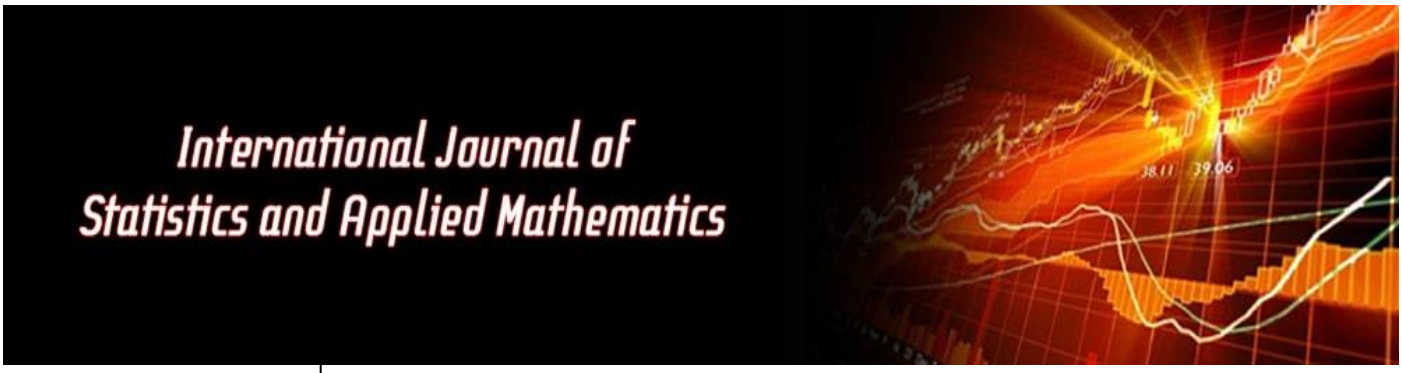
International Journal of Statistics and Applied Mathematics 2022; 7(2): 170-173



ISSN: 2456-1452
Maths 2022; 7(2): 170-173

www.mathsjournal.com
Received: 16-01-2022
Accepted: 21-02-2022

Vishakha Tiwari
Sam Higginbottom University of Agriculture, Technology, and Sciences, Prayagraj, Uttar Pradesh, India

Pratyasha Tripathi
Sam Higginbottom University of Agriculture, Technology, and Sciences, Prayagraj, Uttar Pradesh, India

Corresponding Author:
Vishakha Tiwari
Sam Higginbottom University of Agriculture, Technology, and Sciences, Prayagraj, Uttar Pradesh, India

Outlier detection in a few catchments of the Godavari river basin: A case study

Vishakha Tiwari and Pratyasha Tripathi

DOI: https://doi.org/10.22271/maths.2022.v7.i2b.813

Abstract
The efficiency of the Grubbs-Beck test and the multiple Grubbs-Beck tests to detect discordant observations are compared in this research using a few catchment areas of the Godavari river basin in Maharashtra, India. The Multiple Grubbs-Beck tests are proven to be more effective than the original Grubbs-Beck tests when applied to the Weibull distribution. When using the Grubbs test, only one lower discordant observation was identified in all of the river stations, and no lower discordant observation was found in the Somanpally river station when using the Multiple Grubbs Beck test (MGBT) five survey catchments.

Keywords: Grubb's test, discordant observation, MGB test, weibull sample

Introduction
Flood frequency analysis is a method for estimating the design flood necessary for bridges, culverts, flood embankments, and other water infrastructure, as well as for flood management and flood insurance research. Flood estimating attempts to extract as much information as can from the supplied data, while also remaining robust to the distribution model and achieving a low discordant. Furthermore, very low values can result in inaccurate estimates of huge inundation quantiles. In comparison to current levels of stream gauges with longer annual peak flow records, Griffs and Stedinger (2008) [18] discovered that flood size and frequency predictions utilizing river monitoring stations with short yearly peak flow data records have a greater standard error or uncertainty. Probability distributions are frequently chosen in flood frequency analysis based on statistical tests or graphical methods, which play a useful role in the selection of correct distributions.
Cohn *et al*, (2013) [17], $p(k;n)$ function accurately explains whether the k_{th} smallest observation in a normal sample of n variates is rare. Because inward sweep tests are more sensitive to masking, an outward sweep is preferable to avoid the problem. A flood record may contain more than one low discordant, causing the original GB test statistic to fail to distinguish the smallest observation as a discordant, causing the original GB test statistic to fail to distinguish the smallest observation as a discordant. ROSNER employs a two-sided outward sweep. The MBG test is also based on a one-sided critical value of 10% significance of a normally distributed sample, but it's designed to look at ordered sets of data, which aren't included in the other tests. The detection of lower discordant in flood data is an important stage in flood frequency analysis. Lower discordant is a very tiny observation of flood data that deviates dramatically from the patterns of the rest of the data. Lower discordant identification and treatment is a critical topic in flood frequency analysis, as such observations might have a major impact. Discordant observation is an extra variable that disturbs the whole observation in flood data series. This arises due to different causes, when the observation is different from the selected samples, measurement error or missing of observation, mostly observation is from the different population of most of the data.
The way of detecting single discordant observation is Grubb's test. It was first proposed by Frank E. Grubbs in 1950, and it was later generalized by Frank E. Grubbs in 1969 and 1972. It's also known as the maximum normed residual test or the ESD test (extreme studentized deviation). It is a statistical test for spotting outliers in a single variate of data that follows the probability distribution. For the analysis, the mean, standard deviation, and sample are employed.

Many methods discussed in Bulletin 17 B for the detection of discordant observation are Z score test, Graphical method-Box plot, Histogram, Run sequence plot, Probability plot technique, Lag plot, Quality control Stedigner test, Grubb's-beck test, and Multiple Grubb's-beck tests. The 10% significance test with a single outlier threshold was applied in Bulletin 17B. It's also recommended determining to classify one or more of the smallest observations as low outliers in order to improve the frequency analysis robustness.

The Weibull probability distribution with three parameters was used in this investigation for single or multiple discordant detection approaches. The continuous distribution is one of the three parameters of the Weibull distribution, which is mostly employed in extreme field events such as flood data, survival analysis, and so on. The sample features of quantile estimates based on the maximum likelihood (ML), Method of Moments (MOM), and probability-weighted moments (PWM) approaches were used to compute the Weibull distribution used in the regional flood frequency study (2001) [21]. The researchers conducted an empirical performance evaluation of discordant identification techniques for Weibull or extreme–value distributions (2007). The Grubb statistic G' scored well in tests for lower discordant. For the labelled slippage, Mann's W was much poorer than the others [22].

The purpose of this research is to use the Grubb's test and the multiple Grubb's-beck tests with the Weibull distribution to detect discordant observations in the five rivers stations.

Methodology

Weibull distribution

The three parameters Weibull distribution is mainly used for this analysis. The Weibull distribution was first distinguished by (Frechet Maurice 1927) [20]. This distribution is a continuous probability distribution named given by Swedish mathematician Waloddi Weibull. It can work efficiently, even with small sample sizes, and can be used precisely in frequency analysis and obtaining parameter estimates.

Let $X_1, \dots, X_n$ be independent random variables from a Weibull distribution with location, scale and shape parameters as $\mu, \sigma, and\ \gamma$ respectively.

For case I: When the observations are smaller than the location parameter μ, $e. \forall, x_i < \mu; i = 1, \dots, n$, then *pdf* is given by

Probability density function *(pdf)*

$$f(x_i; \mu, \sigma, \gamma) = \frac{\gamma}{\sigma}\left(\frac{\mu - x_i}{\sigma}\right)^{\gamma-1} e^{-\left(\frac{\mu - x_i}{\sigma}\right)^{\gamma}} \forall\ i = 1, \dots, n$$

Cumulative distribution function *(cdf)*

$$F(x_i) = e^{-\left(\frac{\mu - x_i}{\sigma}\right)^{\gamma}} \forall\ i = 1, \dots, n, \mu < x_i < \infty, -\infty \leq \mu \leq +\infty, \sigma > 0, -\infty \leq \gamma \leq \infty$$

For case II: When the observations are larger than the location parameter $\mu, i. e. \forall, x_i > \mu; i = 1, \dots, n$ then the pdf is given by Probability density function *(pdf)*

$$f(x_i; \mu, \sigma, \gamma) = \frac{\gamma}{\sigma}\left(\frac{x_i - \mu}{\sigma}\right)^{\gamma-1} e^{-\left(\frac{x_i - \mu}{\sigma}\right)^{\gamma}} \forall\ i = 1, \dots, n$$

Cumulative distribution function (CDF)

$$F(x_i) = 1 - e^{-\left(\frac{x_i - \mu}{\sigma}\right)^{\gamma}} \forall\ i = 1, \dots, n, \mu < x_i < \infty, -\infty \leq \mu \leq +\infty, \sigma > 0, -\infty \leq \gamma \leq \infty$$

Where, μ is the location parameter, σ is the scale parameter that decides the appearance or shape of the distribution and γ is the shape parameter.

Proposed test statistic

The original Grubbs-Beck tests uses the field logarithms of the peak flow data, to calculate a critical value at a significance level of 10% on one side of a normally distributed sample. Multiple peak flows recorded by the flow meter can be smaller than the critical Grubbs-Beck test, but usually only non-zero peak flows recorded in the test are identified as lower discordant. The original GB test recommended by Bulletin 17B. The original GB test identifies only one discordant from a particular dataset, but the data may have more discordant available. A method has been developed to statistically detect multiple discordant using the generalized Grubbs-Beck test. Grubbs test (1950) is a test for detecting a single discordant observation.

$$G = max\left|\frac{X_i - \bar{X}}{s}\right|$$

X_i Is the i_{th} observation and $\overline{X}$ is the sample mean, s is the sample standard deviation.

Multiple Grubb's-beck test is a generalization of Grubb's-beck test, it is used for detecting the multiple discordant observation that is based on the extremes value distribution or approximate normal distribution. In this k different discordant was detected by using any data set. Implementing the recommended MGB test for Bulletin 17C consists of two steps, firstly test each observation at the significance level α_{out} of the MGB test, starting from the median and outward toward the smallest observation. When the k-minimum observation is identified as a low outlier, the outward sweep stops and everything Observations smaller than k-smallest (that is, $j = 1, \dots, k$) are also identified as low outliers. Inward sweeps always start with the smallest observation and move towards the median. The significance level isα_{in}. If the observed value m ≥ 1 cannot be identified by the sweep, the sweep will stop. In that case, the total number of low discordant identified by the MGB test will be the maximum of k and m-1. The algorithm has two parameters that need to be specified, outward sweep significance level α_{out} for each comparison. Inward sweep significance level α_{in} for each comparison.

Table I explains the comparison of no of discordant observation found by applying the Grubbs test, and Multiple Grubbs beck test for the five river stations of Godavari basin In Multiple Grubbs beck test, Weibull distribution are used for the estimation of quantile or critical value of the data of river stations. It can be observed that multiple Grubbs-Beck is performing well as compared to that of Grubbs-beck test for detecting single discordant. Evidence can be seen for the Upper river station where multiple Grubbs-Beck test detected 20 discordant while Grubbs-beck test detected none of them.

Table 1: Comparison of single Grubbs test and multiple Grubbs Beck test in five rivers station of Godavari Basin

S. No	River Station	No of discordant observation for Grubbs test	No of discordant observation for MGB Test
1	Bhadrachalam	1	0
2	Tumnar	1	8
3	Sirpur	1	12
4	Sabari	1	15
5	Somanpally	1	12

Table 2: Estimated Quantile value by Weibull Distribution of five river station for Godavari Basin using Grubbs test and Multiple Grubbs Beck test with outlier or without outliers at 10% level of significance.

Estimated quantiles (m^3 /s) using the Weibull distribution (discordant removed by original GB test)					Percentage difference between the Weibull distribution with MGB test and the Weibull distribution with original GB test)		
S. No.	River station	With discordant for Grubbs test (1)	Removing discordant by Grubbs test (2)	Removing discordant by MGBT (3)	% difference between (1) and (2)	% difference between (3) and (2)	% difference between (1) and (3)
1	Bhadrachalam	0.8967	0.8654	0.8828	3%	2%	1%
2	Tumnar	2.0756	0.9752	0.8956	110%	-8%	118%
3	Sirpur	0.9210	0.9432	0.9282	-2%	-1%	-1%
4	Sabari	0.8765	0.8867	0.8786	-1%	-1%	0%
5	Somanpally	0.9764	0.9645	0.8654	1%	-10%	11%

In Table II, the comparison of estimated quantile values of single discordant observation and multiple discordant observation by using the single Grubbs test method and Multiple Grubbs beck test method for the five rivers station of Godavari basin are estimated. In Tumnar river station, high quantile value 2.0756 estimated with discordant by Grubbs test, but in Sirpur Godavari basin estimated the high quantile value 0.9210 with removing the discordant by Grubbs test. When using the Weibull sample in MGB test in different river discharge, it also estimated the high quantile value 0.9764 in Somanpally river station.

It also shows the flood quantiles using Weibull distribution where the lower discordant are identified and shown by the original GB test and MGB test. It was found that there are % difference between the Weibull distribution with MGB test and the Weibull distribution with original GB test between the five different river stations of Godavari Basin. It shows the variation between -8% to 11% percentage difference of flood quantiles between the methods of Grubbs test and Multiple Grubbs beck test. Table II shows the variation between the flood quantiles estimated by two methods: Weibull distribution with discordant by GB test, Weibull distribution without discordant by GB test and Weibull distribution without discordant by MGB test. It is found that for Station Tumnar, Weibull distribution overestimates flood quantile by 110%, underestimates the flood quantiles by -8%, 10%, 3%, and -1% respectively.

Result and Discussion

This paper used maximum discharge flow of rainfall in five river station of Godavari Basin. In Bhadrachalam, Tumnar, Sirpur, Sabari and Sirpur and Somanpally single discordant detected but when using the test of multiple grubbs beck test, different number of discordant detection for the different river station likewise Tumnar, Sirpur, Sabari and Sirpur and Somanpally. Quantile values with discordant, without discordant for Grubbs test and quantile value without discordant for MGBT using Weibull distribution are also estimated for the Bhadrachalam, Tumnar, Sirpur, Sabari and Sirpur and Somanpally 0.8967, 2.0756, 0.9210, 0.8765 and 0.9764 quantile value for Bhadrachalam, Tumnar, Sirpur, Sabari, and Somanpally using Grubbs test with discordant. 0.8654, 0.9752, 0.9432, 0.8867, and 0.9645 estimated the quantile value for Bhadrachalam, Tumnar, Sirpur, Sabari, and Somanpally using Grubbs test without discordant. 0.88282, 0.89566, 0.92823, 0.87865, 0.86543 quantile value for Bhadrachalam, Tumnar, Sirpur, Sabari, and Somanpally using Multiple Grubbs-beck test without discordant in Weibull distribution.

Conclusion

This paper estimates the discordant observation for the five river stations of the Godavari basin. To find out a single discordant and multiple discordant, two methods Grubbs test and Multiple Grubbs beck test was used for comparison and its quantile value ware also observed with the help of simulation technique. For all the river stations single lower outlying observation is found with the help of the Grubbs test while using the method of Multiple Grubbs Beck test (MGBT). It also concluded that the when the variation between the flood discharge is low then the Grubbs test and Multiple Grubbs Beck test not found any discordant observation.

References

1. Bobee B, Cavidas G, Cavidas Ashkar F, Ashkar J, Bernier JJ, Rasmussen P. Towards a systematic approach to comparing distributions used in flood frequency analysis. Journal of Hydrology. 1993;142(1-4):121-136.
2. Rosner B. On the detection of many outliers, Technometrics, B. 1975;17(2):221-227.
3. Saf B. Assessment of the effects of discordant sites on regional flood frequency analysis, Journal of Hydrology. 2010;380(3-4):362-375.
4. Spencer C, McCuen R. Detection of PILFs in Pearson type III data. Journals of Hydrology Engineering. 1996;1(1):2-10.
5. Cunnane C. Factors affecting choice of distribution for flood series, Hydrol Sci. J. 1985;30(1):25-36.
6. Cunnane C. Statistical distributions for flood frequency analysis, in Proc. Operational hydrological Report No. 5/33, World Meteorological Organization (WMO), Geneva, Switzerland; c1989.
7. Grubbs FE. Procedures for Detecting Outlying Observations in Samples, Technometrics. 1969;11(1):1-21.
8. Tietjen G, Moorie R. Some Grubbs-type statistics for the detection of several outliers, Technometrics. 1972;14(3):583-597.
9. Interagency Advisory Committee on Water Data (IAWCD), Guidelines for Determining Flood Flow

Frequency: Bulletin 17-B. Hydrol. Sub comm., Washington, DC; c1982.
10. Interagency Advisory Committee on Water Data (IAWCD), Robust National Flood Frequency Guidelines: What is an Outlier? Bulletin; c2013, p. 2454-2466.
11. Lamontagne JR, Stedinger Cohn JR TA, Barth NA. Robust national flood frequency guidelines: What is an outlier? In Proceeding World Environmental and Water Resources Congress, ASCE; c2013.
12. Beard LR. Statistical methods in hydrology, Civil works investigation project CW-151, U.S. Army Corps of Engineers, Sacramento, California. 1962;2:62.
13. Benson MA. Uniform flood-frequency estimating methods for federal agencies, Water Resource Research. 1968;4(5):891-908.
14. Vogel RM, McMahon TA, Chiew FHS. Flood flow frequency model selection in Australia, J Hydrol. 1993;146(421):449.
15. Merz R, Bloschil G, Humer G. National flood discharge mapping in Austria, Nat Hazards. 2008;46(1):53-72.
16. Nathan RJ, Weinmann PE. Application of at-site and regional flood frequency analyses, in Proc. International Hydrology Water Resources Symposium, Perth. 1991;3(22):769-774.
17. Cohn TA, England JF, Berenbroc CE, Mason RR, Stedinger JR, Lamontagne. A generalized Grubbs-Beck test statistic for detecting multiple potentially influential outliers in flood series, Water Resources Research. 2013;49(8):5047-5058.
18. Griffs VW, Stedinger JR. The LP3 distribution and its application in flood frequency analysis, 2. Parameter estimation methods. Journal of Hydrology Engineering. 2007;12(5):492-500.
19. Thomas WO. A uniform technique for flood frequency analysis. Journals Water Resources Planning Management. 1985;111(3):321-337.
20. Fréchet, Maurice. Sur la loi de probabilité de l'écart maximum, Annales de la Société Polonaise de Mathematique, Cracovie. 1927;6:93-116.
21. Jun-Haeng Heo, Salas JD, Boes DC. Regional flood frequency analysis based on a Weibull model: Part 2. Simulations and applications. Journal of Hydrology. 2001;242(3–4):171-182.
22. Vishakha Tiwari, Pratyasha Tripathi, Vishal Vincent Henry. Grubbs-back test and multiple Grubbs beck test compared for the Godavari basin: A case study. International Journal of Applied Research 2019;5(11):288-291.