Nordic Hydrolgy, 26, 1995, 73-90

Regional Flood Relationships by Nonparametric Regression

D. Gringras, M. Alvo and K. Adamowski

University of Ottawa, Ontario, Canada K1N 6N5

Since some theoretical assumptions needed in linear regression are not always fulfilled in practical applications, nonparametric regression was investigated as an alternative method in regional flood relationship development. Simulation studies were developed to compare the bias, the variance and the root-mean-square-errors of nonparametric and parametric regressions. It was concluded that when an appropriate parametric model can be determined, parametric regression is preferred over nonparametric regression. However, where an appropriate model cannot be determined, nonparametric regression is preferred. It was found that both linear regression and nonparametric regression gave very similar regional relationships for annual maximum floods from New Brunswick, Canada. It was also found that nonparametric regression can be useful as a screening tool able to detect data deficient relationships.

Introduction

Estimation of design floods for a watershed with no data can be improved by regional analysis procedures which incorporate relevant information from other watersheds. Regional analysis involves three basic steps: single station flood frequency analysis, homogeneous region delineation, and regional relationship development. In this paper, the latter step is further investigated, considering the multiple regression approach.

In regional flood frequency analysis, homogeneous region delineation attempts to define regions having similar hydrological characteristics and/or flood related variables in order to allow information transfer to ungauged sites. Of the numerous approaches to this delineation, some of the data-based techniques involve using multivariate statistical methods (Cavadias 1990) while others involve finding regions with a single flood frequency distribution (Hosking and Wallis 1993). More physically-based approaches involve delineating areas with floods of similar mechanisms (Gingras and Adamowski 1993). In the multiple regression approach, linear relationships between logarithmically transformed design floods and physiographic/climatic characteristics in each of the delineated regions are then developed (Kite 1977).

Inferences based on linear regression and least-squares approaches require several assumptions. Ordinary least-squares analysis requires that the relationship be linear in the parameters, that there be constant variance of the "errors" about the regression line, and that these errors be uncorrelated and normally distributed with mean zero. All of these assumptions are often violated in practical applications (Holder 1985). Generalized least-squares analysis (Tasker and Stedinger 1989) will overcome some of the deficiencies of ordinary least squares because it takes into consideration the fact that the standard errors of the dependent variables (the design floods) are different, mainly due to differing record lengths, and that some correlation exists among the annual maximums at the various sites within a homogeneous region.

A major concern in regional analysis is finding the appropriate relationship between the design floods and the relevant physiographic/climatic variables. There exists no physical justification for the selection of a linear relationship between logarithmically transformed data as is typically chosen. A nonlinear relationship may be appropriate in some instances in describing the variations of logarithmically transformed design floods and physiographic/climatic variables. A misspecification of the regional relationship, by using a linear model instead of a nonlinear one, for example, will result in a systematic error or bias which ultimately results in misspecified design floods at the ungauged sites.

Another concern is that the scatter around the regression line for a particular homogeneous region is often not uniform. Because of differences in basin slopes and in the amount of wetlands, there is a large variation in flood magnitude for a given basin size, particularly for the smaller basins. For example, a steep and rocky ten kilometre square basin can generate a given design flood perhaps as much as ten times larger than a ten kilometre square basin on a flat slope comprised of a large percentage of lakes. If part of the homogeneous region is much steeper or wetter than the remainder of the region, the variations around the regression line are not expected to be normally distributed and one may then consider the use of indicator variables or of weighted least squares. As well, one may often wish to incorporate further independent variables in the regional relationship but there may exist degrees of freedom limitations because of small sample size.

An alternative approach seeking to alleviate some of these deficiencies is nonparametric regression, which is based on fewer assumptions. Nonparametric regression requires no specification as to the form of the relationship between the variables. As well, when the error variance changes monotonously with the explanatory variables, nonparametric regression is expected to perform better than a least-squares fit due to its local character. It is not claimed, however, that nonparametric regression can solve all the problems raised by the application of linear regression.

Nonparametric regression has been applied in hydrology to predict groundwater levels from runoff (Adamowski and Feluch 1991), in a long-range streamflow forecasting model (Smith 1991), and to identify relationships between local daily precipitation and average pressure height (Matyasovszky *et al.* 1993). This paper explores the use of nonparametric regression for regional flood relationship development and compares the results obtained with those of parametric regression. The bias, the variance, and the root-mean-square-errors of the estimates from parametric and nonparametric regression are evaluated through simulations. Nonparametric regression is also found to assist in model selection and is useful as a screening tool for pointing out data deficient relationships.

Theoretical Development

Nonparametric Frequency Analysis

The nonparametric regression estimate is defined as a conditional mean which can be expressed as the ratio of two integrals involving a probability density function. Therefore, theoretical developments dealing with nonparametric density functions are presented first.

Nonparametric density estimation has been successfully employed in single station frequency analysis by several researchers including Adamowski (1989), Bardsley (1989), and Guo (1991). They found the nonparametric approach useful in overcoming some of the drawbacks of conventional parametric methods related to distribution selection, tail behaviour, and unimodality among others. It was concluded that nonparametric methods were particularly suitable for multimodal annual flood data following mixed distributions (Gingras and Adamowski 1992).

The probability density function $f(x)$ can be estimated in a nonparametric manner on the basis of a random sample $x_1 \ldots x_n$ by (Adamowski 1985)

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

where $K()$ is a kernel function, itself a probability density function such as a normal or rectangular distribution, and h is a smoothing factor to be estimated from the data.

The process of nonparametric density estimation is similar to that of building a

histogram with a class interval of width h. In a histogram, a rectangular block of height $1/nh$ is added at the center of the class interval to which a given data point belongs. The final histogram frequency distribution is the sum total of all blocks, each one of area $1/n$. In nonparametric frequency, a kernel function centered at the data point location itself is added and the final nonparametric density is the sum total of all kernels.

The choice of the kernel function is not crucial to the performance of the method as various kernels lead to comparable estimates (Prakasa Rao 1983). However, it must satisfy the following conditions (Silverman 1986)

$$\int K(x)\,dx = 1 \tag{2}$$

$$\int xK(x)\,dx = 0 \tag{3}$$

$$\int x^2 K(x)\,dx < \infty \tag{4}$$

A standard Gaussian kernel was selected in this study and is given by

$$K(x) = \frac{1}{\sqrt{2}} \exp\left(-\frac{x^2}{2}\right) \tag{5}$$

The selection of the smoothing factor h in Eq. (1) is, however, critical. One method of computing h is to minimize, by means of a cross-validation technique, the integrated mean-square-error (IMSE) (Silverman 1986)

$$\mathrm{IMSE} \equiv E\left[\int (\hat{f}(x) - f(x))^2 dx\right] \tag{6}$$

$$= E\left[\int \hat{f}(x)^2 dx - 2\int f(x)\hat{f}(x)\,dx + \int f(x)^2 dx\right] \tag{7}$$

where expectation is taken over the random variables $x_1 \ldots x_n$ defining $\hat{f}(x)$. Because IMSE must be minimized with regard to h, the last term, which does not involve $\hat{f}(x)$, can be discarded.

In cross-validation, estimates of $f(x)$ are constructed each time using all the data points but one. Thus, $\hat{f}_{-i}(x)$ is the nonparametric kernel estimate ignoring a single data point x_i

$$\hat{f}_{-i}(x) \equiv \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{x - x_j}{h}\right) \tag{8}$$

It has been shown (Silverman 1986) that

$$E\,\frac{1}{n}\sum_i \hat{f}_{-i}(x_i) = E\int \hat{f}(x) f(x)\,dx \tag{9}$$

Inserting Eq. (9) into Eq. (7) and ignoring its last term, the risk function to be mini-

mized, $R(h)$, which depends on the smoothing factor h, is

$$R(h) = E\left(\int \hat{f}(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x_i) \right) \tag{10}$$

The basic principle of least squares cross-validation is to construct an estimate of $R(h)$ from the data themselves and then to minimize this estimate over h to give the smoothing factor. It has been shown (Rudemo 1982), that Eq. (10), for a normal kernel as defined by Eq. (5), is equivalent to

$$R(h) = \frac{1}{2\sqrt{\pi}nh} \left[1 + \sum_{i=1, i\neq j}^{n} \sum_{j=1}^{n} \frac{2}{n} \exp\left(\frac{d_{ij}}{4}\right) - \sum_{i=1, i\neq j}^{n} \sum_{j=1}^{n} \frac{4\sqrt{2}}{n-1} \exp\left(\frac{d_{ij}}{2}\right) \right] \tag{11}$$

where

$$d_{ij} = -((x_i - x_j)/h)^2$$

Setting the derivative of $R(h)$ with respect to h equal to 0 results in the following equation

$$\frac{1}{2\sqrt{\pi}nh} \left(\sum_{i=1, i\neq j}^{n} \sum_{j=1}^{n} \exp\left(\frac{d_{ij}}{4}\right) \left\{ \left(1 - \frac{4\sqrt{2n}}{n-1} \exp\left(\frac{d_{ij}}{4}\right)\right) \left(\frac{x_i - x_j}{h} - 1\right) \right) \right.$$
$$\left. \left(\frac{x_i - x_j}{h} + 1\right) - 1 \right\} \Bigg) = 0 \tag{12}$$

Therefore, the value of h can be determined numerically by solving Eq. (12). Scott and Terrell (1987) have shown that the cross-validation procedure leads to consistent and asymptotically optimal nonparametric density estimates.

Nonparametric Regression

Nonparametric regression, which does not require strong assumptions about the shape of relationships, is considered a supplement to parametric analyses (Altman 1992). In nonparametric regression, the predicted value of the response variable is a conditional mean defined as the ratio of the integral of a probability density function and of a probability density function, each of which is estimated nonparametrically. The optimal estimate of a function Y given that variable X equals x can be expressed as follows (Muller 1988)

$$E(Y|X=x) \equiv \frac{\int y\, f(x,y)\, dy}{f_X(x)} \tag{13}$$

where $f(x,y)$ is the joint density function of X and Y, and $f_X()$ is the marginal density function of X.

A nonparametric estimate of the joint bivariate density function can be expressed as

$$\tilde{f}(x,y) = \frac{1}{n h_x h_y} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right) K\left(\frac{y-y_i}{h_y}\right) \tag{14}$$

where h_x and h_y refer to the smoothing factors associated with x and y.

An estimate of the conditional mean of Y given $X = x$ is obtained by substituting Eq. (14) into the right-hand side of Eq. (13), and is given by

$$\frac{\sum_{i=1}^{n} y_i K((x-x_i)/h_x)}{\sum_{i=1}^{n} K((x-x_i)/h_x)} \tag{15}$$

Similarly, in multivariate form, with a single dependent variable y and $x = (x_1, x_2 \ldots x_p)$ a vector of dimension p corresponding to the physiographic/climatic variables, the nonparametric density function is

$$\tilde{f}(x,y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_y} K\left(\frac{y-y_i}{h_y}\right) \prod_{l=1}^{p} \frac{1}{h_l} K\left(\frac{x-x_{li}}{h_l}\right) \tag{16}$$

where l is the dimension counter going up to p variables.

Incorporating the multivariate version of $f(x)$ from Eq. (16), in nonparametric form, the equation for the optimal estimate from a regression of Y on X becomes (Muller 1988)

$$\frac{\sum_{i=1}^{n} y_i \prod_{l=1}^{p} \frac{K((x_i-x_{li})/h_l)}{h_l}}{\sum_{i=1}^{n} \prod_{l=1}^{p} \frac{K((x_l-x_{li})/h_l)}{h_l}} \tag{17}$$

with the variables as defined earlier.

To obtain the values of the smoothing factor h_l, least squares cross-validation is once again used. A risk function, similar to Eq. (11) but in a multivariate form, must be developed. Its derivative with respect to the smoothing factors must equal zero, resulting in an equation to be solved for all p values of h_l. This equation is (Adamowski and Feluch 1991)

$$\frac{-n}{2} + \frac{1}{2} \sum_{l=1, i \neq j}^{n} \exp\left(\frac{d_{ij}}{4}\right) \left[\left(1 - \frac{4n\sqrt{2}}{n-1} \exp\left(\frac{d_{ij}}{4}\right)\right) \left(\frac{x_{il}-x_{jl}}{h_l^{-1}}\right) \left(\frac{x_{il}-x_{jl}}{h_l^{+1}}\right) - 1\right] = 0 \tag{18}$$

where

$$d_{ij} = -((x_{il}-x_{jl})/h_l)^2$$

An important observation is that the nonparametric approach does not require the assumptions of linearity, constant variance and normality for the distribution of errors about the regression line, which rarely applies in regional flood relationship development based on linear regression.

One concern in the use of nonparametric regression is related to the sample size available for regional analysis. It has been shown (Silverman 1986) that for density estimation the sample size required to obtain a given mean-square-error increases substantially as the number of independent variables increases. Although most regional equations rarely involve more than two significant variables, there exist many applications of multiple regression in hydrology involving more than two significant variables (Haan 1977). The results in this article apply to these cases as well.

Numerical Analysis

Simulations

In order to compare parametric and nonparametric regression, data from three models (linear: $y \equiv 3x + 1 + \epsilon$; quadratic: $y = x^2 + 1 + \epsilon$; exponential: $y = e^{0.8x} + \epsilon$) were generated. The simulated data consisted of sets of 30 data points corresponding to an independent variable x varying from 0.1 to 3.0 by increments of 0.1. The errors ϵ were generated from a uniform distribution ranging from either -0.1 to 0.1 or from -0.5 to 0.5, to be referred to as the small and large errors respectively.

Five hundred sets of 30 data points were generated from the three models with small and large errors. A stepwise polynomial model (either $y \equiv a_0 + a_1x$ or $y = a_0 + a_1x + a_2x^2$), an exponential model ($y \equiv ae^{bx}$), and nonparametric regression were fitted to each set of 30 data points. Because a logarithmic transformation was employed to fit the exponential model, it is understood that the estimates will be biased (McCuen *et al.* 1990). The bias, variance, and root-mean-square-errors were computed in each case at locations $x = 0.7$, 1.5 and 2.3, resulting in Tables 1, 2 and 3. The first order polynomial model was correct for the linear data, the second order polynomial was correct for the quadratic data, while the exponential model was correct for the exponential data.

An estimate of the bias is given by

$$B = \frac{\sum \hat{y}_i}{n} - y_{\text{true}_i} \tag{19}$$

Where y_{true}, is the true value of the dependent variable at a given x_i based on the chosen data generating model, and $\hat{y}_i$ is the value predicted by the regression for that same x_i. The sample size, n, is equal to five hundred.

The sample variance V is defined as

$$V = \frac{\sum (\hat{y}_i - \sum \hat{y}_i/n)^2}{n-1} \tag{20}$$

Table 1 – Bias from Simulations

Bias at		x=0.7			x=1.5			x=2.3	
	PM	EM	NR	PM	EM	NR	PM	EM	NR
L1	0.00008	0.20463	-0.00057	0.00016	-1.69341	-0.00073	0.00024	-3.59144	0.00085
L2	0.00044	0.19217	0.04107	0.00081	-1.70262	0.00101	0.00118	-3.59144	0.02028
Q1	0.00007	0.97274	0.03412	-0.00030	-0.10943	0.03728	0.00012	-2.47160	0.03847
Q2	0.00035	0.95912	0.01446	-0.00150	-0.11289	0.02412	0.00061	-2.46490	0.03254
E1	-0.14202	-0.19036	0.02160	-0.04743	-1.12016	0.04172	0.20415	-3.45996	0.07699
E2	-0.14173	-0.20324	0.01252	-0.04862	-1.12551	0.03979	0.20463	-3.45777	0.09425

L1 – $y=3x+1+\epsilon_1$ L2 – $y=3x+1+\epsilon_2$ ϵ_1 – U(-0.1,0.1)
Q1 – $y=x^2+1+\epsilon_1$ Q2 – $y=x^2+1+\epsilon_2$ ϵ_2 – U(-0.5,0.5)
E1 – $y=e^{0.8x}+1\epsilon_1$ E2 – $y=e^{0.8x}+\epsilon_2$

Table 2 – Variance from Simulations

Variance		x=0.7			x=1.5			x=2.3	
at	PM	EM	NR	PM	EM	NR	PM	EM	NR
L1	0.00021	0.00024	0.00055	0.00010	0.00014	0.00055	0.00019	0.00007	0.00058
L2	0.00533	0.00623	0.09676	0.00247	0.00355	0.09562	0.00484	0.00166	0.09215
Q1	0.00021	0.00008	0.00200	0.00022	0.00001	0.00048	0.00020	0.00002	0.00050
Q2	0.00532	0.00205	0.07419	0.00553	0.00026	0.11931	0.00510	0.00055	0.23834
E1	0.00021	0.00010	0.00211	0.00022	0.00002	0.00046	0.00020	0.00001	0.00049
E2	0.00532	0.00244	0.07671	0.00553	0.00054	0.12159	0.00510	0.00021	0.15363

L1 – $y=3x+1+\epsilon_1$ L2 – $y=3x+1+\epsilon_2$ ϵ_1 – U(-0.1,0.1)
Q1 – $y=x^2+1+\epsilon_1$ Q2 – $y=x^2+1+\epsilon_2$ ϵ_2 – U(-0.5,0.5)
E1 – $y=e^{0.8x}+1\epsilon_1$ E2 – $y=e^{0.8x}+\epsilon_2$

Table 3 – Root-Mean-Square-Errors from Simulations

RMSE's		x=0.7			x=1.5			x=2.3	
at	PM	EM	NR	PM	EM	NR	PM	EM	NR
L1	0.0146	0.2052	0.0234	0.0100	1.6935	0.0235	0.0139	3.5915	0.0241
L2	0.0730	0.2078	0.3138	0.0498	1.7037	0.3092	0.0696	3.5976	0.3042
Q1	0.0146	0.9728	0.0562	0.0149	0.1095	0.0432	0.0143	2.4716	0.0445
Q2	0.0730	0.9602	0.2728	0.0743	0.1140	0.3463	0.0714	2.4650	0.4893
E1	0.1428	0.1906	0.0507	0.0497	1.1202	0.0469	0.2047	3.4600	0.0801
E2	0.1594	0.2092	0.2772	0.0888	1.1257	0.3510	0.2167	3.4578	0.4031

L1 – $y=3x+1+\epsilon_1$ L2 – $y=3x+1+\epsilon_2$ ϵ_1 – U(-0.1,0.1)
Q1 – $y=x^2+1+\epsilon_1$ Q2 – $y=x^2+1+\epsilon_2$ ϵ_2 – U(-0.5,0.5)
E1 – $y=e^{0.8x}+1\epsilon_1$ E2 – $y=e^{0.8x}+\epsilon_2$

PM – polynomial model; EM – exponential model; NR – nonparametric regression

The estimate of the root-mean-square-error RMSE is defined as

$$RMSE = (V + B^2)^{0.5} \quad (21)$$

Table 3 reveals that, when fitting a correct polynomial model to the generated data, as is the case of the first order linear model for data coming from $y = 3x + 1 + \epsilon$ or the second order polynomial for data coming from $y = x^2 + 1 + \epsilon$, the RMSE's are uniformly less for the parametric fit over the range of x. While both the correct parametric and nonparametric regressions have small biases, the larger variance of the nonparametric regression leads to a larger RMSE.

Fitting an exponential model to any of the generated data sets resulted in large RMSE's because of the large bias present due to the logarithmic transformation. This is even the case for data generated from the exponential function $y = e^{0.8x} + \epsilon$. Even though the presence of bias is known (McCuen *et al.* 1990), logarithmic transformations are nonetheless still commonly employed in hydrology.

Because of the poor fit of the exponential model, as shown in Table 3, either nonparametric regression or the polynomial fit resulted in lower RMSE's for the data generated from an exponential function. For the generated exponential data with small errors, the nonparametric regression, which always exhibits low bias, followed the pattern of the data and resulted in the lowest RMSE's.

Because nonparametric regression is sensitive to variation around the points due to its local character, it exhibited a large variance under large errors. As a consequence, as shown in Table 3, an incorrect but not entirely inappropriate second order polynomial model provided the lowest RMSE's for the generated exponential data with large errors.

Because nonparametric regression always provides an estimate with low bias, when the variation around the data points is small, it will provide a better estimate in terms of mean-square-errors except for the correct parametric model, as long as the latter can be fitted in an unbiased fashion. When the variation around the points is large, nonparametric regression will tend to follow the data points too closely, resulting in large mean-square-errors.

In hydrology, the variations around the regression line are often small, especially after a logarithmic transformation. Therefore, if an appropriate parametric model can be selected for the data, a parametric regression with parameters estimated from that data is preferable to a nonparametric regression. If selection of a correct model is difficult or impossible, then nonparametric regression should be considered. The selection of the appropriate model when one independent variable is involved can be accomplished by a visual fit of the model to the data as well as by residual analysis.

In any number of dimensions, one way to determine the appropriateness of a postulated model consists of performing both parametric and nonparametric regressions. Since the nonparametric regression will always exhibit a low bias, if both regressions provide very similar estimates, then most likely the postulated model is appropriate. Otherwise, the postulated model is inappropriate. Therefore, nonpara-

metric regression can also assist in model selection. More sophisticated tests using nonparametric regression exist to assess the performance of parametric models (Azzalini *et al.* 1989).

Regional Relationships

Regional analysis was performed using annual maximum floods from 53 hydrometric stations available in or surrounding the province of New Brunswick in Atlantic Canada. The stations, which had at least 10 years of record and were from natural flow stations or those with slight regulation, are shown in Fig. 1. They are also listed in Table 4 along with their drainage area and mean annual precipitation. Regional relationships were developed on a province-wide basis and for four homogeneous regions found in a earlier regional study (Inland Waters/Lands Directorate, Environment Canada and the New Brunswick Department of Municipal Affairs and Environment 1987).

Because of the great uncertainty involved in estimating floods for return periods much greater than the sample length, estimates of the 50 and 100-year floods for records less than 15 years in length were not included. Thus, the province-wide equations for the 50 and 100-year floods included 45 stations. The smaller homogeneous regions were also similarly reduced.

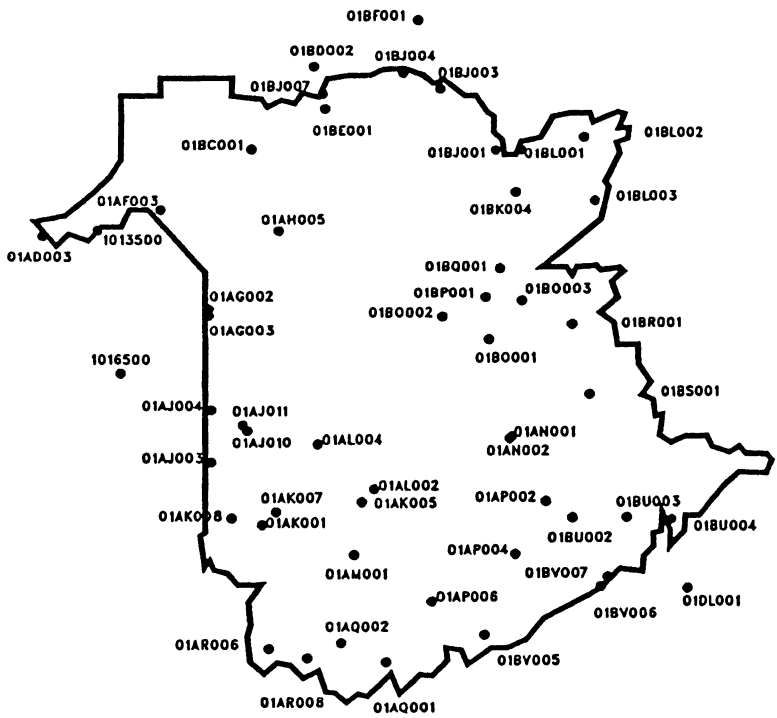

Fig. 1. New Brunswick hydrometric stations.

Table 4 – New Brunswick Hydrometric Station Information

Station Number	Drainage Area (km^2)	Mean Annual Prec.(mm)	Two-Year Flood (m^3/s)
Northwestern			
01AD003	1350	1060	209
01AF003	1150	1070	223
01AG002	199	975	32.9
01AG003	6060	934	906
01AH005	230	1030	41.0
01AJ003	1210	958	241
01AJ004	484	925	90.8
01AK001	234	1120	37.9
01AK007	240	1060	50.1
01AK008	531	1070	70.0
01BC001	3160	1140	575
01BE001	2270	1080	341
01BJ004	88.6	1100	27.6
01BJ007	7740	1120	1470
01BD002	2770	1040	441
01BF001	1140	1060	263
1013500	2290	928	238
1016500	855	943	170
Southern			
01AM001	557	1150	121
01AP004	1100	1190	244
01AP006	293	1140	79.5
01AQ001	239	1240	65.2
01AQ002	1420	1175	227
01AR006	115	1160	25.1
01AR008	43.0	1180	11.7
01BU003	129	1310	38.8
01BU004	34.2	1210	12.5
01BV005	29.3	1410	13.3
01BV006	130	1390	57.9
01BV007	181	1380	84.8
01DL001	63.2	1250	21.7
Central			
01AJ010	350	1130	81.5
01AJ011	156	1100	33.6
01AK005	26.9	1220	5.95
01AL002	1450	1210	295
01AL004	3.89	1230	1.12
01BK004	2090	1010	360
01BO001	5050	1090	840

cont.

Table 4 – New Brunswick Hydrometric Station Information cont.

Station Number	Drainage Area (km^2)	Mean Annual Prec.(mm)	Two-Year Flood (m^3/s)
Central			
01BO002	611	1180	130
01BP001	1340	1180	219
01BQ001	948	1130	182
Eastern			
01AN001	34.4	1180	8.85
01AN002	1050	1130	204
01AP002	668	1040	139
01BJ001	363	988	72.7
01BJ003	510	1050	112
01BL001	175	1010	40.0
01BL002	173	1130	32.6
01BL003	383	1090	69.7
01BO003	484	1080	93.5
01BR001	177	1050	33.8
01BS001	166	1070	45.8
01BU002	391	1030	90.9

The design floods were estimated nonparametrically by kernel density function (Adamowski 1989) and were then subjected to a linear regression using physiographic and climatic parameters such as drainage area, mean annual precipitation, percentage of lakes and swamps, and average water content of snow on March 31 as computed for the basins draining to the hydrometric stations in a previous study (Inland Water/Lands and New Brunswick Department of Municipal Affairs and Environment 1987). The following linear parametric regression models were assumed

$$\log Q = a + b \log x_1 \tag{22}$$

or

$$\log Q = c + d \log x_1 + e \log x_2 \tag{23}$$

where Q is the design flood of a given return period, x_1 and x_2 are physioclimatic variables, while a, b, c, d and e are regression coefficients for the linear regression. For nonparametric regression, a relationship between the logarithms of the variables was found. The most significant factor to enter the equation was always drainage area, with mean annual precipitation always coming second. The addition of further variables was not statistically significant.

On a visual basis, as shown for example by Fig. 2, nonparametric regression and linear regression provide relationships not very different from each other on a province-wide basis where there are 53 data points. Residuals of both the parametric and

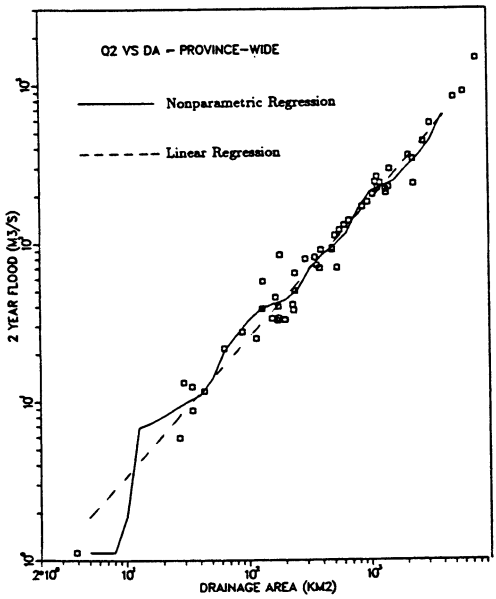

Fig. 2. Province-wide two-year flood linear and nonparametric regressions.

the nonparametric regressions, shown in Figs. 3 and 4 for the one- and two- variable regressions, have no pattern, suggesting a correct relationship. While a lack of pattern is typical of nonparametric regression residuals at all times, it is only typical of parametric regression residuals when an appropriate model has been chosen. The study of the residual plot for a parametric regression helps in detecting structural problems in the model.

For the smaller homogeneous regions, where the number of data points varied from 8 to 18, there is a greater difference between the linear and the nonparametric regressions, as the nonparametric regressions tend to follow very closely the smaller quantity of data points. However, the flood estimates from both approaches tend not to be very different, again implying the adequacy of a linear model.

The above would suggest that the postulated linear model is appropriate. Thus, nonparametric regression for regional relationship development in New Brunswick does not lead to a significant improvement over linear regression and is not adding to the confidence of the latter. In situations where the relationship is nonlinear, however, nonparametric regression has been shown to be an improvement over linear regression (Adamowski and Feluch 1991; Smith 1991).

Screening Ability

The capabilities of nonparametric regression as a screening tool can be illustrated by means of some examples. Fig. 5 shows the nonparametric regression for the 2-year flood of the Eastern region which exhibits a large gap in the data. Linear regression

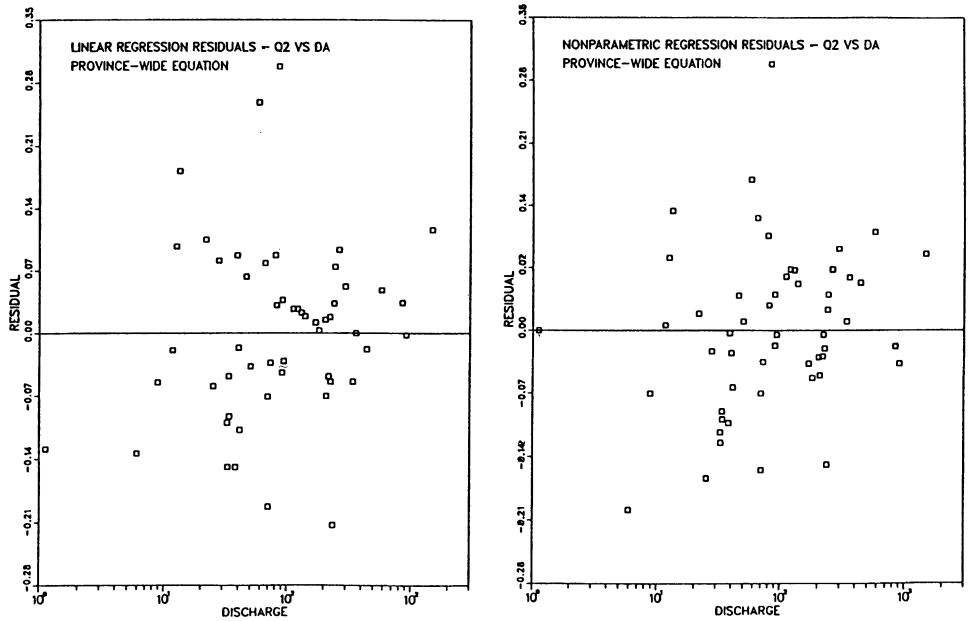

Fig. 3. Residuals for province-wide two-dimensional regressions.

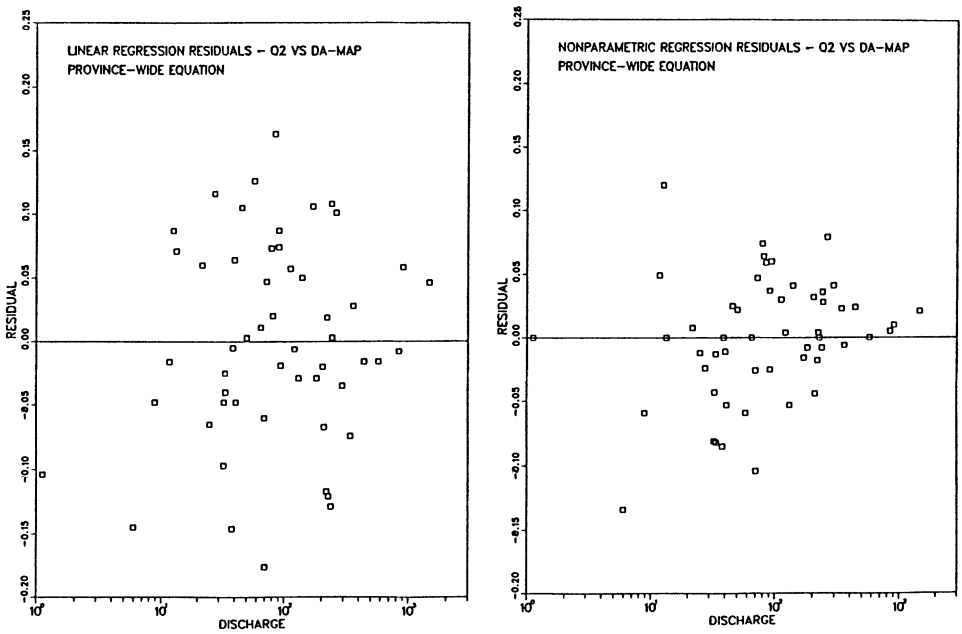

Fig. 4. Residuals for province-wide three-dimensional regressions.

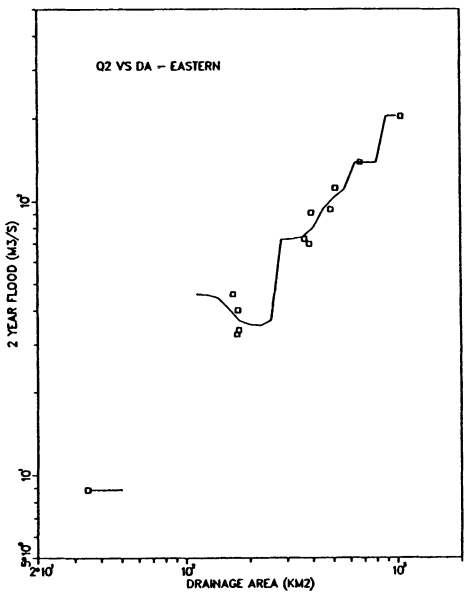

Fig. 5. Eastern Region two-year flood nonparametric regression.

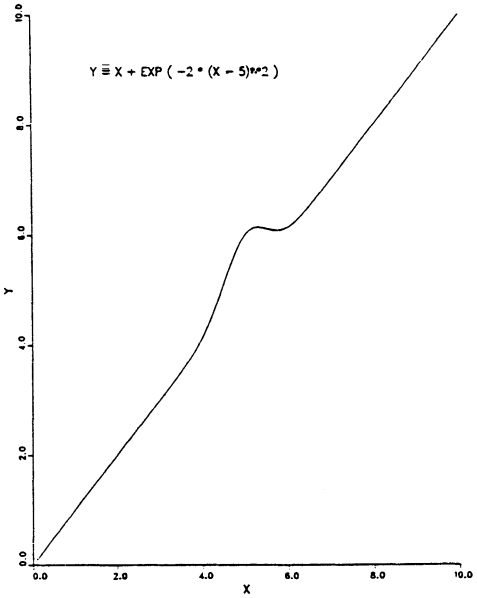

Fig. 6. Linear relationship over a partial range.

imposes a linear relationship over the range of the data independently of any gap which may exist.

Fig. 6 shows a relationship where the variation between the variables is not linear over the entire range. If one had data only for values of x less than 4.0 and greater than 6.0 and were to fit a linear regression, then one might erroneously conclude that the entire relationship was linear. Nonparametric regression, of course, would not yield a correct relationship between 4.0 and 6.0, but by not providing one it would inspire caution in the use of linear regression at that location.

It might be argued further that with sufficient data for values of x ranging from 0.0 to 10.0 in Fig. 6, a nonlinear component would be noticed. If one were unable to fit an adequate nonlinear model for the central portion of the data, then a nonparametric regression might be chosen over a parametric model.

Figs. 7 and 8 show the three-dimensional relationships for, respectively, linear and nonparametric regressions. The linear relationships are parallel, as expected, while the nonparametric relationships meet and cross each other for concurrently low values of drainage area and of mean annual precipitation, and for concurrently very high values of these two variables. Fig. 9 reveals that there are little data with both low drainage area and low mean annual precipitation, or both very high area and precipitation. Thus, linear regression provides an unrealistic assurance in predicting floods for concurrently low and concurrently high values of area and precipitation as it yields a relationship in instances where there exist no data to support it.

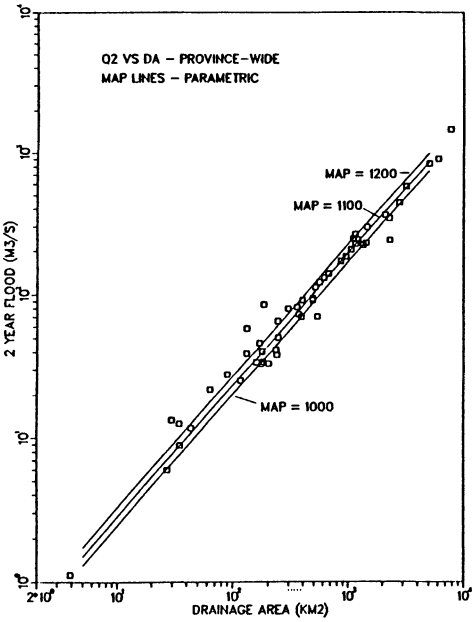

Fig. 7. Province-wide three-dimensional two-year flood linear regression.

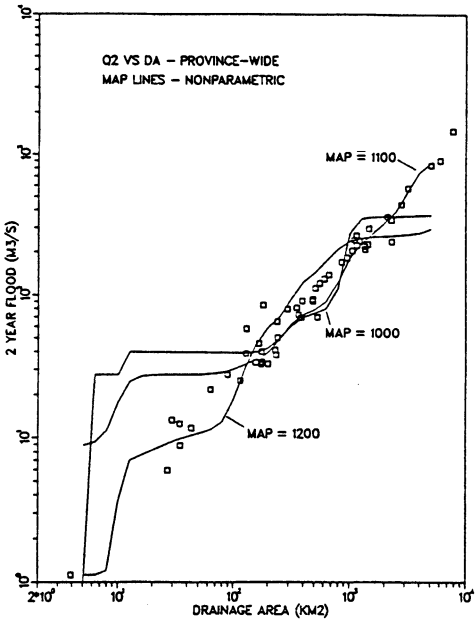

Fig. 8. Province-wide three-dimensional two-year flood nonparametric regression.

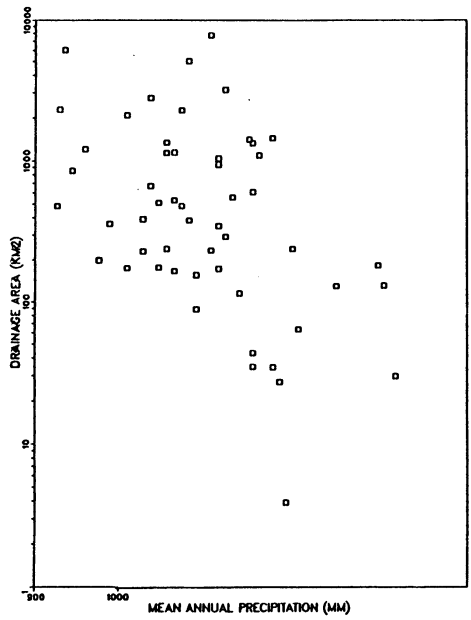

Fig. 9. Drainage area and mean annual precipitation variations.

By use of a scattergram and careful investigations of the ranges of data, the failings of linear regression in the above examples could have been detected without using nonparametric regression as a screening tool. However, when the number of dimensions in a regression increases, such investigations can become very tedious. This could lead to the use of a linear relationship where it should not be employed and to providing the user with an unwarranted assurance in its application. A comparison of linear and nonparametric regressions should quickly point out potential problem areas when both regressions are not very similar.

Conclusion

Simulations showed that if an appropriate parametric regression model can be chosen, its use will lead to lower root-mean-square-errors than a nonparametric regression. If an appropriate model cannot be identified, nonparametric regression should be employed. In regional flood frequency analysis with data from New Brunswick, linear regression and nonparametric regression were found to provide equally good regional relationships. However, nonparametric regression was found to be useful as a screening tool pointing out data deficient relationships.

Acknowledgements

Financial support from Natural Sciences and Engineering Research Council of Canada operating and strategic grants as well as the Ontario Graduate Scholarship is gratefully acknowledged. Valuable comments from two anonymous referees are also acknowledged.

References

Adamowski, K. (1985) Nonparametric Kernel Estimation of Flood Frequencies, *Water Resources Research, Vol. 21 (11)*, pp. 1585-1590.

Adamowski, K. (1989) A Monte Carlo Comparison of Parametric and Nonparametric Estimation of Flood Frequencies, *Journal of Hydrology, Vol, 108,* pp. 295-308.

Adamowski, K., and Feluch, W. (1991) Application of Nonparametric Regression to Groundwater Level Prediction, *Canadian Journal of Civil Engineering, Vol.* pp. 600-606.

Altman, N.S. (1992) An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, *American Statistician, Vol. 46 (3),* pp. 175-185.

Azzalini, A., Bowman, A.W., and Hardle, W. (1989) On the Use of Nonparametric Regression for Model Checking, *Biometrika, Vol. 76,* pp. 1-11.

Bardsley, W.E. (1989) A Simple Parameter-Free Flood Magnitude Estimator, *Hydrological Sciences Journal, Vol. 34 (2),* pp. 129-137.

Cavadias, G.S. (1990) The Canonical Correlation Approach to Regional Flood Estimation. In *Regionalization in Hydrology,* (ed.) M.A. Beran, M. Brilly, A. Brilly, A. Becker and O. Bonacci, Proceedings of Lubljana Symposium, IAHS Publication No. 191, pp. 171-178.

Draper, N.R., and Smith, H. (1966) *Applied Regression Analysis,* John Wiley and Sons.

Gingras, D., and Adamowski, K. (1992) Coupling of Nonparametric Frequency and L-Moment Analyses for Mixed Distribution Identification, *Water Resources Bulletin, Vol. 28 (2),* pp. 263-272.

Gingras, D., and Adamowski, K. (1993) Homogeneous Region Delineation Based on Annual Flood Generation Mechanisms, *Hydrological Sciences Journal, Vol 38 (2),* pp. 103-121.

Guo, S.L. (1991) Nonparametric Variable Kernel Estimation with Historical Floods and Paleoflood Information, *Water Resources Research, Vol. 27 (1),* pp. 91-98.

Haan, C.T. (1977) *Statistical Methods in Hydrology,* Iowa State University Press.

Holder, R.L. (1985) *Multiple Regression in Hydrology,* Institute of Hydrology Wallingford, United Kingdom.

Hosking, J.R.M., and Wallis, J.R., (1993) Some Statistics Useful in Regional Frequency Analysis, *Water Resources Research Vol, 29 (2),* pp. 271-281.

Inland Water/Lands Directorate, Environment Canada and the New Brunswick Department of Municipal Affairs and Environment (1987) Flood Frequency Analyses New Brunswick.

Kite, G.W. (1977) *Frequency and Risk Analysis in Hydrology,* Water Resources Publications, Colorado, U.S.A.

Matyasovszky, I., Bogardi, I., Bardossy, A., and Duckstein, L., (1993) Estimation of Local Precipitation Statistics Reflecting Climate Change, *Water Resources Research, Vol. 29 (12),* pp. 3955-3968.

McCuen, R.H., Leahy, R.B., and Johnson, P.A., (1990) Problems with Logarithmic Transformations in Regression, *ASCE Journal of Hydraulic Engineering, Vol. 116 (3),* pp. 414-428.

Muller, H.G. (1988) Lecture Notes in Statistics Series: *Nonparametric Regression Analysis of Longitudinal Data,* Springer-Verlag.

Prakasa Rao, B.L.S. (1983) *Nonparametric Functional Estimation,* Academic Press.

Rudemo, M. (1982) Empirical Choice of Histogram and Kernel Density Estimation, *Scandinavian Journal of Statistics, Vol. 9,* pp. 65-78.

Scott, D.W., and Terrell, G.R. (1987) Biased and Unbiased Cross-validation in Density Estimation, *Journal of the American Statistical Association Vol. 82,* pp. 1131-1146.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis,* Chapman and Hall.

Smith, J.A. (1991) Long-range Streamflow Forecasting Using Nonparametric Regression, *Water Resources Bulletin, Vol 27 (1),* pp. 39-46.

Tasker, G.D., and Stedinger, J.R. (1986) An Operational GLS model for Hydrologic Regression, *Journal of Hydrology, Vol, 111,* pp. 361-375.

First received: 15 February, 1994
Revised version received: 3 October, 1994
Accepted: 19 October, 1994

Address:
Denis Gringras and Kaz Adamowski, Civ.Eng.Dept.,
Mayer Alvo, Dept. of Math.,
University of Ottawa,
P.O.Box 450 Stn.A.,
Ottawa, On. K1N 6N5, Canada.