Research Journal of Recent Sciences _____ ISSN 2277-2502
Vol. **4(3),** 150-153, March **(2015)** *Res.J.Recent Sci.*

Detection of Outlier in Flood Observations: A Case Study of Tamer Watershed

Heidarpour B.[1], Panjalizadeh Marseh B.[1], Ekramirad A.[2], Hosseinnezhad A.[1] and Ghasemian Langroudi A.[2]
[1]Department of Civil Engineering, Young Researcher Club ,Roudsar and Amlash Branch, Islamic Azad University, Roudsar, IRAN
[2]Department of Civil Engineering, Langroud Branch, Islamic Azad University, Langroud, IRAN

Available online at: www.isca.in, www.isca.me
Received 18th December 2013, revised 29th May 2013, accepted 28th October 2014

Abstract

In this research for determining outlier data by traditional methods, has used [Mean ± 3S], Box plot, Grubbs' Test and Grubbs and Beck Test. Based on the traditional methods and the box plot (boxplot), discharge of 783 m^3/s was determined in 2004-2005 water year. as the outlier data in the data series of Annual maximum instantaneous peak streamflow in Tamer hydrometric station. The station is located in the outlet of Tamer watershed in Iran in an area about 1,531 square kilometers southern Caspian Sea coastline. Moreover, the outlier data in Grubbs' Test was achieved in two modes – observational data series and normalized data series. Based on this, discharge of 783 m^3/s in observational data series was determined as the outlier data. After normalizing the data using Box-Cox transformation, no outlier data were achieved using Grubbs' Test. By using Grubbs-Beck relation, no outlier data were obtained in Tamer hydrometric station.

Keywords: Outlier data, Tamer watershed, statistical tests, Grubbs' Test, Grubbs-Beck Test, Box-Cox transformation.

Introduction

Statistics and information of the recorded maximum Floods in a dam construction site play a decisive role in design flood estimation. Meanwhile, before making any form of calculations, we should be confident about the accuracy of information, we should closely determine weight and value of each recorded quantity - as real dimensions within the desired time span - and specify its position as much as possible. However, unfortunately, value and position of registration statistics is forgotten in some cases, all pieces of information are given an equal value and floods with different return periods are calculated using common techniques. As a result, the obtained figures (design floods) has no consistency with the case study watershed and the costs born to construct massive concrete structures for the floods can be resembled in the fortune premium that should be paid for the fictitious and imaginary accidents.

Outlier data are the data that are single data points that appear to depart significantly from the trend of the other data flow. They are usually divided into three groups: i. Observations made by collection error and/or data registration ii. Observations made by natural factors iii. Observations made by unnatural factors such as dam failure[1]. Both high outlier floods and historical floods are considered as exceptionally large floods, the former were observed during the period of systematic registration and the latter were observed out of this period. The systematic record can be used directly in flood frequency analysis. The non-systematic record cannot be used unless additional information can be supplied to relate it to the population of all flood peaks[2]. The present research aims to compare different methods for Detection of Outlier in Flood Observations.

Methodology

Case Study Region: The area of Tamer watershed is about 1531 square kilometer located in southern Caspian Sea coastline in Iran. The watershed is of the main subwatersheds of Golestan watershed. This area is located between 56°.4' and 55°.30' E longitude and 37°.48' and 37°.24' N latitude. Figure 1 shows position of the subwatersheds and drainage network of this area.

Determining Outlier Data: As proposed by the American Water Resources Council in 1981, if data skewness coefficient exceeds 0.4, outlier data test should be carried out for high values. When data skewness is lower than -0.4, the test is conducted for low values. In case skewness coefficient is between 0.4 and -0.4, the test is conducted for high and low data. On peak floods which are considered as the outlier data - necessary studies should be carried out first to see if the errors caused by preliminary calculations on statistics sheets and transfer of data to different sheets and/or computer were avoided. They should then be compared with the historical data and/or data of the neighboring watersheds. If the available data show that an outlier data can be accepted as maximum data in a long time, it can be considered as historical data.

Many methods have been offered for outlier data detection, but no suitable and comprehensive method has been introduced for this purpose so far[3]. To be more confident and to conduct further studies, we may use several methods for outlier data detection and compare their results.

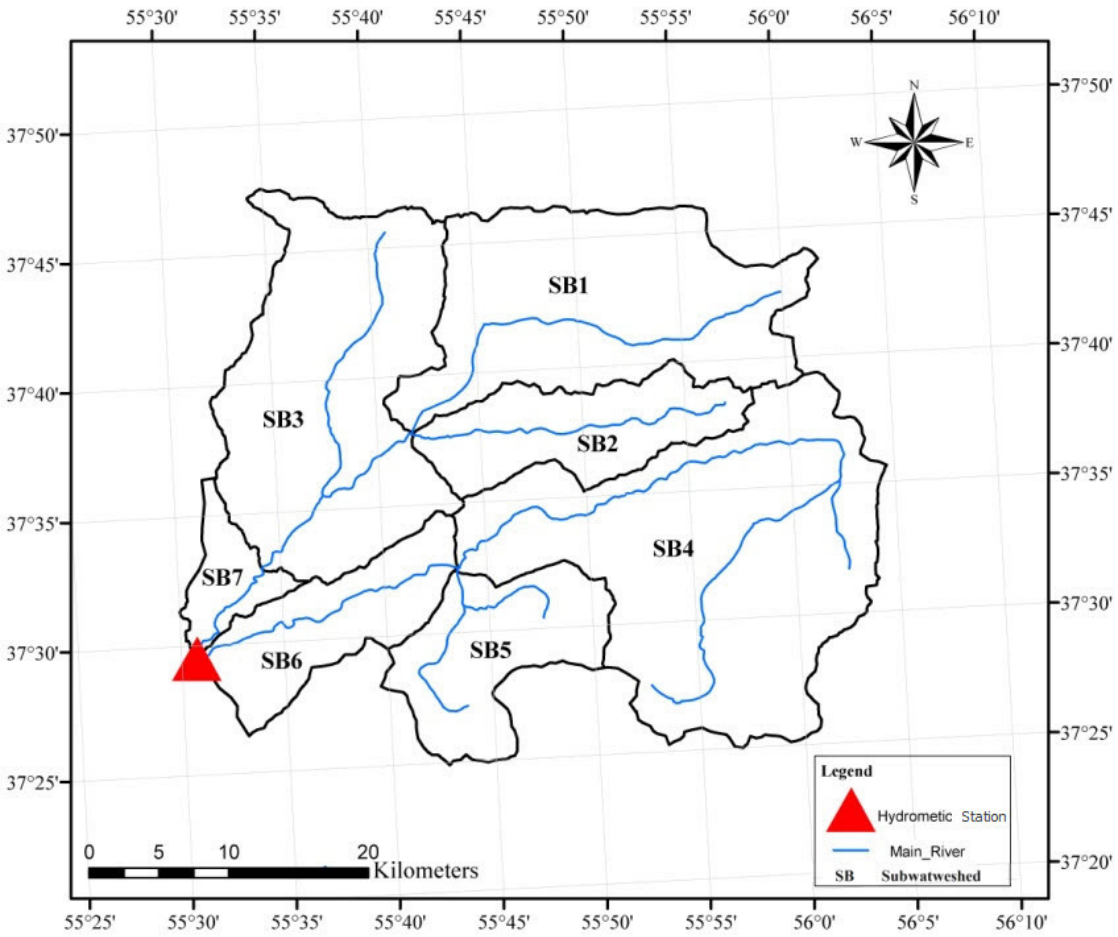

Figure-1
Position of the subwatersheds and hydrometric station in the case study region

Methods for outlier data detection can be divided into 2 groups – Interval and statistical tests. In Interval methods, distribution of reviewed observations and the data out of a specific Interval are considered as outlier data. The major issue in this concern is relationship of the determination of the Interval to specify outlier data. The traditional method in this concern is the mean ± 3 times standard deviation ($\bar{X} \pm 3S$). The data larger than the mean plus 3 times of the standard deviation and the data smaller than the mean subtracted by 3 times of the standard deviation are considered as outlier data[4].

Box plot is also considered as the Interval methods. This plot method is to show position, dispersion and skewness of data, which is frequently used to detect outlier data. Based on this, the observations within $X_i<Q_1+1.5IQR$ or $X_i>Q_1+1.5IQR$ inequalities are among the weak outlier data and the observations within $X_i<Q_1+3IQR$ or $X_i>Q_1+3IQR$ equalities are among the strong outlier data. In these relations, QI, Q3 and IQR are the first quartile, the third quartile, and interquartile range (IQR=Q3-Q1), respectively[3].

Grubbs' Test is of the tests used for detection of outlier data. In this test, it is assumed that data follow a normal distribution and outlier data are identified in each step. If an outlier is detected, the relevant outlier is deleted and the test is conducted again for the remaining data so that no outlier data remain. Grubb's Test statistic (G) is calculated by the following relation:

$$G=\frac{\max\left|X_i-\bar{X}\right|}{S} \quad (1)$$

Where: X_i is the lowest or the highest data, $\bar{X}$ is mean of data and S is standard deviation of data.
For Grubb's Test the hypothesis of no-outlier is rejected if

$$G>\frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2_{\left(\frac{\alpha}{2n},n-n\right)}}{n-2+t^2_{\left(\frac{\alpha}{2n},n-n\right)}}} \quad (2)$$

With $t_{\left(\frac{\alpha}{2n},n-n\right)}$ denoting the $\frac{\alpha}{2n}$ percentile of a t-distribution with (n-2) degrees of freedom[3].

The Grubbs and Beck test (G-B) may be used to detect outliers. In this test the quantities X_H and X_L are calculated by using equation-3 and 4,

$$X_H = \exp(\bar{X} + K_N . S) \quad (3)$$

$$X_L = \exp(\bar{X} - K_N . S) \quad (4)$$

$$K_N = -3.62201 + 6.28446N^{\frac{1}{4}} - 2.49835N^{\frac{1}{2}} + 0.491436N^{\frac{3}{4}} - 0.037911N \quad (5)$$

Where $\bar{X}$ and S are the mean and standard deviation of the natural logarithms of the sample, respectively, and K_N is the G-B statistic tabulated for various sample sizes and significance levels by Grubbs and Beck. At the 10% significance level, the following approximation proposed by Pilon et al. is used, where N is the sample size. Sample values greater than X_H are considered to be high outliers, while those less than X_L are considered to be low outliers[5].

Results and Discussion

The present research used Annual maximum instantaneous peak floods in Tamer hydrometric station in the case study watershed outlet. This area is located at 59˚29'30" E longitude, 37˚28'30" N latitude, and 132 meters above sea level. Figure 2 shows changes of Annual maximum instantaneous peak floods in Tamer hydrometric station during the years the present case study was conducting.

Based on the Interval traditional methods and box plot, an outlier was obtained in Tamer hydrometric station. In the traditional method of Interval and box plot, the maximum outlier values were 479.3 and 444.9, respectively (maximum value of the outlier data is strong). Based on this, discharge of 783 m^3/s was determined as outlier data in 2004- 2005 water year.

Grubbs' Test was performed in two modes - Observational data series and normalized data series - to determine outlier data. Based on this, discharge of 783 m^3/s in the Observational series of data was determined as the outlier data. As it is assumed in the Grubbs's test that data follow a normal distribution, ($Y=X^{\lambda}$) transformation was used to normalize data[6]. Figure 3 shows the probability plot for the states before and after Box-Cox transformation for λ=-0.08. After normalization of data, no outlier data was obtained from Grubb's Test. Statistic of Grubbs' Test in Observational series and the series transformed to normal series were 5.32 and 1.979, respectively. The critical values of the test for α= 0.01 and 0.05 were calculated as 3.04 and 3.38, respectively. By applying Grubb- Beck relation, K_N and X_H values were calculated as 2.682 and 1507 m^3/s. Based on the test, no outlier data were seen in Tamer hydrometric station.

According to the results of the outlier value determination tests and considerable high differences between the values of maximum flood peak observed (783 m^3/s in 2004-2005 water year) and the following maximum flood (230 m^3/s in 1997-1998 water year) whose ratio is about 3.4, maximum flood peak in the station was considered as the outlier data.

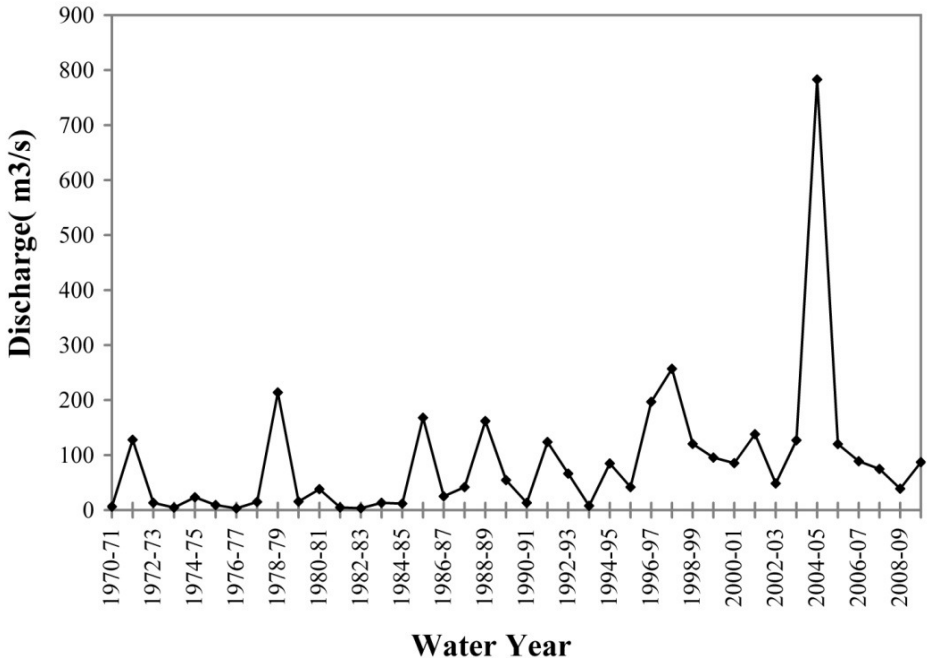

Figure-2

Changes of Annual maximum instantaneous peak floods in Tamer hydrometric station during the water years

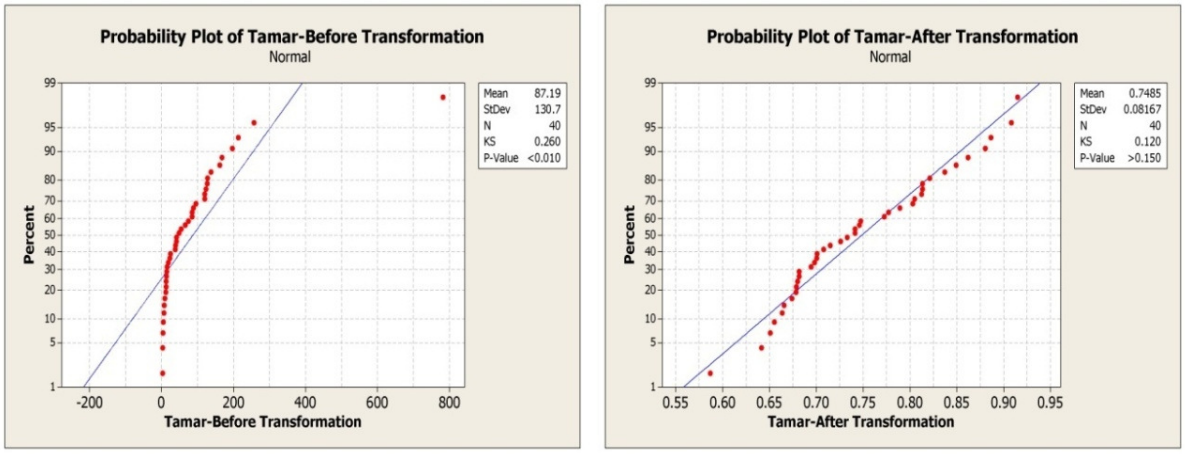

Figure-3
Probability plot before and after Box-Cox transformation

Table-2
Summary of the results of outlier test determination in different methods

No	Name of Method	Parameter Limit	Parameter Value
1	Traditional Interval	$\bar{X}$+3S	479.3082
2	Box Plot	Q_1+3IQR	444.85
3	Grubbs' Test (Statistic G)	Observational Data Series	5.3235
		Normalized Data Series	1.9793
4	Grubbs-Beck Test	K_N	2.682
		X_H	1507

Conclusion

Based on the traditional methods and box plot methods, an outlier was obtained in Tamer hydrometric station in outlet of case study watershed. The upper limit of the outlier value in the traditional method of Interval and box plot was 479.3 and 444.9 (upper value of outlier). Based on this, in 2004-2005 water year, maximum instantaneous observational discharge of 783 m^3/s was determined as an outlier. Determination of outlier data in Grubbs' Test was performed in two modes of observational data series and normalized data series. As it is assumed in the Grubbs's test that data follow a normal distribution, Box-Cox transformation was used for normalization of the data. After normalization of the data, no outlier data was achieved using Grubbs' Test. By applying Grubbs-Beck relation, values K_N and X_H were calculated as 2.682 and 1507 m^3/s. Based on this test, no outlier data was seen in Tamer hydrometric station. In order to be more confident and to conduct further studies, it is proposed to use several methods for outlier data detection and compare their results.

Acknowledgement

This paper is the result of the research plan "Comparison between Flood Frequency Analysis with Probable maximum Flood" in Islamic Azad University, Roudsar and Amlash branch, department of civil engineering.

References

1. Alberta Transportation, Civil Projects Branch., Guidelines on Flood Frequency Analysis, **(2001)**

2. Subcommittee on Hydrology, Bulletin 17-B guidelines for determining flood frequency, frequently asked questions, Water Information Coordination Program, Advisory Committee on Water Information, Hydrologic Frequency Analysis Work Group, http://acwi.gov/hydrology/Frequency/B17bFAQ.html., **(2008)**

3. Garcia F.A.A., Tests to identify outliers in data series, Pontifical Catholic University of Rio de Janeiro, Industrial Engineering Department, Rio de Janeiro, Brazil., **(2012)**

4. Reimann C., Filzmoser P. and Garrett R.G., Background and threshold: critical comparison of methods of determination, *Science of the Total Environment,* **346(1),** 1-16 **(2005)**

5. Hamed Khaled and Ramachandro Rao A., eds. Flood frequency analysis. CRC press, **(2010)**

6. Chou Y.M., Polansky A.M. and Mason R.L., Transforming non-normal data to normality in statistical process control, *Journal of Quality Technology,* **30(2),** 133-141 **(1998)**