Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 4 Issue 5, May - 2017

Outlier Detection in Extreme Value Series

Omer Levend Asikoglu

Ege University
Department of Civil Engineering
Bornova, Izmir, TURKEY
omerasikoglu@gmail.com

***Abstract* — An outlier in a data series is the data that is significantly detached from the rest of the series. Outliers in data series affects sample statistics like mean, standard deviation, coefficient of variation and coefficient of skewness and so the distribution model parameters and convenience level. To make a reliable frequency analyze it is necessary to carefully detect and remove the outliers. In this study, the outliers in annual maximum series of 14 flow gauging stations in the river basins of Seyhan and Ceyhan, in Turkey are analyzed with five different methods. The methods are compared and their findings are discussed briefly.**

Keywords— Outlier Detection; Hydrologic Data; Outliers in Extreme Value Series

I. INTRODUCTION

When the probabilities of hydrologic data are plotted, commonly one or two events appear to be from different population since they lie far from the line that represent other points. Dealing with these "outliers" is a controversial problem [1].

Grubbs [2] indicate an outlier as an observation that is statistically detached from the rest of the data.

Outlier values can arise from three different causes:

1. an error in measurement or recording.

2. an observation from a different population from that of most of the data, such as a flood triggered by a dam break rather than by rainfall.

3. an occasional event from a single population that is rather skewed.

Common application is to eliminate outliers by throwing them away prior to constructing database. But they should not always be eliminated. In some cases outliers may be the most important values in a data series, and should be examined further [3].

A geodetic study in 1838 by Friedrich Wilhelm Bessel, German mathematician and astronomer, is probability one of the early references to the elimination of outliers [4, 5]. Peirce developed a rejection criterion for outliers based on probability principles in 1852. He applied the outlier rejection criterion to 15 observations of the vertical semi-diameters of Venus [6]. This started a lively discussion, which continues until today.

Over the last two decades, many scientific studies have been published on the determination of outlier observations.

Pegram [7] performed a statistics-based outlier detection using regression and underlined the significance of detecting individual elements in data series, which is worth of further investigation.

Zhang et al. [8] used three methods for outlier detection (range, principle component analysis, and auto-association neural network) in an environmental geochemical data series. A mixture of all three methods was suggested for the improvement of a better outlier identifier.

Kondragunta [9] developed a technique, called the Spatial Consistency Check to detect outliers in rain gauge measurements. The outliers flagged using this method are further verified using independent precipitation estimates from radar and satellite.

Feng et al. [10] conducted quality control of daily meteorological data of 726 stations observed in China between 1951 and 2000, and found one or more inconsistent data in 37.9% of the stations.

Asikoglu [11] determined outliers with graphical method in daily maximum rainfall series.

Whitacre et al. [12] used statistical outlier detection methods in accordance with evolutionary algorithms.

Kirk et al. [13] conducted tests for 10 to 100 current hydrological data at various levels of confidence in their studies of outliers in multivariate hydrologic data and compared their findings with Rosner's [14] univariate test.

Filzmoser and Hron [15] used robust methods to detect outliers for composite data.

Sciuto et al. [16] conducted quality control of daily rainfall data by using neural networks. For quality control, they used confidence intervals obtained from

the neural networks based on the contemporaneous data of reference stations.

Cohn et al. [17] studied low-outliers and presented a generalization of the Grubbs-Beck test [18] to provide a reliable standard for detecting multiple potentially influential low flows.

Lamontagne and Stedinger [19] used the Spencer-McCuen test, an extended version of Grubbs-Beck test, to determine whether the three smallest observations are outliers in log-Pearson Type 3 (or Pearson Type 3) distributed data. They evaluated the performance of the Spencer-McCuen test With Monte Carlo experiments.

In this study, five different methods (the *z*-score method, Box Plot method, quality control (QC) test, Stedinger (modified Bulletin 17B) test, and the Grubbs–Beck (G-B) test) were used to detect outliers in maximum flow data series. The methods and their findings are discussed briefly.

II. METHODS

A. *z*-Score Method

This method computes the *z*-scores (the normalized values) $z_i = (x_i - \bar{x})/S_x$, where $\bar{x}$ is the sample mean and S_x is the sample standard deviation. An outlier is defined as the data in the series for which $|z_i|$ exceeds a limit value, typically 2.5. To standardize the data intends to convert them to a unit of the standard deviation so that the distance from the mean is expressed in comparable units.

Since the outlier affects both the mean and standard deviation, no data has a large *z*-score, and therefore none is identified as an outlier. However, using robust statistics in the *z*-score formula is a better way to successfully identify the outlier. Iglewicz and Hoaglin [20] proposed to use the modified *z*-score, as shown in the following formula:

$$z_i = 0.675(x_i - x_{0.5})/MAD \quad (1)$$

with *MAD* denoting the median absolute deviation:

$$MAD = \frac{1}{N}\sum_{i=1}^{N}|x_i - x_{0.50}| \quad (2)$$

and $x_{0.5}$ is the median. The modified *z*-scores with an absolute value of greater than 3.5 should be considered as potential outliers.

The use of *z*-scores to identify the outliers assumes that the considered variable is normally distributed.

B. Box-Plot Method

In box-plot method observations with the values between 1.5 and 3 box lengths from the upper or lower edge of the box are labeled as outliers. The length of the box is the inter-quartile range (*IQR*), which is equal to the difference between the case at the third quartile (Q_3 or $x_{0.75}$) and the case at the first quartile (Q_1 or $x_{0.25}$) in Fig. 1.

$$IQR = Q_3 - Q_1 \quad (3)$$

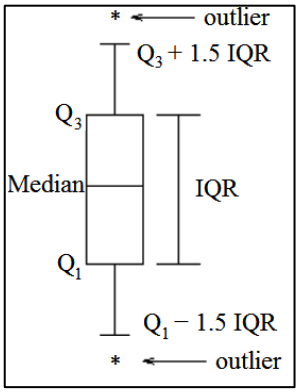

Figure 1 Box-plot

C. Quality Control Test

Quality-Control (QC) test involves four basic steps to identify the outliers [9].

1. Calculation of the median ($x_{0.50}$), and the quartiles ($x_{0.25}$ and $x_{0.75}$) of the data.
2. Calculation of the median absolute deviation (*MAD*).
3. Calculation of the test index as follows [21]:

if *MAD*= 0 , the test index =0
else
if $x_{0.75} \neq x_{0.25}$,
the test index = $|x_i - x_{0.50}|/(x_{0.75} - x_{0.25})$ (4)
else
the test index = $|x_i - x_{0.50}|/MAD$ (5)

4. The test index calculated in Step three is compared to a predefined threshold value (typically, 2). If the test index is bigger than the threshold value, then the data is labelled as an outlier.

D. Grupps-Beck Test

The Grubbs-Beck test [2, 18] defines high- and low-outlier thresholds as:

$$X_H = \bar{x} + k_N.S_x \quad (6)$$

$$X_L = \bar{x} - k_N.S_x \quad (7)$$

where $\bar{x}$ and S_x are the mean and the standard deviation of the sample data and the critical k_N values are given in Grubbs and Beck [18] according to the sample size *N* and significance level α.

If the data in a series is greater than X_H then it is considered high-outlier, if it is smaller than X_L then it is considered low-outlier.

E. Stedinger-Test

Interagency Advisory Committee on Water Data [22] proposed in US Department of Interior Geological Survey (USGS) Bulletin 17B the high- and low-outlier thresholds as follows:

$$X_H = \exp(\bar{y} + k_N . S_y) \quad (8)$$

$$X_L = \exp(\bar{y} - k_N . S_y) \quad (9)$$

where y is the logarithm of the data (y=ln$X$), $\bar{y}$ is the mean and S_y is the standard deviation of the logarithms. For the critical value k_N, Stedinger et al. [23] proposed an accurate approximation related to the sample size N ($5 \le N \le 150$) :

$$k_N = = -0.9043 + 3.345\sqrt{\log N} - 0.4046 \log N \quad (10)$$

The data in the series bigger than X_H and smaller than X_L are flagged as outliers.

III. APPLICATION

In this study, the annual maximum flows of 14 stations in two river basins in Turkey, namely Seyhan and Ceyhan, were examined for outlier detection. Table 1 shows important statistics and information of the stations in two river basins, and Fig. 2 illustrates the locations of the flow gauging stations.

Table 1. Important statistics and information of the stations

No	Station	Basin	N	Mean (m^3/s)	Max. (m^3/s)	Coeff. of variation	Coeff. of Skewness	Latitude	Longitude	Altitude
1	ASLANTAŞ	Ceyhan	35	905	1960	0,62	0,34	37° 15' N	36° 16' E	90
2	KILAVUZLU	Ceyhan	51	547	1629	0,50	1,47	37° 37' N	36° 47' E	450
3	MİSİS	Ceyhan	30	1011	2481	0,43	1,06	36° 57' N	35° 38' E	15
4	TANIR	Ceyhan	39	43	155	0,77	2,04	38° 25' N	36° 55' E	1180
5	KADİRLİ	Ceyhan	32	157	584	0,67	2,27	37° 22' N	36° 05' E	75
6	ÇUKURKÖPRÜ	Ceyhan	41	59	142	0,40	1,50	37° 20' N	35° 55' E	35
7	HANKÖY	Ceyhan	28	67	195	0,77	1,46	38° 15' N	37° 32' E	1349
8	POSKOFLU	Ceyhan	46	84	285	0,62	1,97	37° 08' N	37° 00' E	1040
9	KARAAHMET	Ceyhan	47	53	141	0,53	0,96	38° 01' N	36° 33' E	1324
10	HİMMETLİ	Seyhan	65	200	708	0,61	1,99	37° 51' N	36° 03' E	655
11	GÖKDERE	Seyhan	60	608	1963	0,55	1,80	37° 37' N	35° 36' E	312
12	ÜÇTEPE	Seyhan	34	1179	3278	0,50	1,64	37° 22' N	35° 28' E	130
13	HACILI KÖP.	Seyhan	32	155	528	0,66	2,19	37° 17' N	35° 09' E	167
14	FRAKTİN KÖP.	Seyhan	32	80	185	0,34	1,83	38° 14' N	35° 37' E	1270

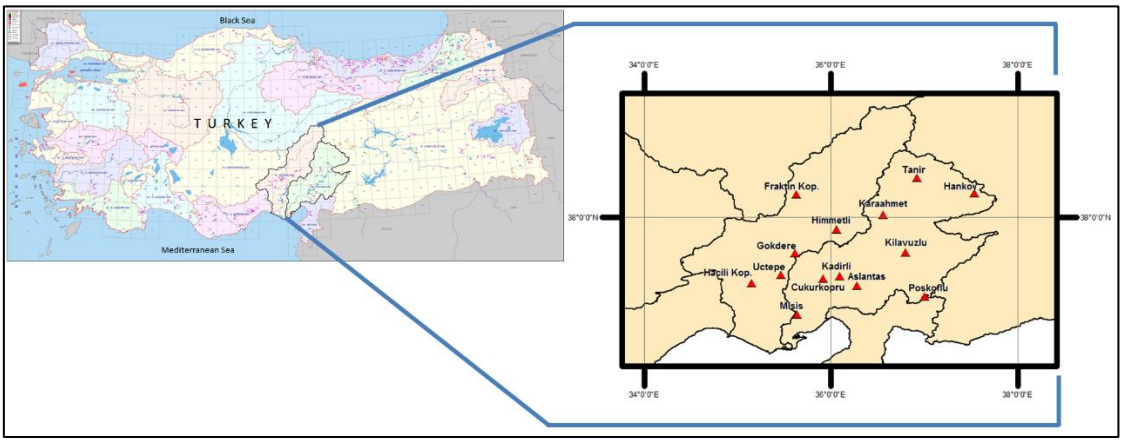

Figure 2 The locations of the flow gauging stations

IV. RESULTS and DISCUSSION

The potential outliers in the annual maximum flow series of 14 stations in Seyhan and Ceyhan basins are detected with five methods described above. Table 2 presents the results of five tests showing the year and the rank number of the outlier in the series.

Table 2 shows that,

- All the tests showed different precision in outlier detection.
- Box-plot and Q-C tests were the most precise ones and gave almost same results in outlier detection.
- *z*-score test was the second precise test by outlier detection.
- Stedinger and G-B tests detected minimum number of outliers compared to other tests.

When Table 2 was examined in detail, it is remarkable that,

- Amost in every station the observations of the years 1979 and/or 1980 were detected as outliers. So they are not based on measurement or recording errors, and they cannot be considered as outliers that should be removed from the series.
- The outliers in 1979 and 1980 have almost in every observation series the highest rank. So the outliers with lower values than the outliers in 1979 and 1980 also should not be removed from the series.
- As a result none of the high outliers detected in the series should be removed.

Table 2 The results of five outlier detection tests for 14 flow stations

Station	ASLANTAŞ	KILAVUZLU	MİSİS	TANIR	KADİRLİ	ÇUKURKÖPRÜ	HANKÖY	POSKOFLU	KARAAHMET	HİMMETLİ	GÖKDERE	ÜÇTEPE	HACILI KÖP.	FRAKTİN KÖP.
Station Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
z-score test	*Outlier(s) in Year(s)*													
		1980	1980	1980 (2)			1980	1980 (1)	1980	1980 (1)	1980 (2)		1980	1980
					1979			1979 (2)		1979 (3)	1979 (1)	1979		
				1969 (1)										
										1968 (2)				
						1966 (2)								
						1958 (1)								
box-plot test	*Outlier(s) in Year(s)*													
										2000 (4)				
												1996 (2)		
													1992 (2)	
		1980	1980	1980 (2)			1980 (1)	1980 (1)	1980	1980 (1)	1980 (2)		1980 (1)	1980
				1979 (3)				1979 (2)		1979 (3)	1979 (1)	1979 (1)		
													1977 (3)	
							1975 (2)				1975 (6)			
							1974 (3)							
													1972 (4)	
				1969 (1)				1969 (3)						
										1968 (2)				
						1966 (2)					1966 (5)			
						1963 (3)					1963 (3)			
						1958 (1)					1958 (4)			
QC test	*Outlier(s) in Year(s)*													
										2000 (4)				
												1996 (2)		
													1992 (2)	
		1980	1980	1980 (2)			1980 (1)	1980 (1)	1980	1980 (1)	1980 (2)		1980 (1)	1980
				1979 (3)	1979			1979 (2)		1979 (3)	1979 (1)	1979 (1)		
													1977 (3)	
				1975 (4)			1975 (2)				1975 (6)			
							1974 (3)							
													1972 (4)	
				1969 (1)				1969 (3)						
										1968 (2)				
						1966 (2)					1966 (5)			
						1963 (3)					1963 (3)			
						1958 (1)					1958 (4)			
G-B test	*Outlier(s) in Year(s)*													
														1994 (1)*
														1980 (1)
Stedinger test	*Outlier(s) in Year(s)*													
														1980

*: The solely low-outlier detected in the tests

- Other than this, one low outlier in the Fraktin Koprusu station detected with the G-B test, which can be considered as the only outlier within the framework of the study.

V. CONCLUSION

As mentioned by Helsel and Hirsch [3], outliers may be the most important points in the data set, and should be investigated further before they thrown away. Although the five tests carried out in this study have detected potential outliers in 14 annual maximum flow series, none of them should considered as outliers and they shouldn't be rejected from the series as a result of detailed interpretation.

REFERENCES

[1] Haan, C.T., Statistical Methods in Hydrology, Iowa State Press, 2002.

[2] Grubbs, F. E., "Procedures for detecting outlying observations in samples". Technometrics, 11, 1969, pp.1–21

[3] Helsel, D.R., Hirsch, R.M., Statistical Methods in Water resources. Techniques of Water Resources Investigations Description. Book 4, Chapter A3. U.S. Geological Survey. 2002, 522 pp.

[4] Bessel, F.W., "Determination of the distance of the 61st star of the Swan". Astronomische Nachrichten, 16,1838, pp. 65–96.

[5] Anscombe, F.J., "Rejection of outliers". Technometrics. v2., 1960, pp. 123-147.

[6] Peirce B., "Criterion for the Rejection of Doubtful Observations." The Astronomical Journal, 2(45), 1852, pp. 161-163.

[7] Pegram, G., "Patching rainfall data using regression methods. 3. Grouping, patching and outlier detection", Journal of Hydrology, 198, (1), 1997, pp. 319–334.

[8] Zhang, C., Wong, P.M., Selinus, O. "A Comparison of Outlier Detection Methods: Exemplified with an Environmental Geochemical Dataset", 6th International Conference on Neural Information Processing, Proceedings. ICONIP '99, 1999.

[9] Kondragunta, C. R., "An outlier detection technique to quality control rain gauge measurements". Eos Trans. Amer. Geophys. Union, 82 (Spring Meeting Suppl.), Abstract H22A-07A, 2001.

[10] Feng, S., Hu, Q., Qian, W., "Quality Control of Daily Meteorological Data in China", 1951–2000: A New Dataset, Int. J. Climatol, 24, 2004, pp. 853–870.

[11] Asikoglu, O. L., Generalized Intensity-Duration-Frequency Models for Maximum Rainfall of Standard Durations, PhD. Thesis, Ege University, Turkey, 2005.

[12] Whitacre, J.M., Pham, T.Q., Sarker, R.A., "Use of Statistical Outlier Detection Method in Adaptive Evolutionary Algorithms", GECCO'06, July 8–12, Seattle, Washington, USA, 2006.

[13] Kirk, A.J., McCuen, R. H., "Outlier Detection in Multivariate Hydrologic Data", Journal of Hydrologic Engineering, Vol.13, No. 7, 2008.

[14] Rosner, B. "On the Detection of Many Outliers", Technometrics, 17, 1975, pp. 221–227.

[15] Filzmoser, P., Hron, K., "Outlier Detection for Compositional Data Using Robust Methods", Math. Geosci., 40, pp. 233–248, 2008.

[16] Sciuto, G., Bonaccorso, B., Cancelliere, A., Rossi, G., "Quality control of daily rainfall data with neural networks", Journal of Hydrology 364, 2009, pp. 13– 22

[17] Cohn, T. A., England, J. F., Berenbrock, C. E., Mason, R. R., Stedinger, J. R., Lamontagne, J. R., "A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series", Water Resources Research, 49, 2013, pp. 5047–5058.

[18] Grubbs, F.E.; Beck, G., "Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations", Technometrics, 14 (4), 1972, pp. 847-854.

[19] Lamontagne, J. R., Stedinger, J. R., "Examination of the Spencer-McCuen Outlier-Detection Test for Log-Pearson Type 3 Distributed Data", J. Hydrol. Eng., 21 (3), 2016.

[20] Iglewicz, B., Hoaglin, D., How to detect and handle outliers. ASQC Quality Press, 1993.

[21] Madsen, H., "Algorithms for corrections of error types in a semi-automatic data collection". Precipitation Measurement and Quality Control. B. Sevruk & M. Lupin (eds.), Proc. of Symp. on Precipitation and Evaporation, Vol. 1, Bratislava, Slovakia, September 20-24, 1993

[22] Interagency Advisory Committee on Water Data (IACWD). Guidelines for determining flood flow frequency, Bulletin 17B of Hydrology and Subcommittee. U.S. Geological Survey, Office of Water-Data Coordination, 1982.

[23] Stedinger, J. R., Vogel, R.M. and Foufoula-Georgiou, E., Frequency Analysis of Extreme Events, chap. 18, p. 99, McGraw Hill, New York, 1993.