VERSITA

Mathematical Modelling in Civil Engineering
Vol. 9-No. 3-2013
Doi: 10.2478/mmce-2013-0008

ESTIMATION OF THE STATISTICAL DISTRIBUTION USED IN HYDROLOGY USING KERNEL FUNCTIONS

Romică TRANDAFIR - professor, Technical University of Civil Engineering Bucharest, Mathematics and Computer Science Department, e-mail: romica@utcb.ro
Daniel CIUIU - assistant professor, Technical University of Civil Engineering Bucharest, Mathematics and Computer Science Department; associate researcher, Institute for Economic Forecasting, Bucharest, e-mail: dciuiu@yahoo.com
Radu DROBOT - professor, Technical University of Civil Engineering Bucharest, Hydrology Department, e-mail: drobot@utcb.ro

Abstract: In this paper we will study the statistical distributions for the extreme discharges of the Danube River using kernel functions. We will also compare the results with those obtained using classical cumulative distribution functions (Pareto, Weibull, etc).

Keywords: kernels, discharges, cumulative distribution functions, confidence intervals, quantiles.

1. Introduction

In hydrological processing of maximum discharges some parametric statistical distributions are used, from which the most common are: Pearson III, log-normal, GEV (Generealized Extreme Values), GPD (Generalized Pareto Distribution) etc. Based on the processing performed to obtain maximum flow rates with different probabilities of exceedance the synthetic flood waves that are used in sizing or checking waterworks. In this way the use of maximum discharge is presented in order to determine the rate of the crest dam, the size of large waters' spillways or flood extension corresponding to standard exceedance probabilities (10%, 1%, 0.1% or 0.01%). The delimitation of floodrisk zones has the goal to prevent the damages produced due to flooding.

The main criticism of this approach is the fact that, because we have low volume samples, the appropriate probability density function is unknown. To circumvent this difficulty, in this paper we propose the use of kernel functions for estimating the empirical pdfs.

Using these kernel functions we will estimate the quantiles and the confidence intervals for the quantiles, starting from a maximum annual discharges sample of the Danube in the region of Budapest.

Definition 1 [8,10]. *It is called kernel function a pdf that allows the estimation of another unknown pdf starting from a given sample.*

If we have a sample of volume n, $X_1,...,X_n$, we estimate the pdf $f(x)$, for an arbitrary x by

$$f(x)=\frac{1}{n\cdot b_n}\cdot\sum_{i=1}^{n}K\left(\frac{x-X_i}{b_n}\right), \tag{1}$$

where K is the kernel function, and b_n is the bandwidth.

There exist several types of kernel functions $K=K_j$ used in literature.

$$\begin{cases} K_0(x) = \chi_{-\frac{1}{2},\frac{1}{2}}(x) = \begin{cases} 1\ if\ |x| \le \frac{1}{2} \\ 0\ if\ |x| > \frac{1}{2} \end{cases} \\ K_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ K_2(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\chi_{-\sqrt{5},\sqrt{5}}(x) \\ K_3(x) = \left(1 - |x|\right)\chi_{-1,1}(x) \\ K_4(x) = \frac{3}{4}\left(1 - x^2\right)\chi_{-1,1}(x) \\ K_5(x) = \frac{15}{16}\left(1 - x^2\right)^2 \chi_{-1,1}(x) \end{cases}, \tag{2}$$

where K_0 is the rectangular kernel, K_1 is the Gaussian kernel, K_2 is the Ephanetchnikoff kernel, K_3 is the triangular kernel, K_4 is the Bartlett-Priestley-Ephanetchnikoff kernel and K_5 is the bi-quadratic kernel.

The Ephanetchnikoff kernel K_2 minimizes the integrated square error (*MISE*=Min Integrated Square Error).

$$MISE\left(\hat{h}\right) = \int_{-\infty}^{\infty} MSE_x\left(\hat{h}\right)dx, \tag{3}$$

where $MSE_x\left(\hat{h}\right)$ is the minimum square error (*MSE*=Min Square Error; see [8,10]):

$$MSE_x\left(\hat{h}\right) = E\left\{\left(\hat{h}(x) - h(x)\right)^2\right\}. \tag{3'}$$

The bandwidth must be chosen such that $\lim_{n\to\infty} b_n = 0$ and $\lim_{n\to\infty} nb_n = \infty$ [8,10]. In this paper we consider $b_n = \frac{1}{\sqrt{n}}$ and we obtain:

$$f(x) = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^{n} K\left(\sqrt{n}\left(x - X_i\right)\right). \tag{1'}$$

Therefore we can compute the cumulative density function using the formula

$$F(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \tilde{K}\left(\sqrt{n}\left(x - X_i\right)\right), \tag{4}$$

where $\tilde{K}$ is the corresponding cdf of the kernel pdf *K*.

The confidence interval with threshold ε of a cdf *F* is built using the following procedure [1]:

1) We determine fist the confidence interval (st, dr) such that $G(st) = \frac{\varepsilon}{2}$ and $G(dr) = 1 - \frac{\varepsilon}{2}$, where *G* is the cdf Beta, having the parameters $m+1$ and $n-m$, $\beta(m+1, n-m)$.
2) The confidence interval becomes $\left(F^{-1}(st),\ F^{-1}(dr)\right)$.

The above parameter *m* is $\left[n(1-\varepsilon)\right]$, i.e. the integer part of $n(1-\varepsilon)$, and *n* is the volume of the sample [1].

2. Methodology

In order to determine the quantiles of various orders stored into array, we generate 10000 variables of type kernel by means of the compound method [9]. To generate one of these variables we use the equation (4) and we generate a random number k, uniformly distributed on $\{1, 2,..., n\}$ and then we generate X with cumulative distribution function kernel translated with X_k and scaled according to the bandwidth window $\sqrt{n}$.

For the Bartlett-Priestley-Epanecinikov type kernel or biquadratic type kernel we generate a random variable with probability density function K using the inverse method [9]. For the Epanecinikov kernel type one should multiply the result of the case Bartlett-Priestley-Epanecinikov with $\sqrt{5}$.

In [2] one simulation was performed for each quantile threshold for kernel-type distributions. But because there is the possibility to have a reversal of monotony through successive simulations (for different thresholds, but close), we made one simulation for all quantiles in the list. For the Gaussian kernel we can use different methods for generating normal variables [9]. We use the Box-Müller method because it is the most rapid. After we sort ascending the values, then the ε-quantile is on the position $[10000(1-\varepsilon)]$,where $[10000(1-\varepsilon)]$ is the integer part of $10000(1-\varepsilon)$.

The quantiles level of trust given by the start and end is determined according to the above method.

Using the kernels mentioned in the previous section we can not capture the phenomena of asymmetry, or of heavy tail. For instance, in the case of the last kernel, the difference between the quantile is practically negligible. In order to avoid this drawback, we use the Pareto kernel, ie Pareto distribution density (a, b, c), c = 0, [6]

$$f(x)=\begin{cases}\dfrac{1}{b}\left(1-\dfrac{a(x-c)}{b}\right)^{\frac{1}{a}-1}, & for\ \ a\neq 0\\ \dfrac{1}{b}e^{-\frac{x-c}{b}}, & for\ \ a=0\end{cases} \tag{5}$$

The calibration of parameters *a* and *b* was done as follows. First we generated 1001 pairs (a,b) with values in the mentioned intervals. For each pair we computed the coefficient of determination [3,4]

$$R^2 = 1-\frac{\sum_{i=1}^{n}\left(Q^m_{p_i^e}-Q^{calc}_{p_i^e}\right)^2}{\sum_{i=1}^{n}\left(Q^m_{p_i^e}-\overline{Q_{p_i^e}}\right)^2}, \tag{6}$$

where $Q^m_{p_i^e}$ are the sample values having the corresponding non-exceedance probabilities p_i^e, $Q^{calc}_{p_i^e}$ are the quantiles of the level p_i^e (computed using the previously presented methodology), and $\overline{Q_{p_i^e}}$ is the average of $Q^m_{p_i^e}$.

After ordering the sample values, p_i^e is computed using (7) for some types of empirical distributions used in hydrology.

$$p_i^e = \begin{cases} \dfrac{2i-1}{2n} \text{ for the Hazen type} \\ \dfrac{i}{n+1} \text{ for the Weibull type} \\ \dfrac{i-0.3}{n+0.4} \text{ for the Cegodaev type} \\ \dfrac{i-0.375}{n+0.25} \text{ for the Blum type} \\ \dfrac{3i-1}{3n+1} \text{ for the Tukey type} \\ \dfrac{i-0.44}{n+0.12} \text{ for the Gringorten type} \end{cases} \quad (7)$$

The parameters *a* and *b* are chosen such that the coefficient of determination R^2 is maximum.

3. Application

Consider 85 maximum annual discharges recorded on the Danube in the region Budapest. For the Pareto kernels we estimate the maximal pair (a,b) such that *a* belongs to successive intervals $(-5,-0.25)$; $(-0.25, 0)$, and *b* belongs to successive intervals $(0, 100)$; $(100, 500)$; $(500, 1000)$; $(1000, 5000)$; $(5000, 10000)$. We obtain some results for which $R^2 > 0.95$. Table 1 contains the results of the calibration of parameters for the Pareto kernel, after processing the data for the types of the above empirical distributions.

Table 1

The calibration of the parameters for the Pareto kernels

Type	Interval for *a*	Interval for *b*	*a*	*b*	R^2
Weibull	(-5,25)	(0,100)	-1.04759	12.73232	0.99996
Weibull	(-5,25)	(100,500)	-0.41888	122.05878	0.99949
Weibull	(-5,25)	(1000,5000)	-0.30784	1098.63582	0.95799
Hazen	(-5,25)	(100,500)	-0.38815	100.79348	0.99969
Hazen	(-5,25)	(500,1000)	-0.36148	546.38813	0.98832
Hazen	(-5,25)	(1000,5000)	-0.30799	1017.82281	0.96174
Cegodaev	(-5,25)	(0,100)	-0.35263	1.37638	>0.99999
Cegodaev	(-5,25)	(1000,5000)	-0.31146	1011.96326	0.95606
Blum	**(-5,25)**	**(1000,5000)**	**-0.3535**	1103.64086	**0.9248**
Tuckey	(-5,25)	(1000,5000)	-0.368	1051.39317	0.95275
Gringorten	(-5,25)	(1000,5000)	-0.30001	1077.51701	0.96077

Even if we obtain good results for R^2 in the case $a \in (-0.25,0)$, we consider for kernels the other case, in order to capture the "heavy tail" phenomenon.

We estimate the quantiles and the confidence interval for $p = 10\%$, $p = 1\%$, and $p = 0.01\%$. In order to take into account the estimator of *b*, we obtain using the method of the moments [7]

b=1243.99617, and in Table 2 we consider only the cases $a \in (-5, -0.25)$ and $b \in (1000, 5000)$. By the values of *a* we capture the "heavy tail" phenomenon, and the above estimated value of *b* belongs to the interval.

Table 2

Quantiles and confidence intervals for thr Pareto kernels

Pareto parameters	The threshold of the quantile	The value of the quantile	The interval for the quantile
(-0.30784,1098.63582)	10%	7433	(7199,7798)
	1%	8716	(7700,11176)
	0.1%	9816	(7505,11176)
	0.01%	11176	(7402,11176)
(-0.30799,1017.82281)	10%	7417	(7149,7764)
	1%	8686	(7704,14475)
	0.1%	9340	(7490,14475)
	0.01%	14475	(7387,14475)
(-0.31146,1011.96326)	10%	7462	(7235,7897)
	1%	8712	(7826,14061)
	0.1%	9510	(7558,14061)
	0.01%	14061	(7416,14061)
(-0.3535,1103.64086)	10%	7470	(7255,7908)
	1%	8748	(7863,21435)
	0.1%	10507	(7563,21435)
	0.01%	21435	(7421,21435)
(-0.368,1051.39317)	10%	7461	(7227,7898)
	1%	8748	(7870,20245)
	0.1%	10042	(7570,20245)
	0.01%	20245	(7418,20245)
(-0.30001,1077.51701)	10%	7429	(7163,7844)
	1%	8679	(7743,14626)
	0.1%	9974	(7497,14626)
	0.01%	14626	(7392,14626)

We notice that we have a decreasing of the quantile with respect of the threshold, as we expected. We can say the same thing about the limits of the confidence intervals β for the above quantile, until the threshold of 1%. For lower threshold the lower limit increases on threshold,

and the variance of the right limit remains the same, but the decrease becomes very slow. The next graphics (in the case $a = -0.30784$, $b = 1098.63582$) is suggestive in this regard.

An explanation of changing of variance for the left limit of the interval consists in the fact that the parameters of the distribution Beta stabilize at $a = 85$ and $b = 1$. From the fact that we determine the quantiles for the same distributions, the variance of the limits of the confidence interval results.

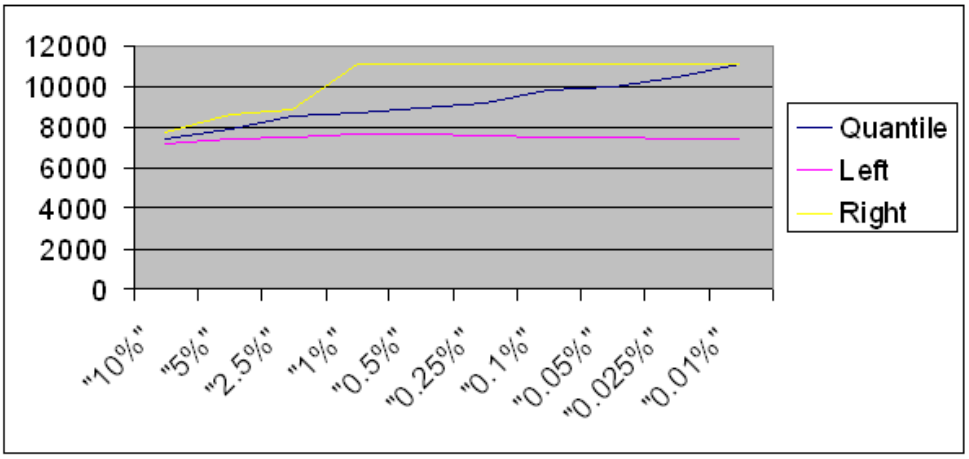

Fig. 1 – The graphics of the quantile and of the limits of confidence intervals, depending on the threshold.

For the right limit of the confidence interval for the quantile we notice a convergence of the value to the value of the quantile with the threshold of 0.01% (1 in 10000). This is due to the fact that we have generated 10000 kernel-type random variables, and the right limits for the Beta distribution are higher than 0.9999. That is why all the values are greater than the maximum generated value.

Comparison of the Values Obtained by Classical Analysis

By classical statistical inference we have obtained the following confidence intervals (Table 3)

Table 3

Confidence intervals using classical statistical inference, and the volume of the sample 85

The exceedance probability	The value of the quantile	Confidence interval for the quantile
10%	7294	(6916, 7667)
1%	8553	(7961, 9390)
0.1%	9270	(7961, 9390)

Comparing the values of the quantiles and the confidence intervals from Table 3 with the similar values in Table 2, we notice that by using the Pareto kernel functions we capture better the heavy tail behavior. For the remaining probability values, both quantiles' values and confidence intervals are very close in terms of specialists in hydrology.

4. Conclusions

Since, on the one hand, the values of the quantiles and confidence intervals are very close to the probabilities of exceedance values covered by the current hydrological analysis, and secondly that it eliminates the need for a priori choice of a distribution function based on an extremely

reduced statistical selection. One can conclude that the approach using the kernel functions is entirely appropriate to this type of hydrological statistical analyzes.

References

[1] Bâ, K., Diaz-Delgado, C., Cârsteanu, A. (2001), Confidence Intervals of Quantiles in Hidrology Computed by an Analytical Method, *Natural Hazards*, 24, 1-12.

[2] Ciuiu, D. (2009), Numerical and Monte Carlo Methods to Make Normal Residues in Regression, *Romanian Journal for Economic Forecasting*, 12(4), 119-131.

[3] Nash, J.E., Sutcliffe, J.V. (1970), River flow forecasting through conceptual models part I — A discussion of principles, *Journal of Hydrology*, 10(3), 282–290.

[4] Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L. (2007), Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, *Transactions of the ASABE*, 50(3), 885–900.

[5] Saporta, G. (1990), *Probabilités, Analyse des Donées et Statistique*, Editions Technip, Paris.

[6] Singh, V. P., Guo, H. (1995), Parameter estimationn for 3-parameter generalized Pareto distribution by the principle of maximum entropy, Hydrological Sciences, *Journal des Sciences Hydrologiques*, 40(2), 165-181.

[7] Trandafir, R., Ciuiu, D., Drobot, R. (2011), The Utilization of Copula in Hydrology, *Scientific Journal Mathematical Modeling in Civil Engineering*, 7(2 BIS), 12-19.

[8] Văduva, I., Pascu, M. (2003), Nonparametric Estimate of the Hazard Rate: A Survey, *Revue Roumaine Math. Pures Appl.*, 58(2), 173-191.

[9] Văduva, I. (2004), *Modele de simulare*, Ed. Universității Bucureşti.

[10] Văduva, I. (1968), Contribuții la teoria estimațiilor statistice ale densităților de repartiție şi aplicații, *Studii şi Cercetări Matematice*, 8, 1027-1076.