OUTLIER TREATMENT IN THE FLOOD FLOW STATISTICAL ANALYSIS

Borislava Blagojević[1]
Vladislava Mihailović[2]
Jasna Plavšić[3]

UDK: 556.532:519.23
DOI: 10.14415/konferencijaGFS2014.081

Summary: *In the hydraulic structures design process, the reliable flood estimation is essential for designing the structure. Statistical analysis of observed flood flows, i.e. the flood frequency analysis is one way to obtain the design flood of a specified probability. The problems that can threaten the credibility of flood estimates are related, among other issues, to the treatment procedure of exceptional values (outliers) in the observed flood flow data sets. In this paper a detection of upper and lower outliers in a series of streamflow annual maxima is performed by the Grubbs and Beck statistical test on 68 hydrological stations in Serbia. The data observation period considered is from the establishment of the stations until 2012. The outlier detection results are compared to the previous results obtained for the data processing period until the year 2006.*

Keywords: *Flood flow, low outliers, high outliers, flood frequency analysis*

1. INTRODUCTION

Standardization of methods and procedures for the analysis of floods is essential for consistent and reliable assessment of design floods for sizing the hydraulic structures. Statistical analysis of the observed flow data is the basis for determining flood estimates for design projects, which in turn are the basis to estimation of design floods and the regional flood frequency analysis. Development of the "Guidelines for flood frequency analysis at hydrologic stations (in gauged basins)" for Serbia is in the final stage [1]. These guidelines propose the way to treat exceptional values (outliers) in the observed flood series of annual maximum flows, which makes the sample for frequency analysis. The proposed approach is new to the existing engineering practice in Serbia. The presence of outliers in the sample can lead to problems in formulating a statistical model and fitting an appropriate theoretical distribution from the observations. Given that finding a correct model that will allow extrapolation of the flow outside the range of

[1] Dr Borislava Blagojević, Assist. Prof., University of Niš, Faculty of Civil Engineering and Architecture, Aleksandra Medvedeva 14, Niš, Serbia, Tel: 018 588 200, e – mail: borislava.blagojevic@gaf.ni.ac.rs; b.blagojevic@eunet.rs
[2] Dr Vladislava Mihailović, University of Belgrade, Faculty of Forestry, Kneza Višeslava 1, Belgrade, Serbia, Tel: 011 30 59 945, e – mail: vladislava.mihailovic@sfb.bg.ac.rs
[3] Dr Jasna Plavšić, Assist. Prof., University of Belgrade, Faculty of Civil Engineering, P.O. Box 42, 11120 Belgrade, Serbia, Tel: 011 337 0206, e – mail: jplavsic@grf.bg.ac.rs

observed values is very important for estimation of design floods, the effect of the high outliers on the choice of the theoretical distribution is usually considered to be crucial. However, low outliers can significantly affect not only the choice of the best distribution, but also the distribution parameter estimates.
The goal of the study presented in this paper is to determine the differences that arise from different approaches to outlier detection and to identify the problems that may arise in application of the proposed procedures.

2. METHODOLOGY

Outlier detection in the annual maximum flood series in this paper is based on the Grubbs and Beck test [2]. This test examines the presence of high and low outliers in the series. It is used under the assumption that the logarithms (or some other transformation) of the original series are normally distributed. When natural logarithm of the variable is used, the upper and lower limits for outliers are given with:

$$X_D = \exp(m_y - K_n \cdot s_y) \text{ i } X_G = \exp(m_y + K_n \cdot s_y)\,. \tag{1}$$

where m_y and s_y are the mean and standard deviation of the natural logarithms of the original variable, and K_n is the frequency factor representing the test statistic that depends on the significance level α and the sample size n. For the 10% significance level and sample size n, approximate expression for K_n is given with [3]:

$$K_n = -0.9043 + 3.345\sqrt{\log n} - 0.4046 \log n\,, \tag{2}$$

The above equation gives approximately the same results as the expression [4]:

$$K_n = -3.6220 + 6.2844 n^{0.25} - 2.49835 n^{0.5} + 0.491436 n^{0.75} - 0.037911 n\,. \tag{3}$$

For the 5% significance level, the following formula has been obtained from information in [5]:

$$K_n = -0.5148 + 3.19\sqrt{\log n} - 0.3837 \log n\,. \tag{4}$$

For flood flow series and their logarithms for which distributions other than normal are assumed, the limits that define outliers are estimated from the given distribution for normal probability of K_n:

$$X_D = F_X^{-1}(1-p), \quad X_G = F_X^{-1}(p), \quad p = \Phi(K_n)\,. \tag{5}$$

where Φ(.) is the cumulative distribution function of the standard normal distribution, and F_X(.) is the assumed parent distribution for the sample data.
The sample data is tested for normality by testing the hypothesis that the skew coefficient is equal to zero. The critical region for this test is defined using the Fisher's asymptotic variance of the sample skewness [6]:

$$\text{var}[c_s] = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}\,. \tag{6}$$

The annual maxima series for which the skewness of the logarithmic flows was not in the critical test region ($|c_s| < Z_\alpha \text{var}[c_s]^{0.5}$, where Z_α is the standard normal variate for the given significance level α) were considered to be log-normally distributed (and denoted LN2). The series for which the normality hypothesis was rejected were fitted with other skewed distributions including Pearson type III (PT3), log-Pearson type III (LPT3) and the general extreme value distribution (GEV). The best distribution was selected on the basis of the goodness-of-fit tests and visual checks. Where it was possible, the same distribution type was adopted for the sites from the same large basin.

An iterative procedure for outlier detection was applied. For positively skewed samples the largest observation X_n is first compared to the upper limit X_G, and for the negatively skewed samples the smallest observation X_1 is compared to the lower limit X_D. If X_n or X_1 turns out to be an outlier, it is removed from the sample. The sample statistics are then recalculated, and the detection procedure starts from the beginning. Such an approach is especially important for samples with two similar extreme values that are not readily detected as outliers. According to [5], in the samples in which the second extreme value was also an outlier, both values were treated as outliers; in the opposite case neither value was treated as an outlier.

The most common recommendation for further frequency analysis of the samples featuring low outliers is to remove them from the sample (the censoring approach) [4]. A quantile x_p is estimated from a censored sample by correcting the new distribution function $F_1(x)$ because of removing n_d values. This is done by conditioning $F_1(x)$ with the reduced sample size:

$$F(x) = P\{X \le x\} = P\{X \le x \mid X > X_D\}P\{X > X_D\} + \\ + P\{X \le x \mid X < X_D\}P\{X < X_D\} = F_1(x)\frac{n-n_d}{n} + 1 \cdot \frac{n_d}{n}. \quad (7)$$

A high outlier can be treated using different approaches according to [7]: a) the outlier is removed from the sample, b) the outlier is replaced with the second largest flow in the same year in which the outlier has occurred, c) the outlier is replaced with the second largest value in the sample, and d) the outlier remains in the sample. Another possible approach is to replace the outlier with the flood of the 50 or 100 year return period from a neighbouring station or region. As a rough estimate, an outlier can be assigned an apparent return period $T^* > n$, estimated as the return period of the same flood event at neighbouring stations or by using information on historical floods [8]. Such approaches are currently not applicable for Serbia, because the results of a regional flood frequency analysis for the considered period are not available.

3. RESULTS AND DISCUSSION

In this paper, the presence of outliers in the annual maximum flow series at 68 hydrological stations on the territory of Serbia is investigated. The observation record at these stations spans from the establishment of the stations until the year 2012 (sample size ranges from 39 to 85 years). Testing for normality of the log-transformed series showed that the LN2 distribution can be applied to 44 out of 68 stations. The PT3

distribution is applied to two stations, the LPT3 distribution to six stations and the GEV distribution to 16 stations. It should be noted that stations on the Drina River are not considered since the well-known historical maximum in 1896 requires special attention and application of the methods for frequency estimation of historical extremes from pre-systematic observations. At the 10% significance level, low outliers were identified at 8 stations and high outliers at 3 stations, while one station (the Studenica River at Ušće) had both a high and a low outlier. At the 5% significance level, almost all of the outliers detected at the 10% significance level were detected again, except for one low outlier (the Ibar River at Lopatnica Lakat) and one high outlier (the Studenica River at Ušće). The results of testing for presence of outliers (equations 2 and 4) are shown symbolically in Figures 1 and 2 along with the results published in [9], where the annual maxima series with data until 2006 were tested at the 10% significance level. It should be noted that the approach used in [9] is somewhat different from this study in that only the series with $|c_s| > 0.4$ were tested for presence of outliers, that only the LPT3 distribution was considered, and that the limits for outlier detection were not determined from equation (5) but rather with an assumption of a normal parent distribution (eq. 1). In addition, the results of an evaluation of the detected outliers' return period that was carried out in [9] is also shown symbolically in Figures 1 and 2 with circles of different sizes.

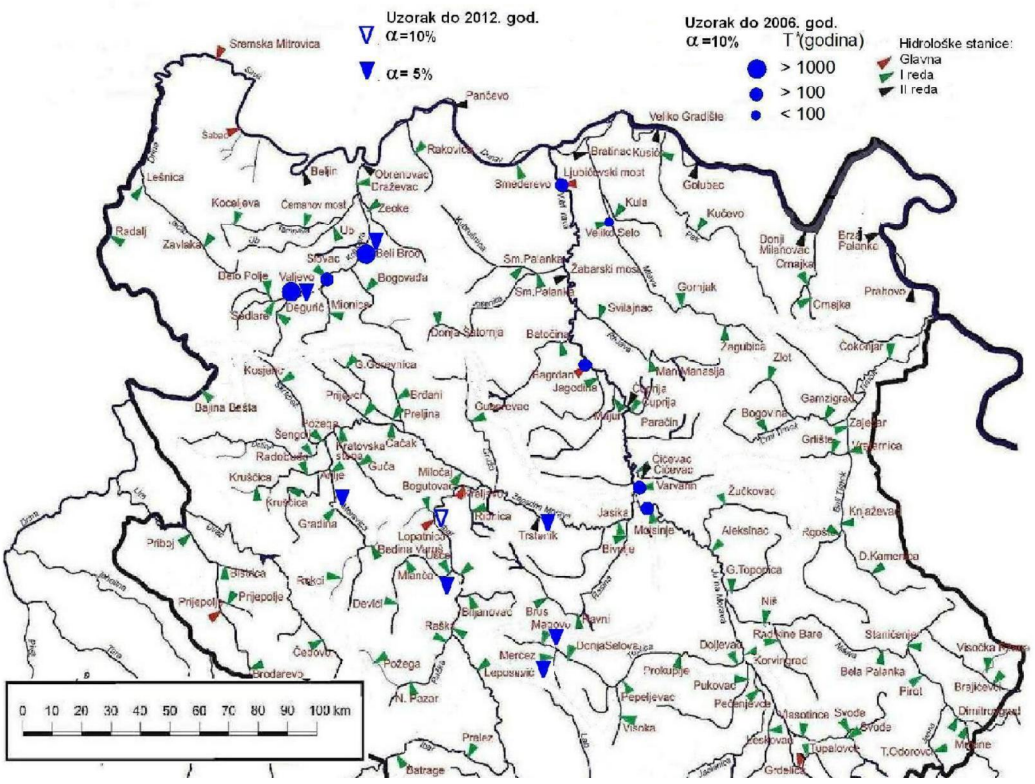

Figure 1. Low outliers detected in the series of annual maxima flows: triangles – the results from this study, circles – the results from [9].

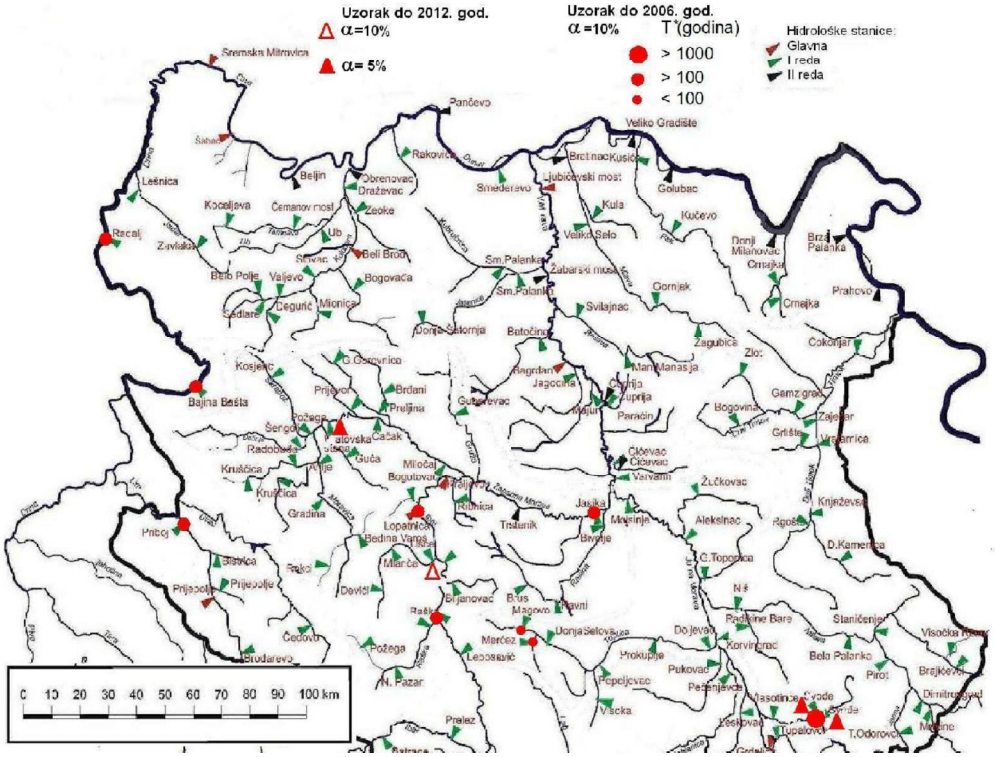

Figure 2. High outliers detected in the series of annual maxima flows: triangles – the results from this study, circles – the results from [9].

In comparison to the results from [9], new outliers are detected in data after 2006 (low outliers in the Toplica River basin), while the high outliers in the Vlasina River basin in 1988 and the low outliers in the Kolubara River catchment that have exceptionally small probability of occurrence are detected again. The differences in detected outliers between the results from this study and the results from [9] can be attributed to different samples and to the fact that the study [9] did not consider possible limitations for the application of the Grubbs-Beck test in regard to the assumption on the log-normal parent distribution.

4. CONCLUSIONS

The main conclusion drawn in this study is that the assumption about the type of the annual maximum flows' parent distribution is extremely important for outlier detection, because transformation (5) yields different limits for the outlier detection. The study has shown that the outliers detected under assumption of the log-normal distribution in cases when this assumption was not supported by the sample properties (because the skew coefficient of the log-transformed data was significantly different from zero) were

generally not detected under assumption of a theoretical distribution identified as the best fit for the given empirical distribution.
The example of the Studentica at Ušće station, for which both high and low outliers were detected, has shown that the order of the outlier detection (high/low or low/high) has an important role. In such cases, the series asymmetry is a useful indicator to select the order of the detection.
Finally, the procedure for outlier detection is certainly of great importance for the subsequent estimation of quantiles from the selected distributions and hence the design flood flows, which is the subject of the ongoing investigation in this study.

ACKNOLEDGEMENTS

"The research on development and improvement of the protection from floods in Serbia: Development of the methodology for standardization of flood assessment procedures in Serbia – Phase One: Guidelines for frequency analysis of floods at hydrologic stations (gauged basins)" has been funded by the Public Water Management Company "Srbijavode". We are grateful to the Republic Hydrometeorological Service of Serbia for providing the data for this study.

REFERENCES

[1] Blagojević, B., Mihailović, V., Plavšić, J.: New Guidelines for Flood Flow Assessment at Hydrologic Stations in Serbia. *Electronic Proceedings of the International Conference on Flood Resilience: Experiences in Asia and Europe*, 5-7 September 2013, Exeter, United Kingdom. Djordjević, S., Butler, D., Chen, A. (Eds.). **2013**.
[2] Bobée, B., Ashkar, F.: *The Gamma Family and Derived Distributions Applied in Hydrology*. Water Resources Publications, Littleton, Colorado, U.S.A., **1991**.
[3] Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E.: Chapter 18: Frequency Analysis of Extreme Events. *Handbook of Applied Hydrology*. Mc-Graw Hill Book Co., New York, **1993**.
[4] ICWD: *Guidelines for determining flood flow frequency: Bulletin 17B* (revised and corrected), Interagency Committee on Water Data, Hydrol. Subcomm., Washington, D.C., **1982**.
[5] NRCS: *National Engineering Handbook*, Part 630: Hydrology, Chapter 18: Selected Statistical Methods, National Resources Conservation Service, USDA, **2012**.
[6] Kottegoda, N., Rosso R.: *Applied Statistics for Civil and Environmental Engineers*, Blackwell, **2008**.
[7] Vukmirović, V., Pavlović, D.: *Utvrđivanje kriterijuma za izbor merodavnih velikih voda*. Tema 2 u okviru naučno-istraživačkog projekta TSI 114 „Savremene metode u hidrotehnici", (in Serbian), Faculty of Civil Engineering, Iniversity of Belgrade, **2000**.

[8] BLFUW: *Leitfaden Verfahren zur Abschätzung von Hochwasserkennwerten*, P. Lorenz (Leitung und Koordination), Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, Wien, **2011**.
[9] Blagojević, B., Ilić, A., Prohaska, S. : Interrelation of Droughts and Floods through Outlier Detection on Rivers in Serbia. *Proceedings of the international conference BALWOIS 2010* , Ohrid, Vol. II. **2010**.

ПОСТУПАЊЕ СА ИЗУЗЕТНИМ ВРЕДНОСТИМА У СТАТИСТИЧКОЈ АНАЛИЗИ ВЕЛИКИХ ВОДА

Резиме: *У процесу пројектовања хидротехничких објеката, поуздана процена меродавних великих вода је од суштинског значаја за димензионисање објеката. Статистичка анализа низова осмотрених протока је један од начина за оцену меродавних великих вода. Проблеми који могу да угрозе веродостојност оцене великих вода повезани су, између осталог, и са начином на који се поступа са изузетним вредностима (изузецима) у осмотреним низовима података. У овом раду спроведена је детекција горњих и доњих изузетака у низовима годишњих максимума статистичким тестом Grubsa i Beka за 68 хидролошких станица у Србији у периоду од оснивања станица до 2012. године. Добијени резултати упоређују се са објављеним резултатима добијеним за период до 2006. године, што је показало да на поступак детекције кључни утицај има претпоставка о расподели коју прате разматрани низови.*

Кључне речи: *Велике воде, доњи изузеци, горњи изузеци, статистичка анализа*