[7]. Hancong Liu Sirish Shah, Wei Jiang (2004), *On-line outlier detection and data cleaning*, Computers and Chemical Engineering 28 (2004) 1635–1647;

[8]. VICTORIA J. HODGE & JIM AUSTIN, (2004), *A Survey of Outlier Detection Methodologies*, Artificial Intelligence Review 22: 85–126, 2004 Kluwer Academic Publishers. Printed in the Netherlands;

[9]. John Oakland (2007), *Statistical Process Control*, sixth edition, publisher Routledge, United Kingdom;

[10]. E. S. Page (1954), *Continuous Inspection Schemes,* Biometrika, vol. 41, no. 1, No. 1/2,Jun. 1954;

[11]. E. S. Page (1954), *An improvement to Wald's approximation for some properties of sequential tests*, Journal of Royal Statistical Society - Series B, vol. 16, no. 1;

[12]. M. V. Koutras & S. Bersimis & P. E. Maravelakis (2007), *Statistical Process Control using Shewhart Control Charts with Supplementary Runs Rules*, Methodology and Computing in Applied Probability, Volume 9, Issue 2, pp 207–224;

[13]. M Best, D. Neuhauser (2006), *Walter A Shewhart, 1924, and the Hawthorne factory*, Qual Saf Health Care 2006;15:142–143;

[14]. D. J. Wheeler and D. S. Chambers. (1992). *Understanding Statistical Process Control*, Second Edition, SPC Press, Inc;

[15]. Vera do Carmo C. de Vargas, Luis Felipe Dias Lopes, Adriano Mendonc ¸a Souza (2004), *Comparative study of the performance of the CuSum and EWMA control chart,* Computers & Industrial Engineering 46 (2004) 707–724;

[16]. Philippe Castagliola, Petros E. Maravelakis (2010), *A CUSUM control chart for monitoring the variance when parameters are estimated*, Journal of Statistical Planning and Inference;

[17]. J.M. Lucas, M.S. Saccucci (1990), *Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements*, Technometrics, vol.32 No.1;

[18]. A. K. Patel, J. Divecha (2011), *Modified exponentially weighted moving average (EWMA) control chart for an analytical process data*, Journal of Chemical Engineering and Materials Science Vol. 2(1), pp. 12-20;

[19]. D. J. Wheeler and D. S. Chambers. (1992). *Understanding Statistical Process Control, Second Editio*n, SPC Press, Inc.

Ha Trung Hieu." Statistical Process Control Methods for Detecting Outliers in Gps Time Series Data" International Refereed Journal of Engineering and Science (IRJES), vol. 07, no. 05, 2018, pp. 08–15.