Journal of Computational Mathematics, Vol.12, No.4, 1994, 330–338.

THE TRANSFORMED NONPARAMETRIC FLOOD FREQUENCY ANALYSIS *

Kaz Adamowski
(*Department of Civil Engineering, University of Ottawa, Ottawa, Canada*)
Wojciech Feluch
(*Institute of Environmental Engineering, Technical University of Warsaw, Warsaw, Poland*)

Abstract

The nonparametric kernel estimation of probability density function (PDF) provides a uniform and accurate estimate of flood frequency-magnitude relationship. However, the kernel estimate has the disadvantage that the smoothing factor h is estimate empifically and is not locally adjusted, thus possibly resulting in deterioration of density estimate when PDF is not smooth and is heavy-tailed. Such a problem can be alleviate by estimating the density of a transformed random variable, and then taking the inverse transform. A new and efficient circular transform is proposed and investigated in this paper.

1. Introduction

Flood magnitude and the corresponding frequency can be estimated from the available date sample by the parametric method whereby various theoretical distributions (i.e., Log-Pearson Type III) are employed. During the past several years parametric modeling has been a subject of intensive investigations by many researchers (Singh, 1986). It is now well recognized that the main problems of parametric procedures are due to the presence of asymmetrical and multimodal densities in the observed flood data. The data also might be of such a type that there is no suitable parametric family that gives a good fit (i.e., the separation effect, Beran et al., 1986) and subsequently will lead to erroneous conclusions.

Many parametric distributions have been recommended for use in hydrology. However, there is no general consensus among hydrologists as to the "best"theoretical frequency distribution for use in flood frequency analysis (Wallis et al., 1985). In order to obtain some degree of uniformity when performing flood analysis (Thomas, 1985), several countries imposed a choice of the procedure (i.e. Log-Peason Type III in A, Generalized Extreme Value in U.K.). In other countries the choice of a distribution

* Received January 14, 1987.

is suggested to be limited to several methods (i.e. Log-Normal Type III, Log-Pearson Type III, Generalized Extreme Value, Wakeby and Weibull Distribution) as is the situation in Canada (Pilon et al., 1985). It is then up to the designer to make the decision as to which method is the most appropriate for a given circumstance. Both of the above administrative recommendations (i.e., imposition of a bass method, or limiting the choice) illustrate a need for a new method which would be uniform, give accurate results, and be suitable for asymmetrical and multimodal distributions.

Under such circumstances, a new nonparametric method was investigated by Adamowski (1985). The nonparametric density estimation does not require assumption of any functional form of density. In fact, very little is assumed, and the assumptions made are mild (Rao, 1984). Adamowski (1985) compared the performance of several parametric and nonparametric estimators and concluded that the nonparametric method is accurate, uniform, and particularly suitable for multimodal data.

The nonparametric method requires a selection of a kernel function $K(.)$, and a smoothing factor h. The choice of a kernel has little effect on the efficiency of the method. Nevertheless, there exists an optimal kernel of Epanechnikov which is in the form of a circular function (Rao, 1983, p.66).

However, the choice of a smoothing factor h plays a crucial role because it affects the bias and variance of the estimator. The optimal choice of a smoothing factor depends on the unknown a priori density and the derivatives of that density. In practical situations since density is unknown, therefore h has to be estimated empirically by various methods (Devroye et al, 1985, p. 191, Adamowski, 1985).

The potential of the nonparametric density estimation is not fully realized in hydrology primarily because of the following two difficulties : a) the value of the smoothing factor h is constant and empirically derived, and b) the nonparametric method places small probability value in the tails of a distribution (thus the extrapolation for return periods exceeding the record length might be influenced too much by the highest observation in the sample).

When the value of h is constant and is not locally adjusted, then the performance of the kernel estimate might deteriorate especially for a density which is not smooth and heavy tailed (skewed). Such problems can be alleviate by the following two modifications, namely a) using the transformation, and b) employing a variable kernel method (Breiman, et al., 1977).

The problem of placing low probability values in the tails of a distribution can be resolved by the introduction of a mixture of parametric and nonparametric methods (Schuster and Yakowitz, 1985).

The purpose of this paper is to introduce the transformed kernel method for estimation of flood frequency and magnitude relationship.

2. The Transformed Kernel Estimate

For a given kernel function $K(.)$, a positive smoothing factor h_*, and a random sample of observations $x_1, x_2, ..., x_n$, the kernel estimate of probability density function is given by (Adamowski, 1985)

$$f_n(x) = (nh_*)^{-1} \sum_{i=1}^{n} K(\frac{x - x_i}{h_*}) \tag{1}$$

where h_* is obtained from observed data.

Based upon a transformation $Y_i = T(X_i)$ the transformed density estimation is (Devroye et al, 1985, page 244)

$$f_n(x) = \tilde{g}_n[T(x)]T'(x) \tag{2}$$

where the transformed data sequence is $y_1, y_2, ..., y_n$, and the density of Y_i sequence is estimated by

$$g_n(y) = (nh)^{-1} \sum_{i=1}^{n} K(\frac{y - y_i}{h}) \tag{3}$$

where h is obtained from transformed data. Since $g_n(y)$ might not be a density on $[0, 1]$ because some portions of $g_n(y)$ can extend beyond 1 or 0, therefore the following "normalization" is introduced

$$\tilde{g}_n(y) = g_n(y) / \int_0^1 g_n(y)dy \tag{4}$$

where f is the density of $x_1, ..., x_n$ (the data), $\tilde{g}$ is the density of $y_i = T(x_i)$ given $x_1, ..., x_n$ and transformation $T(x)$.

The transformation $T(x)$ should be chosen in such a way as to obtain the best rates of consistency, that is

$$\int |f_n - f| = \int |g_n - g| = J_n \to 0, \quad as \quad n \to \infty \tag{5}$$

and

$$E(\int |f_n - f|) = E(J_n) \tag{6}$$

where $E(J_n)$ is given by

$$E(J_n) \approx A(K) \cdot B(g) n^{-2/5} \tag{7}$$

where

$$A(K) = (\int K^2)^{2/5} (\int x^2 K)^{1/5} \tag{8}$$

and

$$B(g) = (\frac{1}{2}(\int \sqrt{g})^4 \int |g''|)^{1/5}. \tag{9}$$

The quantity B(g) has a component $\int \sqrt{g}$ that measures how heavy the tail of g is , and a component $\int g''$ that measures how oscillatory g is. In view of eq. 9, the following is assumed: g is absolutely continuous, bounded, and twice differentiable.

Therefore, these two components that determine the efficiency of the kernel estimate, i.e., discontinuities or sharp oscillations, and large tails, can be measured by quantity $B(g)$. The $\int \sqrt{g}$ is small when g has a small tail and vice versa.

With the transformed kernel estimate, the value of h can be derived analytically for a given kernel and transform, and then the factor $B(g)$ is used as a measure of efficiency of density estimation.

When $g_n(y)$ is a kernel estimate then the convergence limit is proportional to $B(g)$, thus the transformation should be chosen so as to minimize $B(g)$. Devroye et al (1985) has shown that this minimum is attained for the isosceles triangular density on $[0,1]$. If the distribution function F of f were known, then this triangular optimal transformation could be

$$T(x) = \begin{cases} \sqrt{F(x)/2} & \text{for } F(x) \leq \frac{1}{2} \\ 1 - \sqrt{(1-F(x))/2} & \text{for } F(x) \geq \frac{1}{2} \end{cases} \tag{10}$$

and

$$T'(x) = \begin{cases} f(x)/\sqrt{8F(x)} & \text{for } F(x) \leq \frac{1}{2} \\ f(x)/\sqrt{8(1-F(x)} & \text{for } F(x) \geq \frac{1}{2} \end{cases} . \tag{11}$$

The corresponding smoothing factor h is (Devroye et al, 1985,p.106)

$$h = \{\frac{8}{\pi}(\frac{\int \sqrt{f}}{\int f'})^2 \frac{1}{n}\}^{1/3}. \tag{12}$$

However, the density function f is not known, but it can be estimated nonparametrically by

$$f(x) = (nh)^{-1} \sum_{i=1}^{n} K(\frac{x - x_i}{h_*}). \tag{13}$$

The corresponding distribution function is given by

$$F(x) = 1 - p(x) \tag{14}$$

where $p(x)$ is exceedance probability given by

$$p(x) = \int_x^{\infty} f(x)dx = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^{n} W(\frac{x - x_i}{h_*}) \tag{15}$$

where $W(.)$ is an integral of a kernel function $k(.)$.

2.1. The triangular transformed density estimate

Devroye et al (1985, p.110) presents several different transforms (i.e., Uniform on $[0,1]$), isosceles triangular on $[0,1]$, Normal $(0,1)$, Laplace $(exp(-x)/2)$, exponential $(exp(-x))$, Cauchy $(\pi(1+x^2))^{-1}$, and students t-distribution) and concluded that the best transform for the kernel estimate is the isosceles triangular density.

Comparison of relative efficiencies of various kernels Rao (1983, p.66) suggests that the Epanechnikov's kernel is optimal, and is expressed by

$$K(y) = \begin{cases} (1-y^2)3/4 & \text{for } |y| \leq 1 \\ 0 & \text{for } |y| \geq 1 \end{cases}$$

and (see eq.15)

$$W(y) = \begin{cases} -1/2 & \text{for } y \leq -1 \\ y(1-y^2/3)3/4 & \text{for } -1 \leq y \leq 1 \\ 1/2 & \text{for } y \geq 1. \end{cases} \tag{16}$$

It is evident then that the combination of a triangular transform and the Epanechnikov's kernel should provide an optimal estimation of density function, with the smoothing factor (see eq.12) given by

$$h = (5/192\pi n)^{1/5} \tag{17}$$

and the quantity $B(g)$ (see eq. 9) gives

$$B(g) = [\frac{1}{2}(\int \sqrt{g})^4 \cdot \int |g''|]^{1/5} = 1.4460. \tag{18}$$

The denominator in eq. 4 becomes

$$SUM1 = \int_0^1 g_n(y)dy = \frac{1}{n}\sum_{i=1}^{n}[W(\frac{y_i}{h}) - W(\frac{y_i - 1}{h})]. \tag{19}$$

Therefore, eq. 2 is written as

$$f_n(x) = \frac{g_n(y)}{SUM1}T'(x) \tag{20}$$

where $y = T(x)$, and based on eq. 15, the following range on x applies

$$x_{min} - h_* \leq x \leq X_{max} + h_*. \tag{21}$$

The exceedance probability (see eq. 14) is

$$\begin{aligned} P_n(x) &= \int_x^\infty f_n(x)\,dx \\ &= \int_x^\infty \tilde{g}_n[T(x)]T'(x)\,dx \\ &= \int_{T(x)}^1 \tilde{g}_n[T(x)]\,dT(x) \\ &= \frac{1}{SUM1}\int_{T(x)}^1 g_n[T(x)]\,dT(x). \end{aligned} \tag{22}$$

Considering eq. 3, eq. 22 becomes

$$P_n(x) = \frac{1}{n\,SUM1}\sum_{i=1}^{n}[W(\frac{1-y_i}{h}) - W(\frac{y-y_i}{h})]. \tag{23}$$

For a given set of observations $x_i, i = 1, 2, ..., n$, using eq. 23 and the transform $y_i = T(x_i)$ it is possible to compute the exceedance probability for preselected values of x. Alternatively, for a given probability, the value of x can be computed (i.e., design flood with a preselected probability of exceedance).

2.2. The circular transformed density estimate

Devroye et al.(1985) investigated several different transforms and concluded that the triangular transform is most efficient. However, it will be demonstrated in the following that a circular transform (identical to the Epanechinkov's kernel) performs better than a triangular one.

The circular density transform can be expressed as follows (see also eq. 16)

$$K(x) = \begin{cases} (1 - y^2/a)3/4 & for\ |y| \leq 1 \\ 0 & for\ |y| \geq 1. \end{cases} \tag{24}$$

The quantity $B(f)$ for a circular density transform is given by

$$B(g) = (3^3\pi^4/2^9)^{1/5} = 1.3872 \tag{25}$$

where $\int_{-\infty}^{\infty} \sqrt{g(x)}\, dx = \sqrt{3\pi}/4$; $\int_{-\infty}^{\infty} |g'(x)|\, dx = 0$; and $\int_{-\infty}^{\infty} |g''(x)|\, dx = 3$ (see eq . 9). Comparing eqs. 25 and 18 it can observed that the circular transform is more efficient than the triangular one due to the smaller value of $B(g)$. This circular transform has not been investigated and reported elsewhere in the literature.

Based on the circular density transformation, the smoothing factor h is given by

$$h = \left(\frac{15}{2\pi n}\right)^{1/5} \left[\int \sqrt{g} / \int |g''|\right]^{2/5} = \left(\frac{5\pi}{16n}\right)^{1/5} \tag{26}$$

where $g(.)$ is given by eq. 24. It can easily be shown that for a circular density the transform is

$$T(x) = 2\cos[\frac{4\pi + \arccos(1 - 2F(x))}{3}] \tag{27}$$

and

$$T'(x) = \frac{4}{3}\sin[\frac{4\pi + \arccos(1 - 2F(x))}{3}] \cdot \frac{f(x)}{\sqrt{1 - (1 - 2F(x))^2}} \tag{28}$$

where $f(x)$ and $F(x)$ are given by eqs. 13 and 14., respectively.

Introducing the "normalizing" adjustment (see eq. 4) gives

$$\tilde{g_n}(y) = g_n(y) / \int_{-1}^{1} g_n(y)\, dy \tag{29}$$

where the denominator in eq. 29 is given by

$$\begin{aligned} SUM2 &= \int_{-1}^{1} g_n(y)\, dy \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_a^b K(x)\, dx \end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}[W(\frac{1-y_i}{h}) + W(\frac{1+y_i}{h})] \tag{30}$$

where $K(.)$ and $W(.)$ are given in eq. 16. Therefore, eq. 2 becomes

$$f_n(x) = \frac{g_n(x)}{SUM2} \cdot T'(x) \tag{31}$$

or after substituting eq. 3 into eq. 31 results in the following

$$f_n(x) = \frac{T'(x)}{nhSUM2}\sum_{i=1}^{n} K(\frac{y-y_i}{h}). \tag{32}$$

The exceedance probability (see eq. 14) then becomes

$$P_n(x) = \frac{1}{nSUM2}\sum_{i=1}^{n}[W(\frac{1-y_i}{h}) - W(\frac{y-y_i}{h})]. \tag{33}$$

For a given set of observations $x_i, i = 1, 2, ..., n$, using eq. 33 it is possible to compute the exceedance probability for a preselected value of x, or alternatively, for a given probability the value of x can be computed.

3. Numerical studies

A computer simulation experiment was run to compare the performance of the non-parametric transformed method with the parametric method. 50,000 random samples were generated (Pilon, 1986, personal communication) from a Log Pearson type III (LP III) distribution. Of course, the true distribution of floods is not known. Thus, the choice of the LP III distribution for simulation is arbitrary. However, it was made based on the fact that the LP III distribution is widely used in flood frequency analysis, it is a base (or imposed) method in some countries (i.e. USA), and has been used in the past (i.e. Adamowski, 1985) in the evaluation of flood frequency methodology.

The data was generated from an LP III using the Wilson-Hilferty transformation (Adamowski, 1985)

$$LnX_t = C + A\{t/(3B^{1/6}) - (\frac{1}{9}B^{2/3}) + B^{1/3}\}^3 \tag{34}$$

where A, B and C are scale, shape and location parameters, and t is the standard normal deviation. The values of parameters used in generation are: $A = -0.06, B = 25$ and $C = 7$, and are average representative conditions for Canada. The generated sample had the length equal to $50,000$, and the statistics for generated data are: $mean = 255$, $standard\,deviation = 73$, and $coefficient\,of\,skew = 0.443$.

It is obvious that the simulation study is rather limited, and the conclusions are therefore tentative. Nevertheless, an indication of the performance of various methods can be obtained.

The errors between estimates obtained from different methods and the population values were computed using the root mean square error (RMSE) given by

$$RMSE = \sqrt{(BIAS)^2 + Variance}. \tag{35}$$

The simulation results (Table 1) suggest that the differences in the results obtained by the different methods are not remarkable. Comparison of the nonparametric methods indicates that the circular transformed method (CIT) is more accurate than the triangular transformed method (TRT). However, the nonparametric kernel (KE) is more accurate than both the CIT and TRT methods for low return periods of 50 and 100 years, but is less accurate for a 200 year return period.

Certainly, due to the limited number of experiments, the results are not expected to reproduce a wide range of conditions, nevertheless, they do provide an indication of the likely results. It would appear that the transformed kernel estimate gives more accurate results for a higher return period, but it is less accurate for lower return periods. Therefore, there seems to be a compromise between the accuracy and the transformation. The results then indicate that for skewed distribution and long return periods it might be advantageous to use the circular transformed method for higher quantity estimate.

TABLE 1. Comparison of Flood Estimates by LP III Distribution and the Nonparametric Methods

Event	Popul.	Flood Estimates by LP III and Nonparametric Methods			
		Mean (Bias %)[RMSE]			
		LP III	TRT	CIT	KE
Q_{50}	424	420(-0.9)[32]	437(3.1)[42]	443(4.5)[45]	429(1.2)[36.5]
Q_{100}	450	446(-0.9)[41.7]	454(0.9)[43.4]	458(1.8)[44.3]	47(-0.7)[41.8]
Q_{200}	475		474(0.1)[44.9]	475(0.0)[42.1]	461(-2.7)[47.0]

Note: Sample size equals 50, and replication equals 1000. Q_{50}, Q_{100} are the 50 and 100 year return period events; Popul. means Population, and Population = Flood estimates based on 50,000 simulated data from the LP III distribution; LP III = log Pesrson type III (maximum likelihood) distribution; TRT = Triangular transformed; CIT = circular transform; and KE = kernel nonparametric density estimation methods.

4. Conclusion

The triangular and circular transformed nonparametric methods of density estimation for flood frequency analysis have been compared with kernel estimates and LP III distribution used in data generation. It has been demonstrated that the proposed circular transform is more efficient than the triangular transform.

The simulation study however reveals that the transformed kernel density estimates are less accurate than the untransformed kernel for low return periods, but is more accurate for higher return periods. It is suggested then that for skewed distributions and high return periods the circular transform nonparametric kernel method might be a better procedure for estimating flood frequency quantities.

Acknowledgement. We would like to thank Mr. Y. Alila (graduate student at the University of Ottawa) for performing the computer simulation.

This research was supported by the Natural Science and Engineering Research Council of Canada.

References

[1] K. Adamowski, Nonparametric kernel estimation of flood frequencies, *Water Res. Research*, **21** : 11 (1985), 1585-1590.

[2] M. Beran, J.R.M. Hoskings and A. Arnell, Comment on two-component extreme value distribution for flood frequency analysis, *Water Res. Research*, **22** : 2 (1986), 263-266.

[3] L. Breiman, W. Meisel and E. Purcell, Variable kernel estimates of multivariate densities, *Technometrics*, **19** : 2 (1977), 135-144.

[4] L. Devroye and L. Gyorfi, Nonparametric density estimation, *J. Wiley and Sons Ltd.*, 1985

[5] P.J. Pilon, R. Condie and K.D. Harvey, Consolidated frequency analysis package, Water Resources Branch, inland Waters directorate, Environment Canada, July 1985.

[6] B.L.S.P. Rao, Nonparametric functional estimation, academic Press, 1983.

[7] E. Schuster and S. Yakowitz, Parametric/nonparametric mixture density estimation with application to flood-frequency analysis, *Water Res. Bull.*, **21** : 5 (1985), 797-804.

[8] V.P. Singh, International symposium on flood frequency and risk analysis, Louisiana state university, Baton Rouge, USA, (1986), 14-17.

[9] Jr.W.O. Thomas, A uniform technique for flood frequency analysis, *A.S.C.E., Journal of Water Res. Planning and Manag.*, **3** : 3 (1985), 321-337.

[10] J.R. Wallis and E.F. Wood, Relative accuracy of Log Pearson III procedures, *A.S.C.E., Journal of Hydr. Engineering*, **3** : 7 (1985), 1043-1056.