Matematika, 2002, Jilid 18, bil. 2, hlm. 99–107

Nonparametric Kernel Estimation of Annual Maximum Stream Flow Quantiles

Ani Shabri
Department of Mathematics, Faculty of Science
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia

Abstract A nonparametric kernel methods is proposed and evaluated performance for estimating annual maximum stream flow quantiles. The bandwidth of the estimator is estimated by using an optimal technique and a cross-validation technique. Results obtained from a limited amount of real data from Malaysia show that quantiles estimated by nonparametric method using these techniques have small root mean square error and root mean absolute error. Based on correlation coefficient test shown that the nonparametric model approach is accurate, uniform and flexible alternatives to parametric models for flood frequency analysis.

Keywords Bandwidth, cross-validation, kernel, correlation coefficient.

Abstrak *Abstrak Kaedah Nonparametric Kernel dicadangkan dan dinilaikan perlaksanaannya dalam menganggar kuantil aliran tahunan maksimum. Penganggar bagi bandwidth dianggarkan menggunakan teknik optimal dan 'cross-validation'. Hasil keputusan menggunakan data sebenar yang terhad dari Malaysia, menunjukkan bahawa penganggar kuantil berdasarkan model nonparametric menggunakan kedua-dua teknik ini menghasilkan nilai punca min ralat kuasa dua dan punca min ralat mutlak yang kecil. Berdasarkan ujian pekali korelasi menunjukkan bahawa pendekatan model nonparametric adalah tepat, seragam dan ianya boleh dijadikan sebagai kaedah alternatif bagi model parametric dalam analisis frekuensi banjir.*

Katakunci *Bandwidth, 'cross-validation', kernel, pekali korelasi.*

1 Introduction

In flood frequency analysis, one of the more important issues is the choice of the best probability distribution. The true distribution is always unknown in practice and often arbitrary choice of or preference for a given distribution increase the estimation uncertainty.

Some countries have tried to find standard distributions for flood frequency analysis in order to avoid the arbitrariness in the selection of the distribution types. Many distributions

have been used to estimate flood flow frequencies from observed annual flood series. The general extreme value (GEV) distribution is recommended as a base model in the United Kingdom. After appraising many distributions, the U.S. Water resources Council (WRC), issued a series of Bulletins recommending the Log-Pearson Type III (LP3) a base method for use by all U.S. Federal Agencies [1]. While the most commonly used in Canada being the three parameter log-normal (LN3), the Generalized Extreme Value (GEV) and the log-Pearson III (LP3) [2].

Recently, nonparametric density estimation methods have gained popularity in a variety of field of science, including hydrology. This model has the advantage. The shapes of nonparametric density functions are directly determined by the data [3]. It is not requiring assumptions about the distribution of the population of interest [1] and the estimation of parameters (i.e. mean, variance and skew) are not needed. The parametric distributions are limited to certain shapes, however nonparametric densities can adapt to the often-irregular empirical distribution of random variables found in nature [3].

In the present study, the nonparametric kernel method was compared with the GEV, LNIII, GPA and EV1 distributions. The objective of the present study is to introduce and evaluate the performance of annual maximum series using nonparametric techniques and parametric distributions based on real data.

2 Nonparametric Frequency Analysis

The probability density function f(x) estimated by a nonparametric method is given by [2]

$$f(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right) \tag{1}$$

where x_1 to x_n are the observations, $K()$ is a kernel function, itself a probability density function, and h is a bandwidth or smoothing factor to be estimated from the data.

The following condition are impose on the kernel [2]:

$$\begin{aligned} \int K(z)\,dz &= 1 \\ \int zK(z)\,dz &= 0 \\ \int z^2K(z)\,dz &= C \neq 0 \end{aligned} \tag{2}$$

where C is the kernel variance.

The kernel distribution function is the integration of the density Eq. (1) from $-\infty$ to x [2]:

$$F(x) = \frac{1}{n}\sum_{j=1}^{n} K_l\left(\frac{x-x_j}{h}\right) \tag{3}$$

where $K_l(u) = \int_{-\infty}^{u} K(w)\,dw$.

The kernel distribution function may serve to estimate percentiles corresponding to a given probability of exceedance. The flood quantile x_T with a return period of T years, of

the kernel distribution function is [2]

$$x_T = F^{-1}\left(1 - \frac{1}{T}\right) \tag{4}$$

where $F^{-1}()$ represents the inverse of $F()$. The value of x_T can be determined by solving Eq. (3) numerically.

In nonparametric frequency analysis, the choice of kernel function is not critical since various kernels lead to comparable estimates [2]. Gaussian, Gumbel, and Epanechnikov kernels were tested in flood frequency analysis and it was found that the choice of the kernel does not important, and the shape of kernel does not affect extrapolation accuracy [2]. However the calculation and to choice of the bandwidth, h in Eq. 1 is critical. Table 1 shows some of the commonly kernels functions.

Table 1: Some Kernel Function Where $t = h^{-1}(x - x_i)$

Kernel	$K(t)$
Epanechnikov	$K(t) = \frac{3}{4}(1-t^2)$, $\lvert t\rvert < 1$
Rectangular	$K(t) = \frac{1}{2}$, $\lvert t\rvert < 1$
Biweight	$K(t) = \frac{15}{16}(1-t^2)^2$, $\lvert t\rvert < 1$
Gaussian	$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$
Cauchy	$K(t) = \frac{1}{\pi(1+t^2)}$
EV1	$K(t) = e^{-t-e^{-t}}$

In this study only the Gaussian kernel is considered.

3 Estimation Of Smoothing Factor h

3.1 Optimal Technique (OPT)

Several bandwidth estimation techniques are based on minimization of an estimate of the mean square error (MSE) or of the IMSE function. IMSE is obtained by integrating the MSE function over the domain of x.

The criterion of optimality is based on minimizing the IMSE given by [4]

$$IMSE = \int_{-\infty}^{\infty} \left[\hat{f}(x) - f(x)\right]^2 dx \tag{5}$$

where $\hat{f}(x)$ is an estimate of the known density function $f(x)$. An optimal value of h can be obtained by minimizing the IMSE for a given density $f(x)$ and sample. The value of h has to be derived empirically from the observed data. The optimal value of h, can be

determined numerically by differentiating the objective function Eq. (5) with respect to h and then equating it to zero, is given by [4]

$$h = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x_i - x_j)}{\sqrt{5}n(n-10/3)} \tag{6}$$

Based on numerical studies have indicated that the computation of h is also optimal for small samples where $n = 25$ was the smallest sample considered [1].

3.2 Cross-Validation Technique (CV)

Other method of computing h is to minimize the integrated mean of a cross-validation technique [2]:

$$IMSE = \int_{-\infty}^{\infty} E\left[\hat{f}(x) - f(x)\right]^2 dx$$

In cross-validation, estimates of $f(x)$ are constructed each time using every data point except one. Let $\hat{f}_{-i}(x)$ be the nonparametric kernel estimate ignoring a single data point, x_i; that is

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i}^{n-1} K\left[(x - x_j)/h\right] \tag{7}$$

The smoothing factor h can be determined numerically from cross-validation procedure by solving Eq. (8) [2]

$$\sum_{i=1,i=j}^{n}\sum_{j=1}^{n}\exp\left(\frac{d_{ij}}{4}\right)\left\{\left(1 - \frac{4\sqrt{2n}}{n-1}\exp\left(\frac{d_{ij}}{4}\right)\left(\frac{x_i - x_j}{h}\right) - 1\right)\left(\frac{x_i - x_j}{h} + 1\right) - 1\right\} = 0 \tag{8}$$

where $d_{ij} = -((x_i - x_j)/h)^2$. The cross-validation procedure leads to consistent and asymptotically optimal nonparametric density estimates.

4 Data Used For Numerical Analysis

The annual maximum stream flows for the 10 stations in Malaysia was selected for numerical demonstration. The statistical characteristics of these data are given in Table 1. These data were selected on the basis of length completeness and independence of record. The lengths of record are between 15 to 32 lengths and mean of annual maximum stream flows are from 6 to 560 m^3/s. The data was obtained from the Department of Irrigation and Drainage Malaysia.

Table 2: Characteristic of Annual Maximum Stream Flow Data

No	Stations	River/States	Records Length (years)	Mean m^3s^{-1}	Skewness	Kurtosis
1	1737451	Rantau Panjang, Johor	32	259.2	1.066	3.128
2	1836402	Sayong, Johor	18	120.5	1.042	2.916
3	2527411	Muar, Johor	22	171.6	1.308	4.951
4	2928401	Keratong, Johor	15	561.8	2.053	6.532
5	3224433	Triang, Pahang	21	42.0	1.166	3.763
6	3519426	Bentong, Pahang	29	186.2	0.760	3.020
7	3629403	Lepar, Pahang	23	197.1	0.409	1.427
8	4232452	Kemaman, Terengganu	19	354.3	0.071	2.187
9	4732461	Paka, Terengganu	16	6.2	0.368	2.135
10	5229436	Nerus, Terengganu	15	435.2	0.450	1.787

To assess how well the proposed nonparametric method fit to the observed annual maximum flow data, the $X_T - T$ relationships using Eq. (4) are shown in Figure 1. The observed flood discharge corresponds to a return period which was computed by using the Gringorton plotting position formula [5]

$$T = \frac{n + 0.12}{i - 0.44} \tag{9}$$

in which i is the rank assigned to each data point in the sample with a value of one the highest flow, two for the second highest and so on.

5 Results and Discussion

5.1 Root Mean Square Error And Root Mean Absolute Error

Two criteria were used for comparing the two methods of bandwidth h of nonparametric method. The first criteria is defined as the root mean square error (RMSE) between observed and computed flood discharge for entire sample. The RMSE is calculated by the equations

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_T - \hat{x}_T}{x_T}\right)^2} \tag{10}$$

where x_T and $\hat{x}_T$ are observed and computed flood values, respectively for a given value of T. The other criteria is the root mean absolute error and can be calculated by the equations

$$RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_T - \hat{x}_T}{x_T}\right|} \tag{11}$$

5.2 Goodness-Of-Fit Test

The goodness-of-fit test using correlation coefficient (CC) test is used to test the correlation, r between the observations x_T and computed flood value, $\hat{x}_T$. The CC test is measure of linearity of a probability plot. If the sample to be tested is actually drawn from the hypothesized distribution is expected to be nearly linear and the correlation coefficient will be near to one. If $\bar{x}$ denotes the average value of the observations and $\bar{w}$ denotes the average value of computed flood, then correlation coefficient sample can be defined as [6]

$$r = \frac{\sum(x_T - \bar{x})(\hat{x}_T - \bar{w})}{\sqrt{\sum(x_T - \bar{x})^2 \sum(\hat{x}_T - \bar{w})^2}} \tag{12}$$

A comparison of the bandwidths for the Gaussian Kernel using OPT and CV for each stations is shown in Table 3.

Table 3: Comparison of Bandwidth for a Normal Kernel Base on Values RMSE, RMAE and r

Station No	r(Correlaation)		RMSE		RMAE	
	OPT	CV	OPT	CV	OPT	CV
1	0.996	0.996	0.012	0.013	0.065	0.067
2	0.994	0.999	0.011	0.004	0.054	0.038
3	0.995	0.998	0.034	0.007	0.099	0.052
4	0.996	1.000	0.025	0.0004	0.084	0.013
5	0.994	1.000	0.877	0.0007	0.435	0.015
6	0.998	0.996	0.005	0.014	0.045	0.066
7	0.995	0.999	0.122	0.017	0.146	0.069
8	0.997	0.996	0.008	0.014	0.057	0.070
9	0.997	0.997	0.0017	0.00118	0.029	0.031
10	0.998	0.997	0.0014	0.0019	0.029	0.033

Table 3 shows that the nonparametric method using OPT and CV techniques provide a good fit to data because the all stations have produce correlation coefficient, r nearly one. For all 10 stations, Figure 1 shows that the nonparametric kernel estimator using these techniques can fit the real data points closely and provided a sufficiently good approximation to all data.

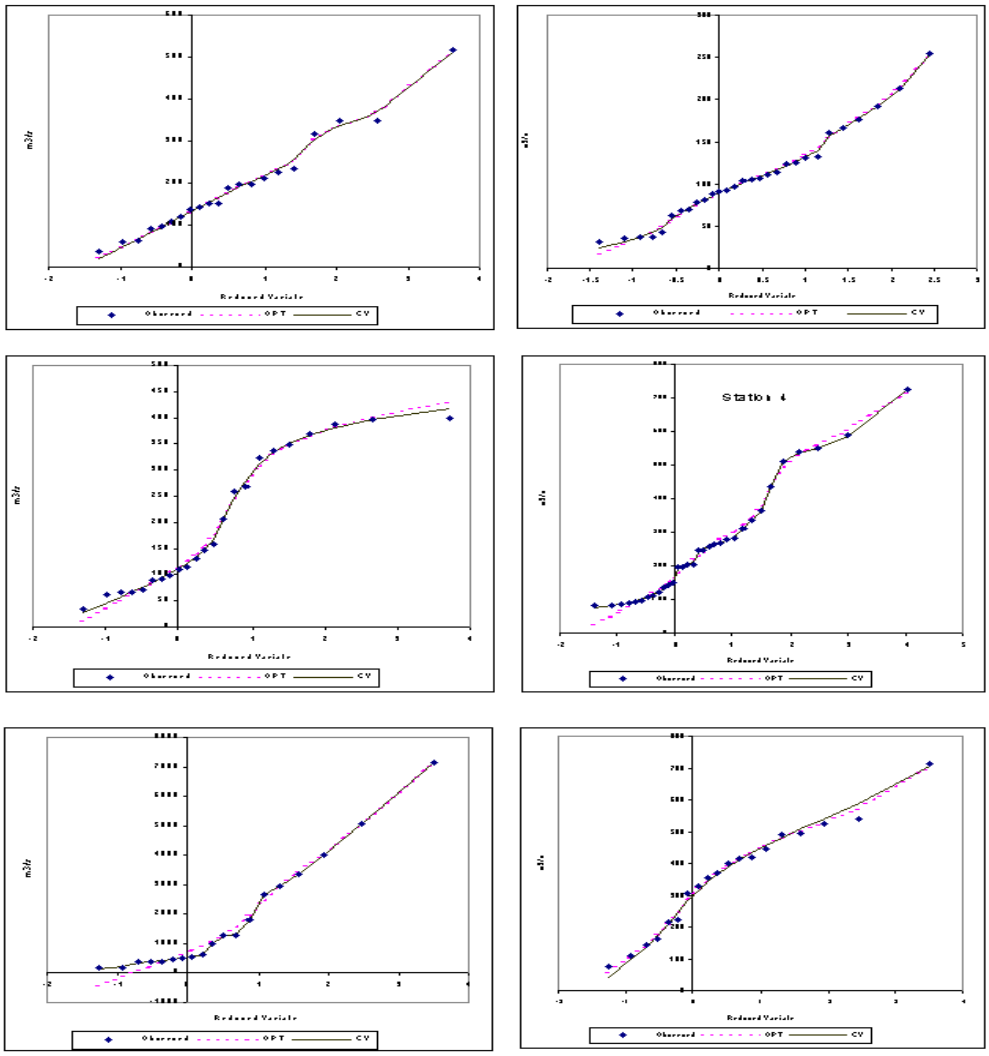

Figure 1: Comparison of Observed and Computed Frequency Curves for the Annual Maximum Stream Flows in Malaysia.

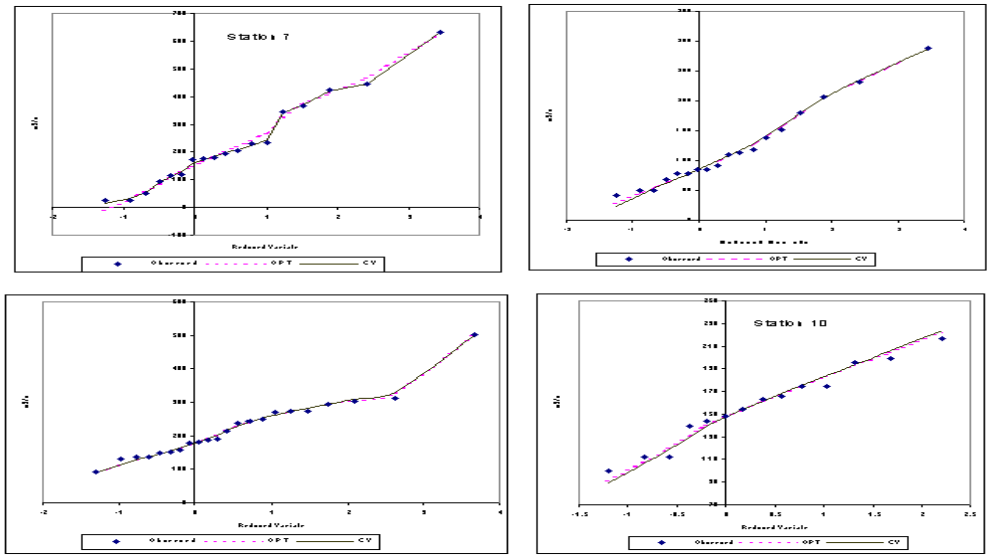

Figure 2: (Continue Figure 1)

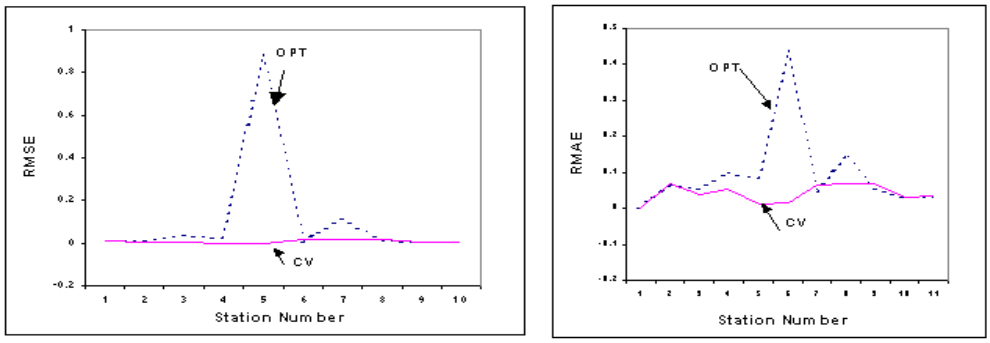

Figure 3: Plot of RMSE and RMAE for OPT and CV Techniques of Nonparametric Method for 10 Stations.

The value of RMSE and RMAE for each stations using OPT and CV technique were presented in Figure 2. It can be seen that the value of RMSE and RMAE fall in nearly zero at all stations except at stations 5 and stations 7 for the OPT technique.

6 Conclusions

The correlation coefficient test statistic as a goodness of fit test is introduced for testing the nonparametric to fit the annual maximum flood flow data. The CC test statistic was found to be useful tool for discriminating among competing bandwidth estimation techniques. Overall these nonparametric method using the OPT and CV techniques consistently produced linear probability plots with r nearly one as measured by the CC test statistics. The nonparametric models with these bandwidth estimation techniques are readily applicable to annual maximum flow frequency analysis.

Based on RMSE and RMSE, clearly, the CV technique is more accurate than the OP technique. However those differences between two methods were not too great and therefore these bandwidth techniques could be considered comparable for practical purpose.

It can be seen from the results presented here, that the nonparametric procedure is useful as it automatically gives stable and accurate estimates of density. It offers a practical and uniform approach to flood frequency analysis.

Acknowledgments

The author would like to thank the Department of Irrigation an Drainage in Malaysia for providing the flood flow data of streams in the Peninsular of Malaysia.

References

[**1**] K. Adamowski. *A Monte Carlo Comparison of Parametric And Nonparametric Estimation of Flood Frequencies*. Journal of Hydrology, 108(1989), 295-308.

[**2**] K. Adamowksi. *Regional Analysis of Annual Maximum and Partial Duration Flood Data by Nonparametric and L-Moment Methods*, Journal of Hydrology, 229(2000), 219-231.

[**3**] D. Faucher, P. F. Rasmussen & B. Bobee. *A Distribution Function Based Bandwith Selection Method for Kernel Quantile Estiamtion*. Journal of Hydrology, 250(2001), 1-11.

[**4**] S.L. Guo, R.K. Kachroo & R.J. Mngodo. *Nonparametric Kernel Estimation of Low Quantiles*, Journal of Hydrology, 185 (1996), 335-348.

[**5**] D. Jain & V.P. Singh. *Estimating Parameters of EV1 Distribution For Flood Frequency Analysi*, Water Res. Resources, 23(1987), 59-70.

[**6**] R.M. Vogel, *The Probability Plot Correlation Coefficient Test for the Normal, Lognormal, and Gumbel Distribution Hpotheses*, Water Resour. Res., 22(4) (1986), 587-590.