ESTIMATION OF CRITICAL STREAMFLOW DISCHARGE LEVEL USING NONPARAMETRIC QUANTILE REGRESSION MODEL

Dr. Francis Kiarie
PhD,
Department of Management Science, Kenyatta University,
P.O Box 43844-00100 Nairobi,
Kenya.
spssnairobi@gmail.com

ABSTRACT
Various parametric models have been designed to analyze volatility in river flow time series data. For maximum likelihood estimation these parametric methods assumes a known conditional distribution. This paper considers the problem of nonparametric estimation of critical streamflow discharge levels of a river regime based on quantile regression methodology of Koenker and Basset (1978). In particular, the paper demonstrates the use of kernel estimators for conditional quantiles resulting from a kernel estimation of conditional distribution function. It is finally proved that the estimate of the nonparametric quantile function is consistent and asymptotically normally distributed and under suitable conditions, the estimator converges uniformly with an appropriate rate.

Key words: Conditional quantile, kernel estimate, quantile autoregression, consistency, asymptotic normality, critical discharge level.

1. INTRODUCTION
Globally, floods impact an estimated 520+ million people per year, resulting in estimates of up to 25,000 annual deaths, extensive homelessness, disaster-induced diseases, crop and livestock damage and other serious harm (UNU, 2004). 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. For hydrological applications, models are usually based on regression relationships derived from paired catchment and catchment treatment experiments, Muthusi (2004).

The conventional method to estimate the conditional distribution function \( F(Y_t|X_t) \) is to assume that \( Y_t|X_t \) follows a particular distribution and then estimate its parameters. However, such specification may lead to wrong inferences and therefore, in this paper, we focus on quantile regression methodology introduced in Koenker and Basset (1978). This is a flexible method that requires no strict assumptions on moments and distribution of the underlying process. Instead of assuming that \( m(X) \) is a conditional mean, it is assumed to be \( \theta \)-th conditional quantile and denoted as \( m_\theta(X_t) \).

In this paper, I use nonparametric approach to model \( m_\theta(X_t) \). I first estimate, non-parametrically, the conditional distribution of \( Y_t \) given \( X_t \) and then invert it at \( \theta \)-level of probability to get conditional \( \theta \)-quantile estimate as in Franke and Mwita (2003).

2. THE STUDY MODEL (CRITICAL STREAMFLOW DISCHARGE LEVEL)
Assume that the underlying hydrological process of interest is of the form

\[
Y_t = m_\theta (X_t) + e_t.
\]  \( (2.1) \)

where \( Y_t \) is the streamflow discharge at time \( t \) measured in cubic meters per second, (cubecs). The variable \( X_t = (Y_{t-1}, \ldots, Y_{t-d}) \) is a \( d \)-dimensional vector consisting of the past observations of \( Y_t \). The conditional quantile function \( m_\theta(X_t) \) is the streamflow discharge level at \( \theta \in (0, 1) \). The errors, \( e_t \), are assumed to be zero quantile with some scale function \( \sigma \). Therefore, the errors may also include heteroscedastic cases, see Franke and Mwita (2003), Mwita (2005) among others.
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 \( Xt = (Y_{t-1}, \ldots, Y_{t-d}, U_{t-1}) \) 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 Estimation of critical streamflow discharge level

We consider the model (2.1), and define a true conditional distribution function \( F_x(y) \) of \( Y_t \) given \( X_t = x \) as

\[
F_x(y) = P(Y_t \leq y \mid X_t = x) = E[I_{t,y} \mid X_t = x]
\]  
(2.2)

where \( I_{t,y} = I\{Y_t \leq y\} \) is an indicator function with \( Pr(Y_t \leq y \mid X_t = x) = 1 \) and 0 otherwise.

For any \( \theta \in (0, 1) \), we define the true critical streamflow discharge level as

\[
m_{\theta}(x) = \inf\{ y \in \mathbb{R} \mid F_x(y) \geq \theta \}
\]  
(2.3)

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

\[
\hat{F}_x(y) = \frac{\sum_{i=1}^{n} K_h(x - X_i)I_{i,y}}{\sum_{i=1}^{n} K_h(x - X_i)}
\]  
(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

\[
\hat{m}_{\theta}(x) = \inf\{ y \in \mathbb{R} \mid \hat{F}_x(y) \geq \theta \} = \hat{F}_x^{-1}(\theta)
\]  
(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 \((Y_t, X_t)\) satisfies \( \alpha \)-mixing conditions. According to Masry and Tjostheim (1995, 1997), both ARCH processes and nonlinear additive autoregressive models with exogenous variables are stationary and \( \alpha \)-mixing under some mild conditions. As Franke and Mwita (2003) demonstrated, if we choose \( X_t = Y_{t-d} \) in (2.1) and assuming the time series \( Y_t \) is \( \alpha \)-mixing, we get an example of a quantile autoregressive process for which \((Y_t, X_t)\) and \( I_{t,y} \) in (2.4) are \( \alpha \)-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 \( X_t \) at point \( x \).

(A1) For all \( u \in \mathbb{R} \)

(i) \( K(u) \geq 0 \)

(ii) \( K \) is Lipschitz continuous i.e. \( \left| K(u) - K(v) \right| \leq C \left| u - v \right| \), for all \( C \in \mathbb{R} \) and \( \alpha > 0 \)

(iii) \( \left| K(u) \right| \leq K_{\infty} \), with \( K_{\infty} \) being a constant

(iv) \( \int K(u)du = 1, \int uK(u)du = 0 \) and \( \int \left| u \right| K(u)du < \infty \)

(A2) For all \( y, x \) satisfying \( 0 < F_x(y) < 1 \), \( g(x) > 0 \)

(i) \( F_x(y) \) and \( g(x) \) are twice continuously differentiable and bounded in \( y, x \)

(ii) \( \hat{f}_x(m_{\theta}(x)) > 0 \), for all \( x \).

(A3) The process \((Y_t, X_t)\) is stationary and \( \alpha \)-mixing with mixing coefficients satisfying \( a(s) = O(s^{-2 + \delta}) \) for some \( \delta > 0 \), \( n \geq 1 \), and \( \{s_n \} \) is an increasing sequence of positive integers.

The consistency 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 conditional quantile estimator is consistent \( \hat{m}_\theta(x) \rightarrow m_\theta(x) \)
that is
\[
E[\hat{m}_\theta(x)] - m_\theta(x) = h^2 B_\theta(m_\theta(x)) + O(h^2)
\]
where \( B_\theta(y) = \frac{B(Y)}{f_\theta(Y)} \) (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,
\[
\sqrt{n}h^2(\hat{m}_\theta(x) - m(x) - h^2 B_\theta(m(x))) \xrightarrow{d} N\left(0, \frac{V(\hat{m}_\theta(x))}{f_\theta^2(m_\theta(x))}\right)
\] (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 establishes 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 \( \varepsilon, \gamma >0 \), such that \( g(x) \geq \gamma \) for all \( x \) in the \( \varepsilon \)-neighborhood \( \{x; ||x - u|| < \varepsilon \text{ for some } u \in G\} \) of G.

(B2) \((Y_t, X_t)\) is stationary and \( \alpha \)-mixing with mixing coefficients \( \alpha(n), n \geq 1 \), and there is an increasing sequence \( s_n, n \geq 1 \), of positive integers such that for some finite \( A \)
\[
(n/s_n) \alpha 2s_n/(3n)(s_n) \leq A, 1 \leq s_n \leq n/2 \text{ for all } n \geq 1.
\]
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 \( n \rightarrow \infty \), the bandwidth \( h \rightarrow 0 \) such that
\[
\hat{S}_n = nh^d (s_n \log n)^{-1} \rightarrow \infty
\]
then (3.2.4) is uniformly consistent on G in the strong sense. That is, for \( x \in G \)
\[
\sup_{x \in G} \left| F_\theta(x) - F_\theta(x) \right| \rightarrow 0 \text{ a.s}
\]

2.4 Summary

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 our estimator is accurate estimation of the critical streamflow discharge level.

3. REAL DATA RESULTS

The application of our 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. Flooding is experienced starting from Ahero plains, down to Lake Victoria through KUSA swamps. 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.

![Monthly maximum Streamflow discharge with 0.95 and 0.99 Quantile](image)

**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 Model validation

To demonstrate that the study model produced good estimates, a model validity test using the Basler Ampel method of Backtesting was performed. This method is mainly applied in financial modeling. However, its basic principles do apply to other applications as well.

The Basler Ampel method suggests that we define a Bernoulli- distributed series of random variables $B_i$, such that $B_i = 1$ if $Y_i > m_\theta(x_i)$ or $B_i = 0$ if $Y_i < m_\theta(x_i)$, $i \in \{1, 2, \ldots, n\}$ where $n$ is the number of backtesting points. $Y_i$ is the streamflow discharge level for the $i$-th month. $m_\theta(x_i)$ is the $i$-th month $\theta$-th quantile level. This number is usually set at 250 or 500, but can be selected arbitrarily.

Using this model validation method, the study model was tested using a sample of size 250 data points. At both 99% and 95% confidence levels, the model results (4 data points at 99% and 16 data points at 95%) lied within the green zone of acceptance. Consequently, the study model was considered adequate in the estimations of critical stream flow discharge levels.

### 4. REFERENCES


