A Study on Outlier Detection Method for GPS Measurement in Bridge Health Monitoring System

Hieu Ha Trung 1*, Tien Yin Chou 2 and Yao Min Fang 2
1 Ph.D. Program for Civil Engineering, Water Resources Engineering and Infrastructure Planning, College of Construction and Development, Feng Chia University, Taiwan.
2 GIS Research Center, Feng Chia University, Taichung, Taiwan.
*Corresponding author email id: hstrunghieu0409@gmail.com

Abstract – Outlier detection, also referred as anomaly detection, always be considered as a primary step in data analysis. This paper provides a synthesis of some statistical-based methods and their potential applications for detecting outlier in bridge monitoring time series data. The results indicate the combination of InterQuartile Range and Mahalanobis Distance is the most appropriate and effective method to identify abnormal points in GPS time series in terms of three-dimensional coordinates data from bridge health monitoring systems.

Keywords – Outlier Detection, Mahalanobis Distance, InterQuartile Range, Bridge Health Monitoring System, GPS Measurement.

I. INTRODUCTION

Outlier detection, also referred as anomaly detection, always be considered as a primary step in data analysis. A well-known definition of an outlier is given by D.M. Hawkins [1]: “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”. The presence of outliers may lead to an incorrect outcome or error result in data processing, but it also possibly contains some useful information [2, 3]. Outlier identify therefore becomes a crucial step for data preprocessing.

The outlier detection has been widely studied in the last decades in different fields with the wide applicability [1, 2, 4, 5]. An example is a novel outlier detection method involving an iterative random sampling procedure is introduced by [6], in which a new measured called observability factor was implemented in their proposed detection method. In other fields, such as a virtual graph is utilized to label potential outliers within the medical dataset [7], outlier detection in stereo matching [8], anomaly detection in industrial processes [9], outlier detection scheme of molten steel temperature in ladle furnace [10], anomaly detection in dynamic process runtime behavior [11], anomalous entities detection in pedestrian flows [12]. These extensive applications have led to the development of numerous outlier detection techniques.

Typically, most outlier detection methodology use some statistical tool based on historical data. One of the most common and straightforward methods for such issue is using the mean (μ) and standard deviation (σ) of the data set, in which if the data points fall outside (μ ± kσ) range, they are labeled as potential outliers. Some researchers have proposed the choice of k criteria depends on different situations [13-15]. Another type of mean – standard deviation based method is Z-scores, \( Z_i = \frac{x_i - \mu}{\sigma} \), whereas an observation would be identified as outlier if it exceeds 3 in absolute value. Nevertheless, the data must obey normal distribution and this method is susceptible to extreme values [16]. Instead of using Z-scores, Iglewicz and Hoaglin [17] recommend the modified Z-scores, \( M_i = \frac{0.6745(x_i - \bar{x})}{MAD} \), in which the absolute value of \( M_i \) greater than 3.5 be marked as outliers,
where $\tilde{x}$ is the Median and MAD denotes the Median Absolute Deviation. Also in [16], the authors show a didactical and small sample to illustrate the computing sequence of using MAD for detecting outliers and recommend the range $\tilde{x} \pm 2.5 \text{MAD}$ as the range for inlier. Another method is InterQuartile Range (IQR), or called Box plot, chooses the first and the third quartile and the difference between these quartiles to define outliers. This method does not require the distribution assumption and also will not be affected by extreme abnormal values [3, 18]. Mahalanobis Distance (MD) is also chosen as an effective tool to identify outlier, in which it computes the distance from a point $x$ to the mean $\mu$ and an observation is considered as outlier if MD exceeds the coefficient $c_k$ depending on the dimension of sample [19, 20].

Bridge Health Monitoring (BHM) systems are used to monitor the physical status of bridge structural elements, structure integrity and usually consists the sensors system which provide the necessary data for processing and the useful information for managers, users in bridge operation management [21]. The data obtains from BHM systems is mostly time series data such as coordinate data from GPS sensors, temperature data from thermometer, wind speed from anemometer, acceleration data from accelerometer. Some previous studies have used data filters to remove noises in measured signals such as in [22] and [23] the moving average filter and Chebyshev filter are employed to smooth the GPS measurements. Wavelet transform is also highly recommended as a tool for eliminating noises to get useful signals in GPS data [24, 25]. In bridge performance assessment, bridge displacement and deformation is significantly indispensable and GPS data is the common choice these assessments [26]. However, GPS measurements obtained from BHM systems may not accurately reveal the behavior of the bridge without detecting outlier prior to data analysis [23].

This study aims to figure out the most appropriate methods for detecting potential outliers in GPS measurements in terms of coordinates using short-period of monitoring time obtaining from BHM system in Can Tho bridge in Vietnam, which is a long span cable-stayed bridge. The combination of IQR method and Mahalanobis Distance method are utilized to identify abnormal observations in such data set. Further details will be discussed in next sections.

II. STRUCTURAL HEALTH MONITORING SYSTEM OF CAN THO BRIDGE AND DATA DESCRIPTION

Can Tho bridge links the two provinces in the South of Vietnam (Can Tho and Vinh Long). The bridge was launched construction in 2004 and opened for traffic in April, 2010 marking it as the longest cable-stayed bridge in the Southeast Asian region, which is 2.75 kilometers long.

Fig. 1. Sensors arrangement in Can Tho Bridge.
The BHM system has been installed in 2010, which comprises many sensors such as GPS sensors, accelerometers, thermometer sensors, anemometers, strain gauge, rain gauge as showing in Figure 1. The GPS system in Can Tho 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. A one-day period of data (January 1st 2015) is chosen for the analysis. Statistical description of the data is shown in Table 1.

Table 1. Data description.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
<th>Shapiro-Wilk Statistic</th>
<th>df</th>
<th>Sig.</th>
<th>Normal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>144</td>
<td>275.3138</td>
<td>0.0032</td>
<td>275.3061</td>
<td>275.3236</td>
<td>.985</td>
<td>144</td>
<td>.106</td>
<td>✓</td>
</tr>
<tr>
<td>y</td>
<td>144</td>
<td>12.28835</td>
<td>0.0051</td>
<td>12.2751</td>
<td>12.3088</td>
<td>.923</td>
<td>144</td>
<td>.000</td>
<td>✓</td>
</tr>
<tr>
<td>z</td>
<td>144</td>
<td>42.62502</td>
<td>0.0373</td>
<td>42.5012</td>
<td>42.7022</td>
<td>.903</td>
<td>144</td>
<td>.000</td>
<td>✓</td>
</tr>
</tbody>
</table>

The normality of the data is checked by Shapiro-Wilk test, which is highly recommended by [27, 28]. In this paper, we use Shapiro-Wilk test in SPSS software to check the normal distribution of data. Therein, if the Significant value of the Shapiro-Wilk test is greater than 0.05, the data is normal. Otherwise, the data significantly deviates from a normal distribution. This result from Shapiro-Wilk test indicates the normal distribution of GPS time series data in x direction.

For evaluating the effect of outlier detection methods, the comparison is made between the numbers of detected outliers by these methods and the real outliers, which is defined by frequency histogram whereby the furthest points in the figure are outliers as following:

The histogram of x-direction data in Figure 2 shows the furthest points for 65, 72, 73 and 79.

The histogram of y-direction data in Figure 3 shows the furthest points for 65, 104

The histogram of z-direction data in Figure 4 shows the furthest points for 10, 11, 81 and 138.
From the frequency histograms, the points that are considered as outliers in GPS data set are 10, 11, 65, 72, 73, 79, 81, 104 and 138 (real outliers).

In the next section, some methods will be employed to label outliers in such data set then be numerically compared to these above real outliers.

III. OUTLIER DETECTION METHODS FOR GPS TIME SERIES DATA IN CAN THO BHM SYSTEM

As mentioned in previous section, some methods such as mean-standard deviation method or Z-scores are potentially applicable to identify outliers in normally distributed data only, so these method is invalid when it comes to the non-normal distribution data. In our case, it can only be applied for x-direction data. For median-median absolute deviation based method, which is proposed by [16], the multiplication by b is crucial in MAD calculation, where b is a constant and the choice of this constant value also depends on the underlying distribution of data. Those methods, as a result, possibly lead to a more complicated sequence of calculation for
detecting outlier in GPS time series data. Thus, the InterQuartile Range and Mahalanobis Distance methods are employed to single out the outliers in our data set and then be evaluated their effectiveness as following.

**The InterQuartile Range**

The InterQuartile Range (IQR) presents the difference between the first and third dataset quartiles: \( IQR = Q_3 - Q_1 \), where 25\% of the data values are less than \( Q_1 \) and 25\% of the data values are greater than \( Q_3 \). An outlier could be defined if any points are outside the range: \( Q_1 - kIQR \leq x \leq Q_3 + kIQR \)

John Tukey [29] suggests \( k = 1.5 \). The range of \( Q_1 - 1.5IQR \leq x \leq Q_3 + 1.5IQR \) is therefore referred as Tukey fences.

![IQR plots for each direction data.](image)

The IQR method is separately used to single out outliers in each direction data. The result shows that only the point 65 (in y-direction data set) is labeled as outliers, the other points 65, 71, 72 (in x-direction data), 55, 56, 57, 60, 64, 68 (in y-direction data) are marked as potential outliers while there is no outlier in z-direction data. This method is obviously ineffective for a single-direction data in GPS measurements, especially to the data with large standard deviation (\( \sigma = 0.0373 \) in z-direction data).

**Mahalanobis Distance**

The Mahalanobis Distance (MD) is calculated as [30]:

\[
MD = \sqrt{(x - \mu)^T V^{-1} (x - \mu)}
\]

Where:
- \( x \) is a vector of variables,
- \( \mu \) is the vector of means of each variable,
- \( V \) is the covariance matrix.

![Mahalanobis Distance data.](image)
The vector of variables is 
\[
\begin{pmatrix}
    x_1 & y_1 & z_1 \\
    x_2 & y_2 & z_2 \\
    \vdots & \vdots & \vdots \\
    x_n & y_n & z_n
\end{pmatrix}
\]
and 
\[
\mu = \left( \frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}, \frac{\sum_{i=1}^{n} z_i}{n} \right)
\]
and the covariance matrix 
\[
V = \begin{pmatrix}
\text{var}(x) & \text{cov}(x,y) & \text{cov}(x,z) \\
\text{cov}(y,x) & \text{var}(y) & \text{cov}(y,z) \\
\text{cov}(z,x) & \text{cov}(z,y) & \text{var}(z)
\end{pmatrix}
\]

The Mahalanobis Distance is computed for GPS data set and be presented in Figure 6. It is virtually seen that there are some points deviating from the bulk of data set, which are potentially outliers. In [20], an observation \( x_i \) will be labeled as an outlier if \( MD > c_k \), where \( c_k \) is a coefficient depending on the number of observations. Accordingly, this criterion may lead to the masking effect. Thus, we propose a more simple method by using IQR method for Mahalanobis Distance dataset.

The Q₁, Q₃ and IQR using Tukey fences are also computed and then graphically shown in Figure 7. IQR method for Mahalanobis Distance data indicates the points 10, 11, 65, 72, 73, 79, 81, 104, 138 are outliers and the points 55, 62, 63, 60, 63 are potential outliers.

In compared to real outliers by frequency histogram in previous section, the combination of using IQR and MD shows an effective result to detect outlier in GPS measurements: all the outliers in data set are correctly detected. Additionally, IQR plots also points out some other abnormal values, which are labeled as potential outlier for further analysis.

Fig. 7. IQR plot for MD data set.

IV. CONCLUSION

Based on the empirical result, we highly recommend using the combination of InterQuartile Range and Mahalanobis Distance method to point out outliers in GPS time series data from BHM system, the computing sequence is suggested as follow:

- From 3-direction coordinates, the Mahalanobis Distance of each observation be computed;
- The Q₁, Q₃ are calculated for Mahalanobis Distance data set and then using IQR with Tukey fences to identify abnormal points. These abnormal points are outliers in GPS measurements.
With respect to outlier detection topic, our proposed method is simply based on the combination of some methods from previous studies but it provides an appropriate and effective method as well as simple calculation for detecting outliers in GPS time series data obtaining from Bridge Health Monitoring system.

REFERENCES


AUTHOR’S PROFILE

First Author
Hieu Ha Trung, PhD student at Ph. D. Program for Civil Engineering, Water Resources Engineering, and Infrastructure Planning, College of Construction and Development, Feng Chia University, Taiwan. Research interests: Civil Engineering, bridge monitoring, Geoinformatics.

Second Author
Tien Yin Chou, GIS Research Center, Feng Chia University, Taichung, Taiwan.

Third Author
Yao Min Fang, GIS Research Center, Feng Chia University, Taichung, Taiwan.