Parametric and Non-Parametric Model Performance Differences in Estimation of Critical Stream Flow Discharge Levels

Kiarie F. K.

[1]Department of Management Science, Kenyatta University, P.O Box 43844-00100, Nairobi

Abstract: ***Nonparametric estimation based on quantile regression methodology of Koenker and Basset (1978) and conventional parametric regression approaches were applied to a river regime to estimate volatility in streamflow discharge levels. Consistency and asymptotic normality properties of estimators obtained from both approaches were given. From the study results non-parametric quantile regression approach yielded better results than other methods. Other than for boundary effects which require boundary modifications, the model validation results implied good performance of the nonparametric model in estimating critical streamflow discharge levels.***

Keywords: Kernel estimate, quantile autoregression, consistency, asymptotic normality, critical discharge level

1. Introduction

Flood processes, by their nature, are inherently complex, nonlinear and many times life threatening. As a natural phenomenon, floods, in real sense, cannot be completely controlled. Engineering designs for flood control structures such as flood control reservoirs and dykes are based on previous flood events. The levels of these events can, however, are exceeded by future floods. One way of reducing losses due to floods is by use of flood early warning systems (FEWS). Such a system consists of streamflow monitoring and forecasting as well as public information system. Various methods are employed in hydrological stream flow monitoring and forecasting. Statistical models for hydrological applications are usually based on regression relationships derived from paired catchment and catchment treatment experiments, see Muthusi (2004). For a response variable of interest, for instance, streamflow discharge levels, Yt given a covariate Xt in Ft-1, both parametric and nonparametric regression methods can be applied to estimate critical streamflow discharge levels. This paper examines the difference in model performances in estimation of critical streamflow discharge levels using both parametric and nonparametric regression approaches. In this paper nonparametric estimation focuses on quantile regression methodology introduced in Koenker and Basset (1978) while parametric estimation is based on the conventional mean-variance regression.

2. The Study Model (Critical Streamflow discharge level) for Nonparametric Estimation

Assume that the underlying hydrological process of interest is of the form

$$Yt = m\theta\ (Xt) + et. \qquad (2.1)$$

where Yt is the stream flow discharge at time t measured in cubic meters per second, (cubecs). The variableXt = (Yt-1, ……., Yt-d) is a d-dimensional vector consisting of the past observations of Yt. The conditional quantile function mθ(Xt) is the streamflow discharge level at θ ∊ (0, 1). The errors, et, are assumed to be zero quantile with some scale function ζθ

Here, model (2.1) can be viewed as a robust generalization of Autoregressive (AR) – Autoregression Conditional Heteroscedastic (ARCH) models introduced in Weiss(1984) and their nonparametric generalizations reviewed by Hardle (1989), see Franke and Mwita (2003) and Mwita (2005) for more details.

If we choose Xt = (Yt-1, ……., Yt-d, Ut-1) where the random vector Ut consists of observations from other time series such as soil moisture budget(SMB), precipitation, evapotranspiration, El Nino Southern Oscillations(ENSO), Pacific Decadal Oscillations(PDO), then model (2.1) would become a quantile autoregressive model with exogenous components.

2.1 Nonparametric Estimation of critical streamflow discharge level

We consider the model (2.1), and define a true conditional distribution function Fx(y) of Yt given Xt = x as

$$Fx(y) = P(Yt \leq y \mid Xt = x) = E[It, y \mid Xt = x] \qquad (2.2)$$

where It, y = I{Yt ≤y} is an indicator function with Pr(Yt≤y|Xt = x) = 1 and 0 otherwise.

For any θ ∊ (0, 1), we define the true critical streamflow discharge level as

$$m\theta(x)= \inf\{ y \in R \mid Fx(y) \geq \theta \} \qquad (2.3)$$

The distribution function in (2.2) can be estimated by the Nadaraya (1964) and Watson (1964) estimator as

$$\hat{F}_x(y) = \frac{\sum_{t=1}^{n} K_h(x - X_t) I_{t,y}}{\sum_{t=1}^{n} K_h(x - X_t)} \tag{2.4}$$

where K(u) is a d-dimensional kernel and Kh(u) = h-d K(u/h) is the rescaled kernel, see Franke and Mwita (2003) and Mwita(2005).

Therefore the kernel estimator for the critical streamflow discharge level is given by

$$\hat{m}_\theta(x) = \inf\{R \mid \hat{F}_x(y) \geq \theta\} \equiv \hat{F}_x^{-1}(\theta) \tag{2.5}$$

where $\hat{F}_x^{-1}(\theta)$ denotes the usual generalized inverse of the distribution function $\hat{F}_x(y)$ which is a pure jump function of y.

2.2 Asymptotic Normality

Assume that the time series (Yt, Xt) satisfies α-mixing conditions. According to Masry and Tjostheim (1995, 1997), both ARCH processes and nonlinear additive autoregressive models with exogenous variables are stationary and α-mixing under some mild conditions. As Franke and Mwita (2003) demonstrated, if we choose Xt = Yt-d in (2.1) and assuming the time series Yt is α - mixing, we get an example of a quantile autoregressive process for which (Yt, Xt) and It, y in (2.4) are α -mixing as well.

The following assumptions are necessary for proving asymptotic normality of
$\hat{m}_\theta(x)$

Henceforth, g (x) denotes the stationary probability density of Xt at point x.

(A1)For all u∈R
K (u) ≥ 0
K is Lipschitz continuous i.e. │K(u)- K(v)│≤ Ck│u - v│, for all Ck, u, v∈R and Ck>0
│K(u) │≤ K∞, with K∞ being a constant
∫K(u)du = 1, ∫uK(u)du = 0 and ∫‖ u ‖ 2 k(u)du < ∞
(A2) For all y, x satisfying 0 <Fx(y) <1, g(x) > 0

1) Fx(y)and g(x) are twice continuously differentiable and bounded in y, x
   fx(mθ(x)) > 0, for all x.
2) (A3) The process (Yt, Xt) is stationary and α- mixing with mixing coefficients satisfying α(s) = O(s-(2 + δ)) for some δ >0, n≥ 1, and {sn }is an increasing sequence of positive integers.

The consistenc y and asymptotic normality properties of $\hat{m}_\theta(x)$, and their proofs can be found in Franke and Mwita (2003).

Here, we only state the theorems.

Theorem 3.1

Assume that (A1)- (A3) hold. As n → ∞, let the sequence of bandwidths h> 0 converge to 0 such that nhd → ∞. Then the conditiona l quantile estimator is consistent , $\hat{m}_\theta(x) \xrightarrow{p} m_\theta(x)$
that is

$$E[\hat{m}_\theta(x)] - m_\theta(x) = h^2 B_m(m_\theta(x)) + O(h^2)\ where\ B_m(y) = -\frac{B(Y)}{f_x(Y)} \tag{2.6}$$

Further if, the bandwidths are chosen such that nhd+4 is either 1 or converges to 0, then
$\hat{m}_\theta(x)$ is asymptotically normal,

$$\sqrt{n}h^d\left(\hat{m}_\theta(x) - m(x) - h^2 B_m(m(x))\right) \xrightarrow{D} N\left(0, \frac{V(\hat{m}_\theta(x)}{f_x^2(m_\theta(x))}\right) \tag{2.7}$$

where, B(y) and V2(y) are the bias and variance expansion for the conditional distribution estimator in (2.4)

2.3 Uniform consistency and uniform convergence

For uniform consistency and uniform convergence of the quantile autoregressive estimate, Franke and Mwita(2003) first establish the uniform consistency of the Nadaraya-Watson kernel estimate (2.4). For this purpose, the following conditions are imposed.

(B1) for some compact set G, there are ε>0, γ >0, such that g(x) ≥ γ for all x in the
ε-neighborhood {x;‖x-u‖< ε for some u∈ G} of G.

(B2) (Yt, Xt) is stationary and α-mixing with mixing coefficients α(n), n≥ 1, and there is an increasing sequence sn, n≥ 1, of positive integers such that for some finite A
(n/sn) α 2sn/(3n)(sn) ≤ A, 1≤ sn≤ n/2 for all n≥1.

Uniform consistency and uniform rate of convergence properties of the estimator under the regularity conditions in Franke and Mwita, (2003) are given in Theorem 3.2.

Theorem 3.2
Assume (A1), (A2), (B1), and (B2). If, as n→ ∞, the bandwidthh→0such that

$$\hat{S}_n = nh^d (s_n \log n)^{-1} \to \infty$$

then (3.2.4) is uniformly consistent on G in the strong sense.
That is, for x∈G

$$\sup_{x \in G} |\hat{F}_x(y) - F_x(y)| \to 0 \qquad \text{a.s}$$

In this section, we have shown that the estimate of our nonparametric quantile function is consistent and asymptotically normally distributed, and under suitable conditions, the estimator converges uniformly with an appropriate rate. The asymptotic normality property is used to construct the required confidence intervals for our estimator. These are strong properties that significantly imply sufficiency of the estimator is accurate in estimation of the critical streamflow discharge level.

3. Real Data Results

The application of the estimator was performed with data from the gauge at River Nyando, in Western Kenya, (River Station No. IGD03) in the wider Nyando Basin, located at 35.2 oE longitude and -0.1oS latitude and covering an area of 3, 587 km2. The drainage area downstream of the outlet of the catchment (IGD03) was found to accommodate all the discharge in the river channel. For this reason, monthly maximum streamflow data from gauging station IGD03 for the period 1970 – 1997 was used for calibrating the model. Also, the twenty-seven year period was considered long enough to capture diverse weather conditions, thus making the model to be a good representative of the basin.

Figure 3.1 gives the daily streamflow hydrograph of ground station gauging data for twenty-seven years, from 1970 – 1997. From the hydrograph, it is clear that the river regime experiences both peak and extremely high flows which are responsible for flood inundations experienced in flood plain areas of the Nyando basin.

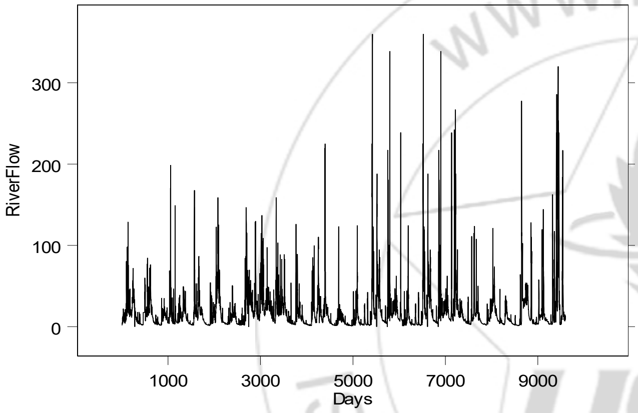


Figure 3.1: Daily streamflow discharges for River Nyando (1970 – 1997) Station (IGD03)

Considering the critical streamflow discharge level to be our target variable, we first present hydrograph for monthly maximum streamflow for the period 1970 – 1997 in Figure 3.2.

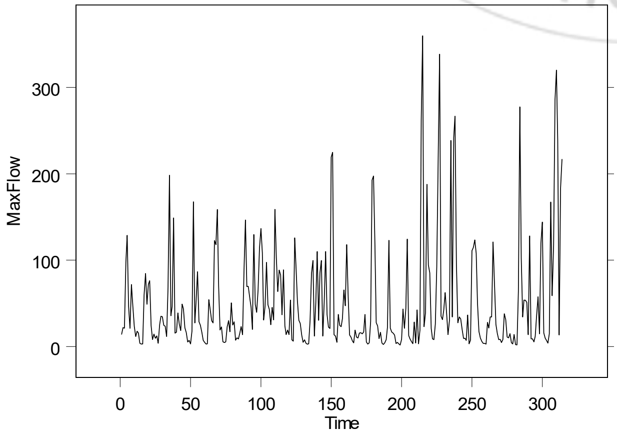


Figure 3.2: Monthly maximum streamflow discharges for River Nyando (1970 – 1997) for Station (IGD03)

The hydrograph of figure 5.2 shows that the river pattern of low flows, peak flows and extremely high flows is preserved by the monthly maximum streamflow time series of our ground station gauging data.

Figure 3.3 gives the volatility of the monthly maximum streamflow discharges. The hydrograph of these deviations depict the turbulence experienced by the Nyando River regime with an observable increase in trend.

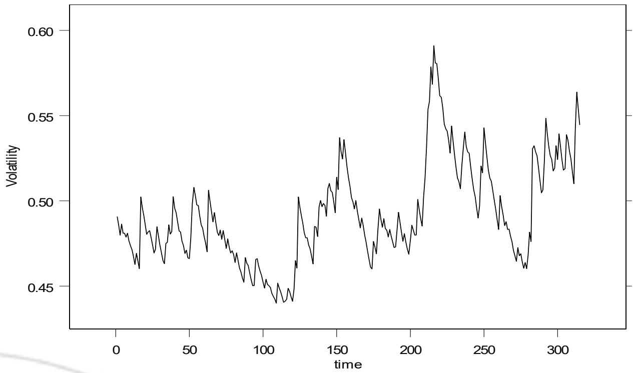


Figure 3.3: Volatility of monthly maximum streamflow discharge for River Nyando.

Figures 3.4 gives the monthly maximum streamflow discharge levels together with 0.95 and 0.99 conditional quantiles respectively.

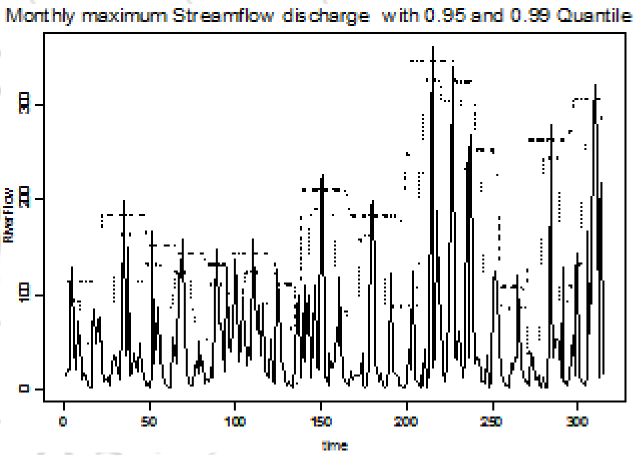


Figure 3.4: Monthly maximum streamflow discharges with 0.95 and 0.99 quantiles.

The dotted curve represents the 0.95 conditional quantile while the dashed curve represents the 0.99 conditional quantile. Streamflow discharge levels above the 0.95-quantile curve represent critical streamflow discharge levels responsible for flood inundations at 95% confidence level. Such a level calls for some site-specific operational instructions to be issued by authorities monitoring the river catchment. The instructions may include shutting of floodgates and other engineering measures. Discharges above the 0.99 quantile curve represent extreme river flow levels. Such levels call for flood control teams to respond to imminent flood conditions and operate a warning system for the public as well as industries.

3.1 Parametric Estimation of critical stream flow discharge level

To compare model performances and the quality of our estimator quantitatively, a comparative assessment was carried out with two different assumptions on the river flow data.

3.1.1 Mean-Variance method

Under this method we assume normality of the streamflow discharge data. Our hydrologic model is then modified to take the following form of a quantile autoregressive-heteroscedastic process

Yt = µ(Xt) + ζ(Xt)et (3.1)

whereµ (Xt) is the conditional mean of Yt given Xt and ζ(Xt) is the conditional standard deviation of Yt given Xt . Using the Nadaraya-Watson estimator of µ (Xt)

$$\hat{\mu}(x) = \frac{\sum_{t=1}^{n} K_h(x - X_t)Y_t}{\sum_{t=1}^{n} K_h(x - X_t)}$$

and the estimator of the variance

$$\hat{\sigma}^2(x) = \frac{\sum_{t=1}^{n} K_h(x - X_t)(Y_t - \hat{\mu}(x))^2}{\sum_{t=1}^{n} K_h(x - X_t)}$$

Therefore, the estimator of the innovations εt is given by

$$\hat{\varepsilon}_t = \frac{(Y_t - \hat{\mu}(x))}{\hat{\sigma}(x)}$$

The conditional quantile function estimator of mθ(Xt) using (3.1) becomes,

$$\hat{\mathrm{m}}_{\theta}(\mathrm{x}) = \hat{\mu}(\mathrm{x}) + \hat{\sigma}(\mathrm{x})\varepsilon_{\theta}$$

Where εθ are the true quantiles of et ~ N(0, 1) i.i.d random variables.

3.1.1.1 Asymptotic properties of the mean-variance estimator

The consistency of the conditional mean and conditional variance estimator is shown in the following preposition, see Hardle (1989).

Proposition 3.1.1.2

Assume the stochastic design model with a one-dimensional predictor variable X and

(D1) ∫|K(u)| du <∞,

(D2) $\lim\limits_{|u|\to\infty} uK(u) = 0$

(D3) EY2 <∞,

(D4) hn→ 0, nhn→∞ as n →∞,

Then, at every point of continuity of µ(x), f(x) and σ2(x), with f(x) > 0

$$\hat{\mu}(x) \xrightarrow{p} \mu(x)$$

For convergence, we first define the mean squared error at a point x as follows

$$d_M(x,h) = E[\hat{\mu}_h(x) - m(x)]^2$$

The following theorem gives the speed of dM(x, h) as a function of h and n. See Hardle(1989).

Theorem 3.1.2(Gasser and Muller 1984)

Assume the random design model with a one-dimensional predictor variable X and define
CK = ∫K2(u)du,
dK = ∫ u2K(u)du.
Assume
(F0) K has support [-1, 1] with K(-1) = K(1) = 0,
(F1) µ∈ C2,
(F2) maxi|Xi– Xi-1| = O(n-1),
(F3) var(εi) = σ2, i = 1, ……….., n,
(F4) n→∞, h → 0, nh→∞.
Then
dM(x, h) ≈ (nh)-1σ2ck + h4dK2[µ″(x)]2 /4.

The mean squared error splits up into the two parts, variance and bias. The theorem shows that the variance and bias goes to zero as n→∞, and h →0, respectively.

3.1.2 Historical simulation

Here, we assume independence and identical distribution (iid) of the streamflow discharge data.

Let Yt, t = 1, 2, ……………be a sequence of iid random variables, with a distribution function F(y), now to represent streamflow discharge level at time t.

We denote Yt = Y since Yt are iid. For some probability θ ∈ (0, 1), consider the
θ -quantile as

qθ y = inf{y∈R | F(y)≥ θ }

If F is continuous, then P(y = qθ y) = 0 and F(qθ y) = θ, while if F is discontinuous in
qθ y, then P(y = qθ y) > 0 and F(qθ y) = P(y ≤ qθ y) > θ.

Consider the order statistics y1, n ≤ y2, n ≤ ………..≤ yk+1, n ≤……………yn, nas the sorted values of n-tuple (y1, y2, ………, yn) and let k = [n(1 – θ)] (= maxm∈N {m≤ n(1 – θ })be the integer part of n(1 – θ). The set of observations which constitute the 100(1 – θ)% largest of the total values in the sample is represented by the largest k observations (outcomes)
{yk, n ……….. y1, n}. As usual, yk+1, n denotes the empirical quantile which we may write as, qθy, n where θ stands for the proportion of observations below yk, n.

3.1.2.1 Asymptotic properties of yk+1, n under the iid assumption

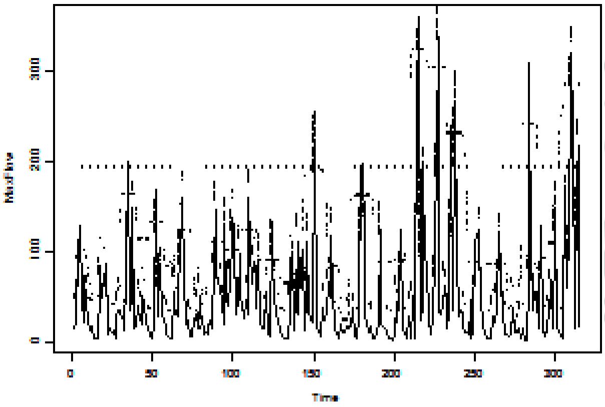
We assume that F has a density function, f and P (yk, n> yk+1, n) = 1.
With n-k = nθ + o(n1/2), i.e k/n = 1 – θ + o(n-1/2), then by central limit theorem

θ (1 – θ)
qθyn~ N(qθy, ————)
nf2(qθy)
where qθy is the asymptotic quantile and θ((1 – θ)

, ———— being the asymptotic variance, nf2(qθy)
See Mwita(2003) and references therein. This asymptotic normality property is used to construct the required confidence intervals for the estimator.

Figure 3.5 gives the 0.95 conditional quantiles obtained under the two assumptions (mean- variance and i.i.d) together with the 0.95-th quantile using model (2.1). The dotted straight line represents the 0.95-th quantile under iid assumption. The dotted curve represents the 0.95-th quantile under model (2.1) while the dashed curve (not visibly clear) with a similar pattern to the solid (actual streamflow discharge curve) represents the 0.95 conditional quantile under the assumption of normality. From the graph, both the iid and mean-variance methods appeared to underestimate the critical streamflow discharge level and therefore performing poorer than model (2.1) at the 95% level.

3.5: 0.95 quantile estimates Under mean variance and iid assumptions together with0.95 quantile under QAR.



Consequently, the nonparametric model is considered to be adequate and better in performance in the estimations of critical stream flow discharge levels.

4. Conclusion

In this paper, a nonparametric quantile regression approach was considered. By avoiding assumptions on the form of the conditional distribution of the streamflow discharge in a river regime, our method of estimating critical streamflow discharge level yields better results than other methods, which make assumptions on the underlying conditional distribution function of the streamflow discharges. Furthermore, the critical streamflow discharge level estimates obtained from the study model fits the quantile data well apart from the extreme ends, which is attributed to data scarcity at the extremes. However, the estimator obtained from our method suffers from boundary effects and therefore requires boundary modifications, see Mwita (2003). Other than the singular drawback of boundary effects, the study model's validation results implies a good performance of the model in estimation of critical streamflow discharge levels.

References

[1] Franke, J. and Mwita, P. (2003). Nonparametric estimates for Conditional Quantiles of Time Series. Report in Economic Mathematics, Nr 87. Technical University of Kaiserslautern.
[2] Hardle, W. (1989). Applied nonparametric regression. Cambridge University press, Cambridge.
[3] Koenker, R. and Bassett, G. (1978). Regression Quantiles. EconometricaVol46, 33-50.
[4] Masry E., and Tjostheim, O. (1995). Nonparametric Estimation and Identification of Nonlinear ARCH time series: Strong convergence and asymptotic normality. Econometric Theory Vol. 11, 258-289.
[5] Masry, E., and Tjostheim, O. (1997). Additive nonlinear ARX time series and projection estimates. Econometric Theory Vol. 13, 214-252
[6] Muthusi, F.M. (2004). Evaluation of the USGS Streamflow Model for Flood Simulation. M.Sc Thesis, Jomo Kenyatta University of Agriculture and Technology, Kenya.
[7] Mwita, P. (2003). Semi-parametric Estimation of Conditional Quantiles for Time Series with Application in Finance. PhD Thesis, University of Kaiserslautern, Germany.
[8] Mwita, P. (2005). On conditional scale function: Estimate and asymptotic properties. African Diaspora Journal of MathematicsVol2(2). In press.
[9] Nadaraya, E. A. (1964). On estimating regression. Theory Probab. Appl. Vol9, 141-142. centers/programmes. Accessed from http://www.update.unu.edu/issue322.htm .
[10] Watson, G. S. (1964). Smooth regression analysis. Sankya ser. A26, 359-372.
[11] Weiss, A. (1984). ARMA models with ARCH errors. Journal of Time Series Analysis Vol.3, 129-143.