FACETS OF UNCERTAINTY
Kos Island, Greece | 17–19 October 2013

European Geosciences Union
International Association of Hydrological Sciences
International Union of Geodesy and Geophysics

5th EGU LEONARDO CONFERENCE • HYDROFRACTALS '13 • STATISTICAL HYDROLOGY—STAHY '13

Outlier Detection of Extreme Value Series

Omer Levend Asikoglu[1] and Ebru Eris[1]
[1]Ege University, Izmir, Turkey
omer.asikoglu@ege.edu.tr, ebru.eris@ege.edu.tr

INTRODUCTION

In statistics, an outlier is an observation that is numerically distant from the rest of the data (Grubbs, 1969).
Outliers can have one of three causes:

1. a measurement or recording error.
2. an observation from a population not similar to that of most of the data, such as a flood caused by a dam break rather than by precipitation.
3. a rare event from a single population that is quite skewed.

Outliers are often dealt with by throwing them away prior to describing data, or prior to some of the hypothesis test procedures. But they should not. Outliers may be the most important points in the data set, and should be investigated further (Helsel and Hirsch, 2002).
In this study, five different methods (the z-score method, Box Plot method, QC (Quality Control) test, Maidment (modified Bulletin 17B) method and the G-B (Grubbs–Beck) test) were used to detect outliers in maximum flow data series.

1 z-Score Method

This method computes the z-scores, which are just the normalized values,

$$z_i = (x_i - \bar{x}) / S$$

where $\overline{X}$ is the sample mean and S_x is the sample standard deviation. An outlier is defined as any observation for which exceeds some cutoff value, typically 2.5. Standardizing variables converts them to a standard deviation unit of measurement so that the distance from the mean for any case on any variable is expressed in comparable units.

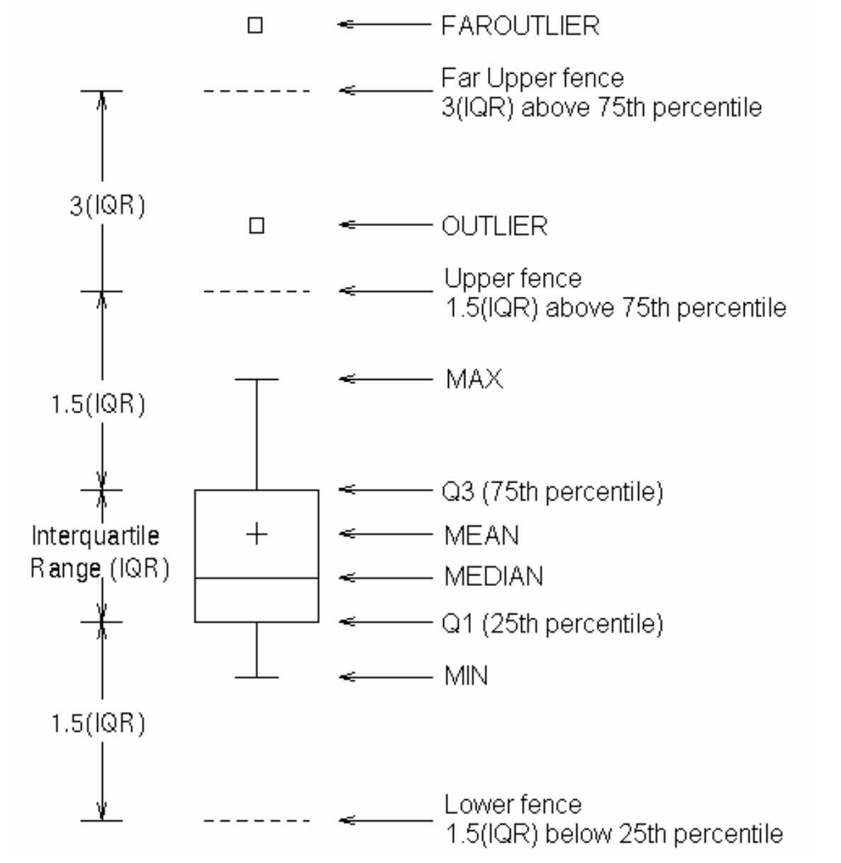
$$z_i = 0.675(x_i - med\ x) / MAD$$

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - X_{0.50}\right|$$

2 Box-Plot Method

A box-plot identifies outliers using a somewhat different criterion. Cases with values between 1.5 and 3 box lengths from the upper or lower edge of the box are identified as outliers. The box length is the inter-quartile range (*IQR*), or the difference between the case at the 75th quartile ($x_{0.75}$ or Q3) and the case at the 25th quartile ($x_{0.25}$ or Q1).

$$IQR = x_{0.75} - x_{0.25}$$



3 QC (Quality Control) Method

This method is used to identify outliers which are not spatially consistent with the neighboring rain gauges (Kondragunta, 2001). There are four steps involved in the identification of outliers:

✓ Determination of median ($x_{0.50}$), 25th ($x_{0.25}$) and 75th ($x_{0.75}$) percentile of the data points.
✓ Calculation of M*AD* (the median absolute deviation).
✓ Calculation of test index (Madsen, 1993) for each station as follows:

if *MAD*= 0 , index =0
else
if $x_{0.75} \neq x_{0.25}$, index = $\left|x_i - x_{0.50}\right| /(x_{0.75} - x_{0.25})$
else
index = $\left|x_i - x_{0.50}\right| / MAD$

✓ The index calculated in Step three is compared to a predefined threshold value (typical value is 2). If the index is greater than the predefined threshold value, then the data is flagged as an outlier.

4 Maidment Method

The USGS (US Geological Survey) proposed in Bulletin 17B the high- and low-outlier thresholds as follows:

$$X_H = exp(\bar{y} + k_N .S_y) \qquad X_L = exp(\bar{y} - k_N .S_y)$$

where *y* is the logarithm of the systematic peaks (*y=lnX*). k_N values are given in Bulletin 17B as Appendix 4 for sample size *N*. Maidment (1993) proposed following Equation for k_N :

$$k_N = = -0.9043 + 3.345\sqrt{log\ N} - 0.4046\ logN$$

5 G-B Test Method

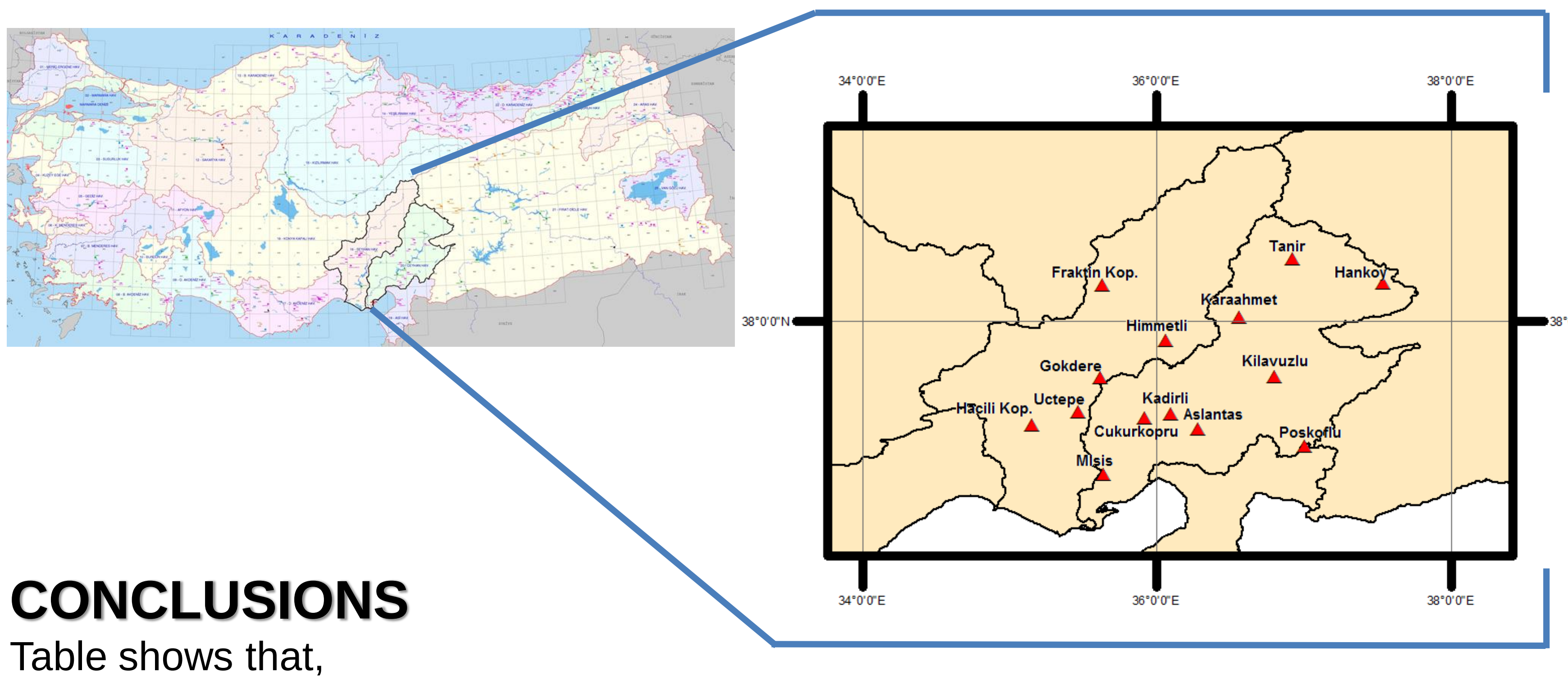
G-B test method (Grubbs and Beck,1972) is also one of the commonly used outlier detection methods. x_H and x_L are the high- and low-outlier thresholds which were described in USGS Bulletin 17B:

$$X_H = exp(\bar{y} + k_N .S_y) \qquad X_L = exp(\bar{y} - k_N .S_y)$$

k_N is the critical value calculated according to the sample size *N* and significance level *α*. The k_N critical values are given in Tables for given sample size and significance levels. Later Pilon et al.(1985) formulized k_N values for = 0.10 as follows:

$$k_N = - 3.62201 + 6.28446N^{1/4} - 2.49835N^{1/2} + 0.491436N^{3/4} - 0.037911N$$

STUDY AREA



RESULTS

The maximum annual flows of 14 stations in two neighbor watersheds, Seyhan and Ceyhan, were examined for outlier detection with five methods. All the outliers detected with the five methods were high outliers except one low outlier in the Fraktin Köprüsü station detected with the G-B test. The results of the tests were given in Table.

Station	ASLANTAŞ	KILAVUZLU	MİSİS	TANIR	KADİRLİ	ÇUKURKÖPRÜ	HANKÖY	POSKOFLU	KARAAHMET	HİMMETLİ	GÖKDERE	ÜÇTEPE	HACILI KÖP.	FRAKTİN KÖP.
N (year)	35	51	30	39	32	41	28	46	47	65	60	34	32	32
River	CEYHAN	CEYHAN	CEYHAN	HURMAN	SAVRUN	SONBOS	SÖĞÜTLÜ	GÖKSU	GÖKSU	GÖKSU	GÖKSU	SEYHAN	KÖRKÜN	ZAMANTI
Catchment	CEYHAN	CEYHAN	CEYHAN	CEYHAN	CEYHAN	CEYHAN	CEYHAN	CEYHAN	CEYHAN	SEYHAN	SEYHAN	SEYHAN	SEYHAN	SEYHAN
z-scale 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)			
Maidment test	Outlier(s) in Year(s)													
														1980
G-B test	Outlier(s) in Year(s)													
														1994 (1)*
														1980 (1)

*The solely low-outlier detected in the tests

CONCLUSIONS

Table 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
❑ Maidment and G-B tests (both based on USGS's Bulletin 17B) detected minimum number of outliers compared to other tests.

If the Table was examined in detail, it is remarkable that,

❑ Almost 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. And 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.
❑ Other than this, one low outlier in the Fraktin Köprüsü station detected with the G-B test, which can be considered as the only outlier within the framework of the study.
❑ As mentioned by Helsel and Hirsch (2002), outliers may be the most important points in the data set, and should be investigated further before they thrown away.

SELECTED REFERENCES

ANSCOMBE, F.J. 1960. Rejection of outliers. Technometrics. v2. 123-147
GRUBBS, F.E.; BECK, G., 1972, Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations, Technometrics, Vol. 14 (4), pp. 847-854.
HELSEL, D.R., HIRSCH, R.M., 2002. Statistical methods in water resources. Techniques of Water Resources Investigations Description. Book 4, chapter A3. U.S. Geological Survey. 522 pp.
KONDRAGUNTA, C. R., 2001, An outlier detection technique to quality control rain gauge measurements. Eos Trans. Amer. Geophys. Union, 82 (Spring Meeting Suppl.), Abstract H22A-07A.
MAIDMENT, D.R., 1993, Handbook of Hydrology, Frequency Analysis of Extreme Events, Chapter 18, McGraw-Hill, Inc., New York.