



ORIGINAL ARTICLE

Determination of A Some Simple Methods for Outlier Detection in Maximum Daily Rainfall (Case Study: Baliglichay Watershed Basin – Ardebil Province – Iran)

Sajad Mirzaei^{1*}, Majid Raoof², Ali Ghasemi³, Abazar Esmali⁴, Hasan Etaati⁵, Maryam Moradnezehadi⁶, Yadollah Mirzaei⁷

¹Department of Range & watershed Management, University of Mohaghegh Ardabili, Iran

²Dept. of Water Engineering, University of Mohaghegh Ardabili, Iran

^{3,5}Department of Watershed Management, University of Tehran

⁴Member of Water shedding Board - University of Mohaghegh Ardabili, Iran

⁶Department of Forestry Engineering –University of Tehran

⁷Dept. Range & Watershed Management, University of Mohaghegh Ardabili, Iran

*Corresponding author: s.mirzayi00@gmail.com,

ABSTRACT

Outliers in maximum daily rainfall can play a considerable role in unreal analysis leading to unreal predictions. Therefore, accurate statistical determination of data to find outliers is very important. In the present work, data related to maximum daily rainfall obtained from 12 rain gauge stations belonging to Regional Water Corporation of Ardebil Province inside and around Baliglichay watershed basin were used. Five methods for outlier detection (namely $3 \times SD \pm M$, Box-Whisker Plot, American Water Resources Association, Grubbs' Test, and Mahalanobis Distance) were adopted. After detecting outliers by the mentioned methods, three control stations were considered in terms of distance, location, and climate. The control stations consisted of 9 rain gauge stations for regional water and 5 meteorological rain gauge stations. Afterwards, outliers in each station were controlled by simultaneous data achieved from the control stations through the five mentioned methods. The results obtained from the present study revealed that only 5 out of 14 outliers detected in 3 out of 7 stations were confirmed. Furthermore, comparison of performance of the methods showed that American Water Resources Association and $3 \times SD \pm M$ methods were more efficient to detect outliers while Mahalanobis Distance had the lowest efficiency.

Keywords: Baliglichay watershed, Grubb's Test, American Water Resources Association, maximum daily rainfall, outlier.

Received 12/11/2013 Accepted 09/01/2014

©2014 AELS, INDIA

INTRODUCTION

Description of meteorological variables such as heavy rainfall is cumbersome due to haphazard climatic variations and sampling error in limited data as a small sample belonging to an indefinite statistical population [21]. Maximum daily rainfall is a meteorological parameter with highly haphazard nature compared with rainfall data such as monthly and annual rainfall [8]. Adoption of maximum rainfall in estimation of flood and amount of runoff in small or without long-term hydrometric data regions have long been in center of attention. According to the relations provided by SCS, hourly rainfall can be estimated from maximum daily rainfall with varying turnover times. The relations are very important in estimation of hydrographs. With regard to lack of data related to short-term rainfall in many parts of the world, estimation of such information is performed through the methods concerning maximum daily rainfall [4]. Several application for annual maximum daily rainfall with varying turnovers can somewhat reflect importance of such rainfall. Therefore, before calculations of floods and designing water structures, more precise daily rainfall data in terms of existence of outliers is necessary. An outlier can be considered an observation which abnormally or haphazardly deviates from normal status of data and analysis [1]. As a matter of fact, one of necessary factors in abundance analysis is existence of real and long-term data where outliers may have considerable adverse effect in data prediction [3]. Such effect leads to unreal estimations in data analyses for a given variable; for instance:

- (1) Outliers lead to unreal probabilistic distribution;
- (2) Estimation of probabilistic distributions parameters is influenced by outliers leading to faulty estimation;
- (3) Low outliers cause underestimation leading to structure collapse in big floods;
- (4) High outliers cause overestimation leading to higher costs of water structures building (2).

Thus, false inputs result in false outputs so that an outlier in some statistical methods can specify final results [6]. Several methods have been proposed for outlier detection, none of which have been globally accepted [18]. Such methods as main attribute [9,11], multivariable regression and artificial network [16], genetic algorithm [25], and Mahalanobis Distance-based method [19,20] were presented with the last one as being the most famous (15). Some Iranian authors have performed studies on effect of outliers in such fields as hydrology and sedimentology. For instance, (3) evaluated effect of outliers in hydrologic abundance analysis by use of linear torque method in Jiroft Dam region in Iran on annual maximum rainfall series and concluded that outliers have no influence on regional distribution; however, they were found to exert a considerable effect on determination of the most suitable statistical distribution for data relating to a station. [6] adopted main attributes method in order to detect outliers in regional analysis of flood in Lake Urmia region in Iran and found that this method has high precision. [2] used univariate detection methods such as $3 \times SD \pm M$, median \pm outliers median, Box-Whisker Plot, and Grubbs' Test and concluded that median \pm outliers median and Box-Whisker Plot are more efficient than $3 \times SD \pm M$ and Grubbs' test. Many of these methods need a lot of calculations and time and are not computable by statistical software. Therefore, more than one method can be adopted for outlier selection to assure higher accuracy. On this basis, five simple methods were used in the present study in order to detect outliers in maximum daily rainfall data in the stations located in and around Baliglichay watershed in northwest of Iran.

MATERIALS AND METHODS

Stations and the region

Baliglichay watershed is one of subsidiary regions of Gharasou watershed basin located in Pol-e-Almas hydrometric station (2014 m high and 1036.18 km²) 85% of which is located in Ardebil Province and the rest is located in Eastern Azerbaijan Province. Baliglichay River originates from Sabalan Mountain in northwest and Bozghoosh in the south. The river is one of the most important rivers in northwest of Iran as it flows through Ardebil City and it is considered as the source of drinking and agricultural water. In order to detect outliers in maximum daily rainfall data, data related to 12 rain gauge stations with a 27-year statistical term (1983-2009) were used because of proper distribution of the station; location of the stations and Baliglichay watershed were depicted in Fig. 1.

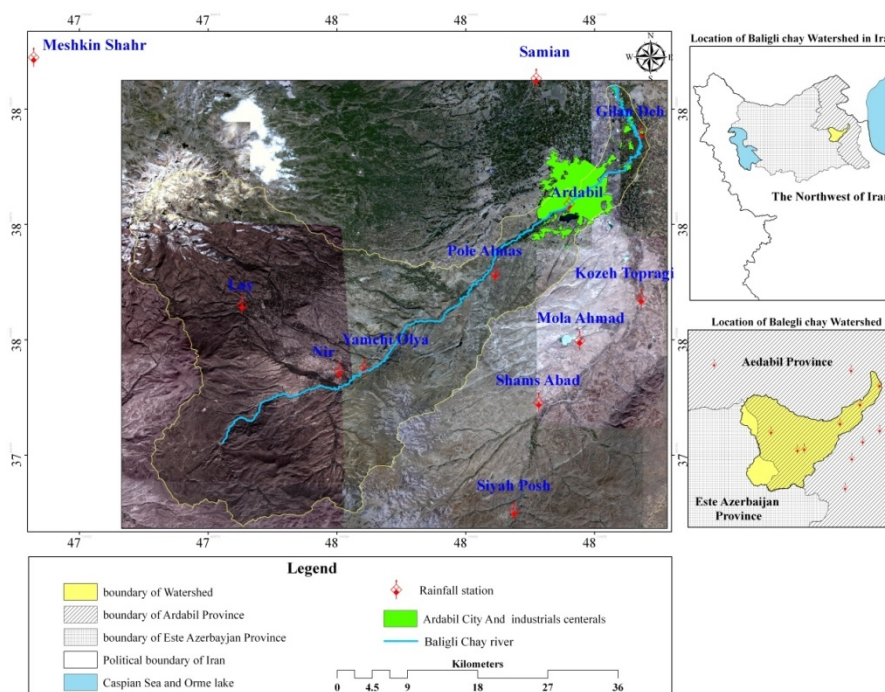


Figure 1: Location of Baliglichay watershed and distribution of the stations

Outlier detection methods

Outlier detection methods can be generally divided into three groups: univariate, two-variate, and multivariate detection methods. Multivariate outlier detection method consists of multivariate determination of each observation according to a blend of variables. Also, there are various methods in order to perform univariate outlier detection which can be classified into range and statistical tests. In range methods, observations distribution is determined and data out of a given range are considered outliers; in statistical methods, all outlier detection methods are based upon the principle that an outlier is a doubtful observation which is somewhat or fully out of the subject because the data were not made by hypothetical models of random variable [7,10]. In the present work, the univariate methods $3 \times SD \pm M$, Box-Whisker Plot, American Water Resources Association, Grubbs' Test, and Mahalanobis Distance were adopted so as to detect outliers. After performing calculation of each method and detection of outliers in maximum daily rainfall data, the nearest stations' data were adopted in order to assure correctness of outliers where for each station with outliers, 3 control station were assigned in terms of the nearest distance, location, and climate. Rainfall data for 9 rain gauge stations for regional water and 5 meteorological rain gauge stations were considered as control stations. In order to make suitable decision, the data exceeding the lower threshold were eliminated from outliers [5]. Then, high outliers were compared with the nearest stations. When there were similar simultaneous data in control stations, the data were not regarded as outliers and introduced into maximum daily rainfall data; otherwise, they were considered as outliers. This way, performance of outlier detection methods was compared.

$3 \times SD \pm M$

One of the most important and well-known univariate outlier detection methods is $3 \times SD \pm M$. This conventional method is based upon the principle that first, data mean and $3 \times SD$ are obtained; then, data higher than $3 \times SD + M$ or lower than $3 \times SD - M$ are regarded as outliers and eliminated from data set [28].

Box-Whisker Plot

The method is widely used in order to detect outliers [20,18]. Box-Whisker is a plot presenting dataset very effectively by aid of central and distribution criteria. The plot was proposed by famous statistician Tukey [24]. Box plot is a method for summarizing estimated data according to a distance index which is adopted for discovering properties and data analysis [10]. In this method, a rectangle (box) and two lines around the rectangle (whisker) are used and the plot is drawn by median, the first and third quarters, and the lowest and highest observed amount. The rectangle length equals interquartile range ($IQR = Q3 - Q1$). Internal and external frontiers are specified by successive steps. When a data is placed between internal and external boundaries, it is called weak outlier (data smaller than $Q1 - 1.5IQR$ and bigger than $Q1 + 1.5IQR$) and if it is placed out of the external boundaries, it is regarded as strong outlier (data smaller than $Q1 - 3IQR$ and bigger than $Q1 + 3IQR$).

2-2-3- Grubbs' Test

The test was invented by Grubbs to detect outliers. In this method, the test has a statistic for each data obtained by the formulae mentioned by Grubbs and then compared with standard amount of each data. If it exceeds the standard, it is considered an outlier (14). There are two hypotheses in this test; null hypothesis is there is no outlier and the other hypothesis is there is at least one outlier. Statistic of this test (G) is derived as follows (23,14):

$$G = \frac{\max |x_i - \bar{x}|}{S} \quad (1)$$

where x_i , \bar{x} , and S stand for the biggest or smallest data, data mean, and standard deviation, respectively; the null hypothesis is rejected when the following relation is achieved:

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

where n and $t^2_{(\alpha/2, n-2)}$ stand for sample size and critical amount of t-student distribution statistic with

freedom degree of $n-2$ and significance level of $\alpha/2n$ ($\alpha=0.05$), respectively.

2-2-4- American Water Resources Association

This method was first proposed by American Water Resources Association and used for elimination of outliers. In this method, data kurtosis is obtained; if kurtosis exceeds +0.4, the test is done for high outliers and vice versa. If kurtosis is between the two ranges, the tests for both high and low outliers are performed. Detecting high and low outliers are done via the relations 3 and 4, respectively (5).

$$y_H = \bar{y} + k_N s_y \quad (3)$$

$$y_L = \bar{y} - k_N s_y \quad (4)$$

where y_H , y_L , \bar{y} , s_y , and k_N stand for high outliers' decimal logarithm, low outliers' decimal logarithm, data mean logarithm, SD of data logarithm, outliers abundance coefficient, respectively; k_N is derived as follows at significance level of 10% and $10 < \text{samples} < 150$:

$$k_N = -3.62201 + 6.23446N^{-.25} - 2.47832N^{-.5} + 0.491436N^{-.75} - 0.037911N \quad (5)$$

2-2-5- Mahalanobis Distance

Mahalanobis Distance acts as a criterion of multidimensional location of each observation relative to centroid of all bservations; in other words, the distance is a criterion of distance between each observation in multidimensional space and mean center of all observations. Advantage of Mahalanobis Distance compared to other distances is its considering covariance matrix (12). For a p-dimensional multivariate sample, Mahalanobis Distance for i^{th} observation is obtained as follows:

$$MD_i = (x_i - \bar{x})^T c^{-1} (x_i - \bar{x})^{1/2} N \quad (6)$$

where x_i , \bar{x} , and c stand for variables' vector for i^{th} observation, mean variables' vector (centroid of observations), and sample covariance matrix, respectively.

RESULTS AND DISCUSSION

The results showed that 7 out of 12 stations suffered from outliers. Figures 2 and 3 depict the results obtained from Box-Whisker Plot and Mahalanobis Distance methods, respectively. Also, Table 1 summarized the results acquired from all five methods. The findings indicate that each method had a different outcome for outlier detection; $3 \times SD \pm M$ and American Water Resources Association methods detected the lowest outliers (one high outlier and 5 low outliers) while Box-Whisker Plot detected the most outliers (9 high outliers). The methods used in the present study detected 14 outliers (10 high and 4 low outliers) altogether. After specifying the stations suffering from outliers, characteristics of these stations (stations of origin) along with control stations are presented in Table 2 and Figure 4.

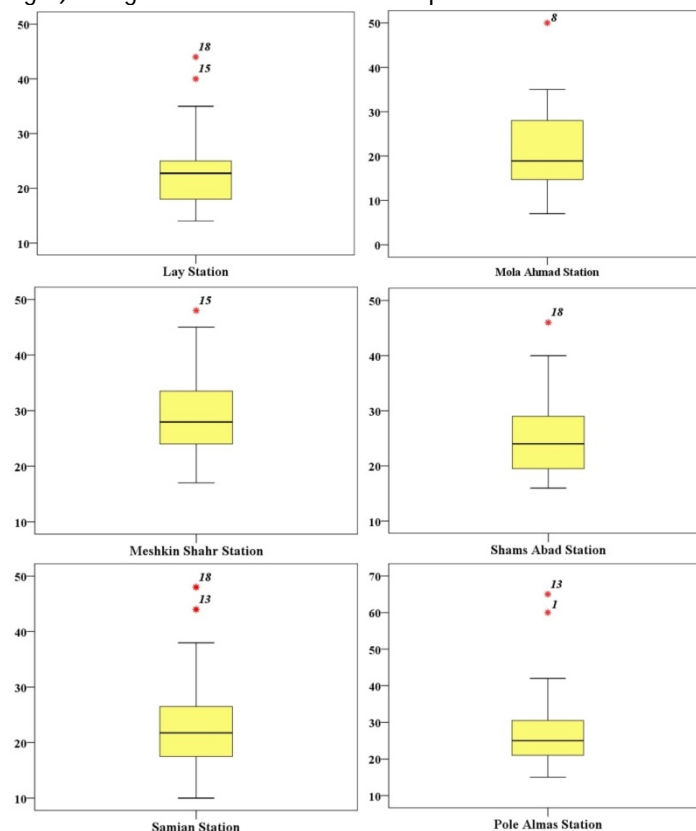


Figure 2: the results obtained from Box-Whisker Plot in the stations with outliers

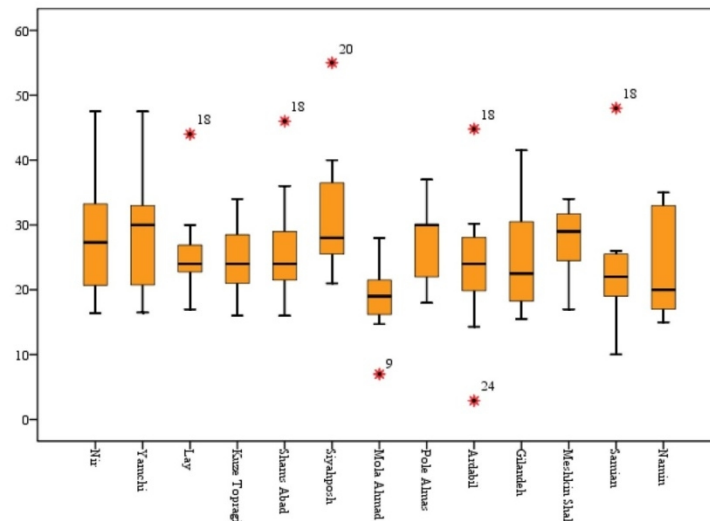


Figure 4: the results acquired from Mahalanobis Distance in the stations

Table 2: The calculated outliers for maximum daily rainfall by all the 5 methods in the stations

	3×SD±M		Box-Whisker Plot		American Resources Association		Grubbs' Test		Mahalanobis Distance	
Station	High outlier	Low outlier	High outlier	Low outlier	High outlier	Low outlier	High outlier	Low outlier	High outlier	Low outlier
Mola Ahmad	-	8.5	-	-	-	8.5	-	-	-	-
	-	7	-	-	-	7	-	-	-	7
	-	8	-	-	-	8	-	-	-	-
	-	-	50	-	-	-	50	-	-	-
Ardebil	-	2.9	-	-	-	2.9	-	2.9	2.9	2.9
	-	-	-	-	-	-	-	-	44.8	-
Lay	-	-	40	-	-	-	-	-	-	-
	-	-	44	-	-	-	44	-	44	-
Shams Abad	-	-	46	-	-	-	46	-	46	-
Almas	-	-	60	-	-	-	-	-	-	-
	65	-	65	-	65	-	65	-	-	-
Samian	-	-	44	-	-	-	-	-	-	-
	-	-	48	-	-	-	48	-	48	-
MeshkinShahr	-	-	48	-	-	-	48	-	-	-
Outliers	1	1	9	3	1	1	6	1	5	2

Table 3: Characteristics of the stations with outliers (stations of origin) and control stations

Station of origin	Altitude (m)	Control stations	Station type	Distance (km)	Altitude (m)
Mola Ahmad	1512	KozehTopragi	Regional water	8.15	1402
		Shams Abad	Regional water	8.9	1534
		Almas	Regional water	12	1480
Ardebil	1365	Almas	Regional water	11.4	1480
		Gilandeh	Regional water	11.4	1341
		KozehTopragi	Regional water	14	1402
Lay	2038	Nir	Regional water	12.8	1623
		Atashgah	Regional water	17.1	1778
		Kalur	Regional water	21	1576
Shams Abad	1534	Koraim	Regional water	5.3	1534
		Mola Ahmad	Regional water	8.9	1512
		Foladlu	Regional water	7.2	1473
Almas	1480	Ardebil	Regional water	11.4	1365
		KozehTopragi	Regional water	15	1402
		Nir	Regional water	20	1623
Samian	1286	Gilandeh	Regional water	12.6	1341
		Ardebil	Regional water	16.2	1365
		Namin	Regional water	21.3	1405
MeshkinShahr	1485	West Meshkin	Regional water	10	1244
		East Meshkin	Regional water	22	1211
		Mashiran	Regional water	32	705

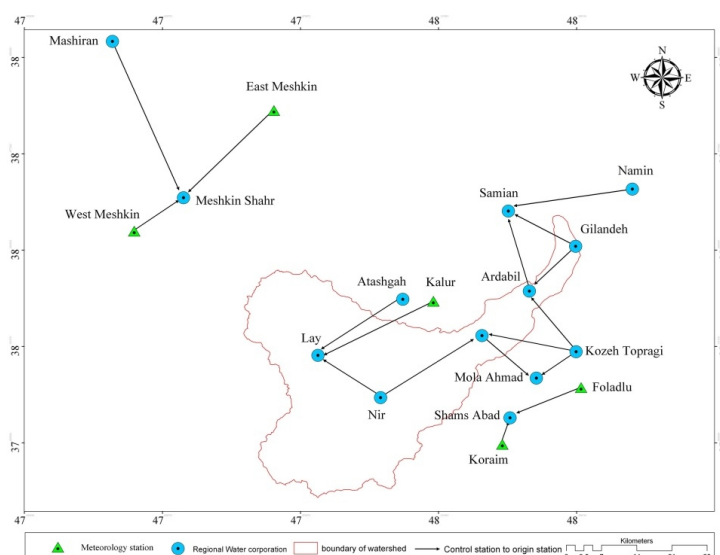


Figure 4: the stations with outliers (stations of origin) and control stations

The results obtained from the comparison of stations of origin and control stations (Table 3) indicate that out of 14 outliers detected in 7 stations by the five methods, only 5 outliers related to three stations (namely Mola Ahmad, Ardebil, and Almas) are confirmed because statistical determination of control stations showed that rainfall rates in 9 cases were close to detected outliers in stations of origin. Therefore, the data are of historical data for the stations and they cannot be considered as outliers. In order to determine efficiency of the methods, the ratios of detected outliers to real outlier (real percentage) and also to unreal outliers were measured for each method.

Table 3: the results obtained from determination of outlier detection for maximum daily rainfall relative to data obtained from control stations

Station	Occurrence date	Outlier	Control station 1	Control station 2	Control station 3
Mola Ahmad	1988//3	8.5	2	-	10
	1992//25	50	19	25	30
	1993//14	7	-	-	-
	1991//28	8	-	-	4
Ardebil	2007//2	2.9	-	-	-
	2001//25	44.8	30	28	27
Lay	1998//23	40	32	-	35
	2001//25	44	37	42	37
Shams Abad	2001//25	46	40	49.8	22
Almas	1984//7	60	42	49	10
	1996//7	65	12.4	21	15
Samian	1996//5	44	31	41	37
	2001//25	48	28	44.8	35
MeshkinShahr	1998//14	48	41	34	26.5

The results shown in Table 4 indicate that American Water Resources Association and $3 \times SD \pm M$ methods had the best efficiency by detecting 5 real outliers and not detecting unreal outliers. Moreover, Grubbs' Test and Box-Whisker Plot were more efficient in detection of high outliers than low outliers; however, these methods suffered from unreal outliers. Furthermore, Mahalanobis Distance was the least efficient method among others.

Table 4: comparison of efficiency of outlier detection methods for maximum daily rainfall in the stations

Outlier detection methods	Number of outliers			True percentage			Unreal outliers		
	High outlier	Low outlier	Total	High outlier	Low outlier	Total	High outlier	Low outlier	Total
$3 \times SD \pm M$	1	4	5	%100	%100	%100	0	0	0
Box-Whisker Plot	9	3	12	%100	%75	%80	6	0	6
American Water Resources Association	1	4	5	%100	%100	%100	0	0	0
Grubbs' test	6	1	7	%100	%25	%50	5	1	6
Mahalanobis Distance	5	2	7	%0	%50	%20	5	1	6

CONCLUSION

With regard to the fact that maximum daily rainfall data is one of the most useful data for designing water structures as well as water and soil conservation programs, outlier detection is one of crucial stages in the relative studies and whatever kind of error in outlier detection will lead to unreal predictions and estimations in designs and simulations. Maximum rainfall data outliers lead to overestimation and consequently, higher design and water and soil conservation costs. On the other hand, if the historical outliers are eliminated,

REFERENCES

1. JabbariNoughabi, M. (2011) Outliers, definition and application; Student Journal of Statistics, Vol. 9, No. 1, p. 1-16.
2. Hakimkhani, S., Alijanpour, A. (2010) Recognition of outliers by sediment origin method; Journal of Water and Soil Conservation Research, Vol. 17, No. 1, p. 23-43.
3. Soleimani Sardou, F., Soleimani, A., Kouhestani, S. (2010) Role of outliers in hydrologic abundance analysis by linear torque in Jiroft Dam; the 2nd conference on comprehensive management of water resources, Kerman, Shahid-Bahonar University of Kerman, Water and Irrigation Engineering Council.
4. Matinzadeh, M., Fattahi R., Shayannejad, M., Abdollahi, K. (2011) Estimation and restoring maximum 24-h rainfalls in Chaharmahal-e-Bakhtyari Province by fuzzy regression; Iranian Journal of Water Research, Vol. 5, No. 8, p. 179-186.
5. Mahdavi, M. (2009) Applied hydrology; Vol. 2, No. 3, Publications of Tehran University, p. 38-114.
6. NouriGhidari, M. (2010) Recognition of outliers in regional flood analysis by main components analysis; the 5th national congress on civil engineering, Mashhad, Ferdowsi University – Mashhad.
7. Anscombe, F. J. and Guttman, I. 1960. Rejection of outliers. *Technometrics* 2, 2, 123{147}.
8. Barker, B. Tailor, G. and Wallilis, G. 2006. Regional precipitation frequency analysis and spatial mapping of precipitation for 24-hour and 2-hour duration in Washington. *Journal of Climatology* 38(2):1871-1884.
9. Caussinus, H., Fekri, M., Hakam, S., and Ruiz-Gazen, A. 2003. A monitoring display of multivariate outliers. *Computational Statistics and Data Analysis*, 44: 237-252.
10. Chandola, V., Banerjee, A., and Kumar, V. 2007. Outlier Detection A Survey. University of Minnesota, pp: 36-45.
11. Chiang, L.H., Pell, R.J., and Seasholtz, M.B. 2003. Exploring process data with the use of robust outlier detection algorithms. *J. Process Control*, 13: 437-449.
12. Filzmoser, P., Garrett, R.G., and Reimann, C. 2005. Multivariate outlier detection in exploration geochemistry. *Computers and Geosciences*, 31: 579-587.
13. Grubbs, F. E., 1969, Procedures for Detecting Outlying Observations in Samples, *Technometrics*, vol. 11, No. 1, pp. 13-14.
14. Grubbs, F. E., 1950, Sample Criteria for Testing Outlying Observations, *Annals of Math. Statistics*, vol. 21, pp. 27-58.
15. Hair, J.F., Andersen, R.E., Tatham, R.L., and Black, W.C. 1998. *Multivariate Data Analysis*. Prentice Hall, Upper Saddle River, New Jersey.
16. Lalor, G.C., and Zhang, C. 2001. Multivariate outlier detection and remediation in geochemical databases. *The Science of the Total Environment*, 281: 99-109.
17. Rao, A. R. And, Kh.H. Hamed 2000; Flood frequency analysis U.S. CRC Press Florida 341pp.
18. Riemann, C., Filzmoser, P., and Garrett, R.G. 2005. Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, 346: 1-16.
19. Reimann, C., and Filzmoser, P. 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environ Geol.* 39: 1001-1014.
20. Rousseeuw, P.J., and Van Driessen, K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41: 212-223.
21. Shahzadi, A., Akhter, A.A., Safe, B. 2013. Regional Frequency Analysis of Annual Maximum Rainfall in Monsoon Region of Pakistan using L-moments. *Pak.j.stat.oper.res.* Vol.IX, No.1, pp111-13.
22. Solberg, H. E. and Lahti, A. 2005. Detection of outliers in reference distributions: Performance of horn's algorithm. *Clinical Chemistry* 51, 12, 2326-2332.
23. Stefansky, W. 1972. Rejecting outliers in factorial designs. *Technometrics* 14, 2, 469-479.
24. Tukey, J.W. 1977. *Exploratory data analysis*. Addison-Wesley publication, pp: 205-235.
25. Wiegand, P., Pell, R., and Comas, E. 2009. Simultaneous variable selection and outlier detection using a robust genetic algorithm. *Chemometrics and Intelligent Laboratory Systems*, 98: 2. 108-114.
26. Yusop, Z., Nasir, H. and Yusof, F., 2013. Disaggregation of daily rainfall data using Bartlett Lewis Rectangular Pulse model: a case study in central Peninsular Malaysia. *Environmental Earth Sciences*, DOI 10.1007/s12665-013-2755-7.
27. Yester, Ld. 2006. Maximum rainfall intensity analysis using I-moments in Spain. The 7th international Conference on Hydro science and Engineering. Philadelphia Drexel University.
28. Zhang, C.S., Selinus, O., and Schedin, J. 1998. Statistical analyses for heavy metal contents in till and root samples in an area of southeastern Sweden. *The Science of the Total Environment*, 2.

Citation of this article

Sajad M, Majid R, Ali G, Abazar E, Hasan E, Maryam M, Yadollah M. Determination of A Some Simple Methods for Outlier Detection in Maximum Daily Rainfall (Case Study: Baliglichay Watershed Basin – Ardebil Province – Iran). Bull. Env. Pharmacol. Life Sci., Vol 3 (3) February 2014: 110-117