Daily flow forecasting with regression analysis

Huynh Ngoc Phien*, Bui Khanh Huong and Phan Dinh Loi
Asian Institute of Technology, PO Box 2754, Bangkok 10501, Thailand

Abstract

Five commonly used regression methods, namely ordinary least squares, ridge, principal components, stepwise, and least absolute value regressions, were considered in this study in the context of daily flow forecasting. From applications to two catchments located in the Lower Mekong Basin, it was found that:

- Ordinary least squares, stepwise, and least absolute value regressions have very good and comparable forecasting capability which is better than that of the remaining two methods.
- Stepwise regression and least absolute value regression are respectively the least and most time-consuming methods.

Having a very good performance, requiring the least computing time, and resulting in simpler equations, stepwise regression is the best among the considered methods.

Introduction

Regression analysis methods have been used quite extensively in river flow forecasting. In many countries, some forms of multiple linear regression are used to forecast the flow at a (downstream) station expressed as a function of the flows at upstream stations. In the literature, the work of Nash and Barsi (1983), Liang and Nash (1988), Phien and Lee (1986), Phien *et al.* (1988a,b) among many others, confirms this popularity. However, in almost all cases, a particular method has been adopted for use without giving any evidence why such a form had been chosen. In other words, the question of appropriateness of the selected method has not been dealt with.

In this study, an evaluation of the most commonly used regression methods in river flow forecasting was made. These comprise ordinary least squares regression (OLS), ridge regression (RIR), principal components regression (PCR), stepwise regression (STR) and least absolute value regression (LAV). For this comparative study, relevant data at two stations, (one at Nam Ngum Dam Site in Laos and the other at Ban Chot of the Chi River Basin in Thailand) were used.

Regression methods

A very brief description of the regression methods to be used is given in this section. More details can be found in Draper and Smith (1981).

Ordinary least squares regression

Suppose that the dependent variable Y can be expressed as a linear function of m predictor variables $X_1, X_2,...,X_m$:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_m X_m + \epsilon \quad (1)$$

where $\beta = [\beta_0\ \beta_1 ... \beta_m]'$ (2)

is the vector of regression coefficients to be estimated (with the prime (') denoting the transpose of a vector or a matrix) from a given set of data points $\{(x_{i1},...,x_{im}y_i),\ i=1,..,n, >m+1)\}$.

Each observation $(x_{i1},...,x_{im},y_i)$ satisfies the equation:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im} + \epsilon_i \\ &= b_0 + b_1 x_{i1} + ... + b_m x_{im} + e_i \end{aligned} \quad (3)$$

where ϵ_i and e_i are the random error and residual, respectively, associated with the response y_i, and

$$b = [b_0\ b_1 ... b_m]' \quad (4)$$

is the estimate vector of β, consisting of the estimates $b_0, b_1,..., b_m$ of $\beta_0, \beta_1,..., \beta_m$, respectively.

By minimising the sum of square errors:

$$SSE = \sum_{i=1}^{n} e_i^2 \quad (5)$$

one obtains

$$b = (X'X)^{-1} X'Y \quad (6)$$

where X is the following nx(m+1)-matrix:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1m} \\ 1 & x_{21} & x_{22} & ... & x_{2m} \\ . & & & & \\ 1 & x_{n1} & x_{n2} & ... & x_{nm} \end{bmatrix} \quad (7)$$

Ridge regression

In the least squares method, if there is an excessive amount of multicollinearity among the predictor variables, the matrix X'X approaches a near singular condition. In this case, the least squares method still yields unbiased estimators for the regression coefficients, but their variances can be very large. One solution to the problem is to abandon the least squares method and accept biased estimation methods. Ridge regression is a method to obtain the estimates by minimising the sums of square errors for the model:

*To whom all correspondence should be addressed.
Received 12 October 1989; accepted in revised form 20 February 1990

$$Y = Xb^* + e \ , \ e = [e_1 e_2 \ldots e_n]' \quad (8)$$

subject to the following constraint:

$$\sum_{j=0}^{m} b^{*2}_j = C \quad (9)$$

where C is a positive constant.

By means of a Langrange multiplier λ, the ridge regression estimates can be shown to be given by:

$$b^* = (X'X + \lambda I)^{-1} X'Y \quad (10)$$

where $\lambda > 0$ and I is the identity matrix of order $k = m+1$. Clearly, the ridge estimates depend on the ridge parameter (λ) and they coincide with least squares estimates when $\lambda = 0$. In this study, the optimum value of λ is determined using the method proposed by Lee (1987).

Principal component regression

When multicollinearity exists among the predictor variables, instead of ridge regression, one can employ regression on principal components. In this case, the predictor variables are transformed into a set of independent variables known as the principal components. These components have been arranged according to the portions of variation in the predictor variables explained by them, with the first component accounting for the largest variation. In other words, from m predictors $X_1, X_2, \ldots, X_m$ in which multicollinearity may exist, one obtains m independent components denoted $w_1, \ldots, w_m$ where all the variation in the X_j, $j = 1,\ldots,m$, is fully explained by the w_j, $j = 1,\ldots,m$, with w_1 accounting for the largest portion, followed by w_2, and so on. Since these principal components are independent, the ordinary least squares method can readily be applied.

In most cases, only a small number of these components can account for a very large portion of the total variation. As such, one may consider only these principal components and the number of independent variables can significantly be reduced. In the present study, all the components which contribute to 90 per cent of the total variation were used in the regression analysis.

Stepwise regression

Stepwise regression method is a standard procedure for searching for the optimum subset of predictor variables. It begins with the smallest subset of predictor variables consisting of only one variable and subsequently increases the number of variables in the equation until no further inclusion is possible. It comprises both forward selection and backward elimination in every step to ensure that only predictor variables which contribute significantly are entered and retained. More information can be found in many statistical textbooks, particularly in Draper and Smith (1981).

Least absolute value regression

Let $\bar{y}$ and $\bar{x}_j$, $j = 1,\ldots,m$ denote the sample means of Y and X_j:

$$\bar{y} = \sum_{i=1}^{n} Y_i/n$$

$$\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$$

then the linear programming formulation of this regression technique is as follows:

$$\text{minimise } Z = \sum_{i=1}^{n} (e_i^+ + e_i^-) \quad (11)$$

subject to:

$$\sum_{j=1}^{m} (b_j^+ + b_j^-)(x_{ij} - \bar{x}_j) + e_i^+ + e_i^- = y_i - \bar{y}$$

$$e_i = e_i^+ - e_i^- \ ; e_i^+, e_i^- \geq 0 \qquad i = 1,\ldots,n$$

$$b_i = b_i^+ - b_i^- \ ; b_i^+, b_i^- \geq 0 \qquad j = 1,\ldots,m$$

$$b_0 = \bar{y} - \sum_{j=1}^{m} b_i \bar{x}_i$$

This formulation makes sure that the centroid $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m, \bar{y})$ lies in the hyperplane determined by the equation:

$$Y = b_0 + b_1 X_1 + \ldots + b_m X_m$$

The details of the solution technique can be found in Narula and Wellington (1977), while the properties of the LAV estimators are given in Dielman and Pfaffenberger (1982).

Models considered

Two models were considered in this study.

The standard linear regression (SLR) model

In this model, the discharge on day t+L is expressed as a linear function of the rainfall data on days t, t-1,... and its past values:

$$Q_{t+L} = Q(t+L) = A + \sum_{j=0}^{r} a_j Q(t-j) + \sum_{j=0}^{s} b_j R(t-j) \quad (12)$$

where A, a_j, b_j are regression coefficients,
r, s are maximum lags in the discharge and rainfall, respectively, and
L is the forecasting lead time (in days).

The extended linear perturbation model

Following the idea introduced by Nash and Barsi (1983), Phien and Lee (1986) developed the extended linear perturbation (ELP) model, particularly useful for flow forecasting. When based upon the model of Eq. 12, the ELP model can be expressed as:

$$U_{t+L} = U(t+L) = A + \sum_{j=0}^{r} a_j U(t-j) + \sum_{j=0}^{s} b_j V(t-j) \quad (13)$$

In this case, U and V are the variables defined as follows:

$$U_i = Q_i - q_i \qquad V_i = R_i - r_i \quad (14)$$

in which q_i and r_i are respectively the daily means of discharge and rainfall data computed from the record used. Explicitly, if N years of records are variable, then:

$$q_i = (1/N) \sum_{k=1}^{N} (Q_i)_k \quad (15)$$

$$r_i = (1/N) \sum_{k=1}^{N} (R_i)_k$$

where $(Q_i)_k$ and $(R_i)_k$ denote, respectively, the discharge and rainfall on day i of the kth year.

The ELP model differs from the model of Nash and Barsi (1983) in that it incorporates the lagged variables $U_t, U_{t-1}, \ldots, U_{t-r}$. Such incorporation can greatly improve the performance of the hybrid scheme introduced by Nash and Barsi (1983), as evidenced by the studies by Phien and Lee (1986) and Phien *et al.* (1988a).

In the following analysis, the lead time was set equal to 1. This means that only forecasting one day in advance was considered.

Case studies

Data employed

Two reliable data sets from two catchments located in the Lower Mekong Basin were given by the Mekong Secretariat for use in this study.

- Nam Ngum Dam site: For this station (catchment area = 8 460 km²), data on the discharge and rainfall are available for 4 years and 11 months, from April 1, 1966 to February 28, 1971. The first four years of data were used for "model calibration", whereby all the regression coefficients were estimated, while the data in the remaining period of 11 months were used for model validation.

- Ban Chot Station: This station (with catchment area of 10 200 km²) is in the Nam Chi River basin, Thailand. Both rainfall and discharge data are available for 11 years (from May 1, 1975 to April 30, 1986). The data for the first eight years were used in model calibration, and those for the remaining three years were used in model validation.

In both cases, the rainfall data employed were computed as the arithmetic means of daily rainfall of all stations located within the corresponding catchment area.

With the forecasting lead time L = 1 day, it was found, from the autocorrelation function of the discharge and cross-correlation function of discharge and rainfall that the values of r and s are 24 and 8, respectively.

Performance indices

The commonly used indices are defined below.

- ***The mean relative error:*** $$MRE = (1/T) \sum_{i=1}^{T} (Q_i - F_i)/Q_i \quad (15)$$

- ***The mean absolute deviation:*** In this study, the MAD statistic is defined as

$$MAD = [1/T\bar{Q})] \sum_{i=1}^{T} |Q_i - F_i| \quad (16)$$

- ***The mean square error:*** $$MSE = (1/T) \sum_{i=1}^{T} (Q_i - F_i)^2 \quad (17)$$

In these equations, T denotes the length of the forecasting period (in days) considered, Q_i and F_i denote, respectively, the observed and forecast discharges on day i, and $\bar{Q}$ is the mean daily discharge over the entire forecasting period:

$$\bar{Q} = (1/T) \sum_{i=1}^{T} Q_i \quad (18)$$

In connection with the MSE, another index, namely the root mean square error with respect to the mean (RMSEM) is also in common use:

$$RMSEM = (MSE)^{1/2}/\bar{Q} \quad (19)$$

- ***The efficiency index***: This index is extensively used. It is computed from the following equation

$$E = (S_0 - S_1)/S_0 \quad (20)$$

where S_0 is the total variation in Q:

$$S_0 = \sum_{i=1}^{T} (Q_i - \bar{Q})^2$$

and S_1 is the sum of square errors

$$S_1 = \sum_{i=1}^{T} (Q_i - F_i)^2 = T^* MSE$$

The MRE indicates the bias involved in the forecasting, and a perfect forecasting model gives rise to a zero value for MRE. The MSE and its accompanying indices like RMSEM and efficiency (E), are closely connected with the least squares approach, which weighs the performance of a model according to the square error $e^2_i = (Q_i - F_i)^2$. However, when there are outliers in the observed data, it may not be appropriate to attach much weight to them. As such, the MAD would be more appropriate as it is more resistant to possible outliers. It is obvious that for the same data set the ordinary least squares (OLS) method will give rise to the smallest values for the MSE, RMSEM and the largest value for the efficiency E, while the least absolute value (LAV) regression will give rise to the smallest value of MAD.

Results and discussions

The results obtained for the first station (Nam Ngum) corresponding to the SLR model are summarised in Table 1, while those obtained for the ELP model are shown in Table 2. For the second station (Ban Chot), the results are shown in Table 3 and Table 4, for the SLR and ELP models, respectively.

General observations

- In all cases, the values of the mean relative error (MRE) are very small, indicating that all the forecasting equations obtained by the SLR and ELP models are almost unbiased.

TABLE 1
RESULTS OF THE STANDARD LINEAR REGRESSION MODEL FOR NAM NGUM DAM SITE

Index	OLS	Method RIR	PCR	STR	LAV
	Calibration				
MRE	-0,0610	-0,1757	-0,0952	-0,0592	-0,0102
MAD	0,1679	0,2118	0,2472	0,1673	0,1529
RMSEM	0,3648	0,4148	0,5141	0,3687	0,3846
E	0,9298	0,9092	0,8606	0,9283	0,9220
MSE	15178	19617	30318	15504	16865
	Validation				
MRE	-0,0376	-0,1204	-0,0609	-0,0379	-0,0004
MAD	0,1373	0,1802	0,2160	0,1340	0,1274
RMSEM	0,3281	0,3700	0,4922	0,3275	0,3372
E	0,9360	0,9186	0,8559	0,9362	0,9324
MSE	21806	27733	49079	21732	23033

Ridge parameter (for RIR) = 0,0048

TABLE 2
RESULTS OF THE EXTENDED LINEAR PERTURBATION MODEL FOR NAM NGUM DAM SITE

Index	OLS	Method RIR	PCR	STR	LAV
	Calibration				
MRE	-0,0081	0,0419	-0,0078	-0,0096	-0,0240
MAD	0,1562	0,2220	0,2276	0,1569	0,1486
RMSEM	0,3141	0,3920	0,4359	0,3171	0,3239
E	0,9479	0,9189	0,8997	0,9469	0,9446
MSE	11262	17545	21696	11480	11890
	Validation				
MRE	0,0042	0,0701	0,0025	-0,0009	-0,0144
MAD	0,1591	0,2282	0,2397	0,1592	0,1571
RMSEM	0,3499	0,4215	0,5230	0,3529	0,3569
E	0,9272	0,8943	0,8374	0,9260	0,9242
MSE	24805	35993	55401	25222	25805

Ridge parameter (for RIR) = 0,0095

- In terms of the MAD and the efficiency (E), one may say that all the resulting equations have a satisfactory performance. With respect to E, all the resulting equations can explain more than 85 per cent of the variation in the daily discharge ($E > 0,85$). In fact, except for the method of regression on principal components (PCR), the value of E exceeds 0,90. As compared to the overall mean discharge, the root mean square error is within 50 per cent in most cases, as indicated by the values of the RMSEM.

- It has been frequently observed that the performance indices have "better" values for the calibration stage than for the validation stage. However, this is **not true all the times** as revealed by several indices employed in the present work. For example, in terms of the MRE, MAD, RMSEM and E, the results obtained for the SLR model by the ordinary least squares (OLS) method for the validation stage are better than those corresponding to the calibration stage. This applies to both stations (Tables 1 and 3). The same observation can also be made for the STR and LAV methods.

Performance evaluation

By examining Tables 1 to 4, one can observe the following:

- As mentioned previously, the OLS method has the best performance during the calibration stage, in terms of the indices which are closely linked to that method, namely E, MSE and RMSEM. Likewise, the LAV regression has the best perfor-

TABLE 3
RESULTS OF THE STANDARD LINEAR REGRESSION MODEL FOR BAN CHOT (NAM CHI)

Index	OLS	Method RIR	PCR	STR	LAV
	Calibration				
MRE	-0,0884	-3,8763	-0,0665	-0,1018	-0,0109
MAD	0,0509	0,2631	0,0949	0,0502	0,0379
RMSEM	0,1994	0,5120	0,2842	0,2009	0,2216
E	0,9911	0,9413	0,9819	0,9910	0,9890
MSE	145	956	294	147	179
	Validation				
MRE	-0,0584	-2,9833	-0,0351	-0,0655	-0,0052
MAD	0,0491	0,3218	0,0865	0,0482	0,0336
RMSEM	0,0877	0,6065	0,1362	0,0880	0,0808
E	0,9963	0,8239	0,9911	0,9963	0,9969
MSE	15	704	35	14	12

Ridge parameter (for RIR) = 0,0001

TABLE 4
RESULTS FOR THE EXTENDED LINEAR PERTURBATION MODEL FOR BAN CHOT (NAM CHI)

Index	OLS	Method RIR	PCR	STR	LAV
	Calibration				
MRE	-0,2780	0,2823	-0,0472	-0,0408	-0,0142
MAD	0,0539	0,2532	0,0955	0,0537	0,0462
RMSEM	0,1876	0,4461	0,2665	0,1892	0,1993
E	0,9921	0,9555	0,9841	0,9920	0,9911
MSE	128	725	259	130	145
	Validation				
MRE	0,0297	0,8117	0,0459	0,0131	0,0027
MAD	0,0627	0,3812	0,1084	0,0627	0,0524
RMSEM	0,1281	0,6267	0,1922	0,1289	0,1247
E	0,9921	0,8120	0,9823	0,9920	0,9926
MSE	31	752	71	31	30

Ridge parameter (for RIR) = 0,0001

mance during the calibration stage in terms of the MAD. However, there is no guarantee that these hold true for the verification stage, because the data used in this stage were not employed in the related optimisation techniques.

- In terms of the indices used, three methods, namely OLS, STR and LAV have comparable performances, which are better than those of the remaining two methods, RIR and PCR.

- The fact that two methods viz. OLS and LAV, perform consistently well indicates that there are no highly influential outliers in the data sets employed (otherwise, different performances would result in the two methods).

- The stepwise regression (STR) method has a similar performance to that of the OLS. This is understood because the procedure employed in the STR is also based upon the maximisation of the portion of variation explained by the resulting equation, which is equivalent to the minimisation of the sum of square errors.

- As shown in Tables 1 to 4, the optimum value of the ridge parameter λ is very small in all cases. This indicates that even with the incorporation of lagged variables for medium-sized catchments (8 460 km^2 and 10 200 km^2), there exists no serious multicollinearity among the predictor variables. However, the allowance for biased estimation (Eq. 10) leads to a significant reduction in the efficiency of the resulting model. In all cases, the results for the RIR are worse than those obtained by the OLS. In fact, they are inferior to those obtained by the STR and LAV also.

- When both the OLS and PCR are applicable, the results obtained by them should be comparable. In the case studies, the results obtained by the PCR are consistently worse than those obtained by the OLS. This is because of the fact that only those components which contribute to 90 per cent of the variation in the discharge have been used. Fortunately, the corresponding reduction in the efficiency is less than 10 per cent. As one of the important reasons for using principal component analysis is the reduction in the dimensionality, it would be unwise to take all the components which are obtained from the original predictor variables.

- In view of the results obtained, it seems that for the data sets employed it is not necessary to adopt the methods intended for use in the existence of multicollinearity. Using these (namely RIR and PCR) would worsen the performance of the resulting forecasting models without any real advantage on compensation.

  In order to have more insight into the performance of the aforementioned regression methods, the execution time of each method for Ban Chot (Nam Chi) is shown in Table 5. Due to the fact that the LAV method consumes too much time on the microcomputer (an APC IV, AT compatible, from NEC), only six years of data were employed for this case. The results collected clearly show that the stepwise regression (STR) method consumes the least time among the five methods considered, followed by the OLS. In all cases, least absolute value regression (LAV) is most time-consuming.

Remarks

- Statistical tests (like the t-test) were used in assessing the significance of each coefficient in the resulting equations for all methods, except the LAV, because the properties of LAV estimators are not so popular (see Dielman and Pfaffenberger, 1982). This means that in the resulting equations for the LAV, some coefficients may not be significantly different from zero. As such the performance of the LAV method, in terms of the indices used, may appear to be better than it actually is.

- With regard to the two methods, namely the standard linear regression and extended linear perturbation models, the results in Tables 1 to 4 show that both have comparable performances in all regression methods.

TABLE 5
EXECUTION TIME OF REGRESSION METHODS FOR THE STANDARD LINEAR REGRESSION MODEL AT BAN CHOT (NAM CHI)

Method	Microcomputer+	Main frame*	
	(1)	(1)	(2)
OLS	2 min 00 s	24 s	1 min 30 s
RIR	5 min 59 s	52 s	1 min 31 s
PCR	4 min 02 s	29 s	2 min 48 s
STR	1 min 42 s	11 s	30 s
LAV	92 min	6 min 20 s	19 min 56 s

Notes: (1) For 6 years of data
(2) For 11 years of data
(+) APC IV (AT compatible)
(*) IBM 3083

Conclusions

This study considered five commonly used regression methods in the context of daily flow forecasting. These are ordinary least squares, ridge, principal components, stepwise and least absolute value regressions. Based on the results obtained from their applications to two reliable data sets along the the standard linear regression model and the extended linear perturbation model, the following conclusions can be drawn.

- The ordinary least squares, stepwise and least absolute value regressions have comparable performance in terms of the commonly used statistical indices. However, least absolute value regression is the most time-consuming method, while stepwise regression is the least time-consuming method.

- For the two catchments considered, which are of medium sizes, the value of the ridge parameter is very small. This indicates that no serious multicollinearity exists among the predictor variables, which include lagged values of both rainfall and discharge. Even so, the bias induced by introducing the ridge parameter may lead to a considerable reduction in the performance of the resulting forecasting equations (obtained by ridge regression).

- Stepwise regression method is the least time-consuming method and has a very good performance. Moreover, the resulting equations involve only the best predictor variables and hence are simple in their form. As such, it is therefore the most suitable method for use in daily flow forecasting.

References

DIELMAN, T and PFAFFENBERGER, R (1982) LAV (Least Absolute Value) estimation in linear regression : A review. In: S Zanakis and J Rustagi (eds.) *Optimisation in Statistics,* North Holland, Amsterdam.

DRAPER, NR and SMITH, H (1981) *Applied Regression Analysis* (2nd edn.), John Wiley and Sons, Inc., New York.

LEE, TS (1987) Optimum ridge parameter selection. *Applied Statistics* **36**(1) 112-118.

LIANG, GC and NASH, JE (1988) Linear models for river flow routing on large catchments. *Journal of Hydrology* **103** 157-188.

NARULA, SC and WELLINGTON, JF (1977) Multiple linear regression with minimum sum of absolute errors. *Applied Statistics* **26**(1) 106-111.

NASH, JE and BARSI, BI (1983) A hybrid model for flow forecasting on large catchments. *Journal of Hydrology* **65** 125-137.

PHIEN, HN and LEE, ST (1986) Forecasting of daily discharges of Burmese rivers. International Journal for Development Technology **4** 173-188.

PHIEN, HN, NGUYEN, VTV and LEE, ST (1988a) Forecasting daily flows of the Mekong River. In: International Water Resources Association: *Water for World Development, Proceedings of the VIth IWRA Congress on Water Resources* **II** 265-273.

PHIEN, HN, AUSTRIACO, NC, PORNPRASERTSAKUL, A and DECHAVICHITLERT, P (1988b) Forecasting of daily discharges for the lower Indus basin during the flood season. *Proc. Sixth Congress of APD-IAHR,* Kyoto. **1** 191-198.