

REGIONALIZATION OF LOW FLOW CHARACTERISTICS USING LOGISTIC AND GLS REGRESSION

Gary D. Tasker

U.S. Geological Survey, 430 National Center, Reston, VA 22092

ABSTRACT Two new methods are described for dealing with practical problems encountered in implementing regional regression models to predict low flow characteristics. First, to determine the probability of encountering a flow of zero at a site as a function of the basin characteristics at the site, a logistic regression is used to relate the probability of zero flow to basin characteristics. A second practical problem in regional regression arises because estimates of low flow characteristics, such as the mean annual minimum flow, at streamgauges are based upon records of different length or at some sites estimates may be based upon a relatively small number of low flow measurements rather than a complete record of flows. In addition, low flows at a site may be autocorrelated resulting in an effectively shorter record. Ordinary least squares regression is not appropriate in such cases. Instead an estimated generalized least squares (EGLS) estimator for the regional regression model is used. This model takes into account differences in accuracy of at-site estimates of low flow characteristics, the sample cross correlation between estimates, and the sample autocorrelation of annual minimum flows.

INTRODUCTION

Low flow characteristics of streams are used in planning and design of water supplies, analyzing environmental and economic impacts, modeling stream water quality, regulating instream uses, and improving the general level of understanding of natural and regulated stream systems. Hydrologists are sometimes called upon to estimate low flow characteristics, such as the 7-day - 10-year low flow, at sites where little or no flow information is available. Often the means for making such estimates is a regional regression model relating flow characteristics at streamgauges to basin characteristics so that one may use the basin characteristics, such as drainage area and soil type, to estimate low flows.

This paper describes two new methods of dealing with practical problems encountered in implementing regional

regression models to predict low flow characteristics in the United States. First, one often encounters minimum flows of zero discharge and would like to determine the probability of encountering a flow of zero at a site as a function of the basin characteristics at the site. This type of problem may be approached as a logistic regression where the probability of zero flow at a site is related to basin characteristics (Cox ,1970).

A second practical problem in regional regression arises because estimates of low flow characteristics, such as the mean annual minimum flow, at streamgauges are based upon records of different length or at some sites estimates may be based upon a relatively small number of low flow measurements rather than a complete record of flows. The accuracy of such measurements vary greatly from site to site. In addition, low flows at a site may be autocorrelated resulting in an effectively shorter record. Ordinary least squares regression is not appropriate in such cases. Instead we use an estimated generalized least squares (EGLS) estimator for the regional regression model which takes into account differences in accuracy of at-site estimates of low flow characteristics, the sample cross correlation between estimates, and the sample autocorrelation of annual minimum flows.

All the techniques are illustrated by a network of gages in central Florida, USA. Logistic regression is used to estimate the probability of zero flow at a site and EGLS regression is used to estimate the mean, standard deviation, and coefficient of skewness of the 7-day annual minimum flows conditioned on the observed flows being nonzero. Combining the logistic regression model with the EGLS model provides a method for estimating the unconditional nonexceedence probability of low flow at an ungauged site. By treating the regional regression model in this way one may avoid the question of what to do with observed responses of zero in a log linear model.

REGIONAL ESTIMATE OF PROBABILITY OF ZERO FLOW

At many streams in the United States the minimum flow for a year is zero. Hydrologists are sometimes called upon to estimate the probability of having a minimum of zero at a site where little or no streamflow data is available. To make such an estimate one may use a regional logistic regression model to relate the probability of the minimum flow at a site being zero to a set of known basin characteristics of the site. Let $1-h_i$ be the probability that the annual minimum (1-day or 7-day) flow at site i is zero. A simple way of representing the dependence of this probability on explanatory variables is the linear logistic model:

$$h_i = \frac{\exp(x_i B)}{1 + \exp(x_i B)} \quad (1)$$

where x_i is a (1xp) row of known basin characteristics at site i , and B is a (px1) column of unknown parameters to be estimated. The B are estimated by using sample data collected at streamgauges in the area to maximize the likelihood function:

$$L(B) = \prod_{i=1}^n \frac{\exp(x_i \hat{B} s_i)}{(1 + \exp(x_i \hat{B}))^{n_i}} \quad (2)$$

where s_i is the number of nonzero minimums observed at streamgauge i and n_i is the total years of record at streamgauge i .

Application to Low Flows in Florida

Data for 55 streams in central Florida were available for analysis (Table 1). The data include the number of years of data, n_i ; the number of years in which the annual minimum 7-day flow was zero, f_i ; and explanatory variables log (base 10) of drainage area, LAREA; log of basin slope, LSLOPE; an index of soil infiltration, SOILS; area of lakes and swamps expressed as a percentage of drainage area, STOR; and a basin shape factor equal to the log of the quantity basin length divided by the square root of drainage area (SHAPE). The logistic regression analysis was performed using BMDP procedure LR, Stepwise Logistic Regression, Dixon (1985). The procedure indicated that the regression coefficients for LAREA, LSLOPE, and SOILS were significantly different from zero (P values <.0001). The regional equation for estimating the probability of zero flow,

$1 - h_i$, is

$$1 - h_i = 1 - \frac{\exp(\hat{y}_i)}{[1 + \exp(\hat{y}_i)]} \quad (3)$$

where $y = -7.6131 + 2.6984 \text{ LAREA} + 3.5281 \text{ LSLOPE} + .7440 \text{ SOILS}$.

REGIONAL ESTIMATES OF LOW FLOW STATISTICS

Low flow at a site is often characterized by an index of low flow, such as the 7-day, 10 year low flow, which is the discharge having a 10-year recurrence interval derived from a frequency curve of lowest average flow for seven consecutive days in a year (Riggs, 1980).

Table 1a. Flow data used for regional regressions for low flows in central Florida. Flow statistics are mean, standard deviation, and coefficient of skewness of the logarithms of observed, nonzero 7-day annual minimum flows.

Station ID	Years of record	Years of zero flow	Mean (Logs)	Standard Deviation (Logs)	Skew. Coeff.
2236500	25	9	-0.6300	0.9180	1.2306
2237000	11	0	0.9590	0.5330	-1.4280
2256000	10	6	0.4100	1.0640	.
2256500	52	35	-0.2120	0.8190	1.1541
2262900	24	0	0.0770	0.6050	-0.2837
2263800	25	4	0.6780	0.8350	-2.3361
2264000	40	31	-0.5100	0.6670	1.6800
2267000	36	0	1.1070	0.3980	-1.8363
2269500	25	1	0.8660	0.5740	-1.4400
2270000	12	0	0.9480	0.2430	-0.4215
2271500	36	0	0.8790	0.4100	-0.4597
2293000	10	6	-1.2360	1.0290	.
2293986	24	0	0.8810	0.3210	-0.1025
2295420	11	0	0.8740	0.4260	0.0046
2296223	31	16	-0.7310	0.4620	-0.5894
2296500	35	0	0.4700	0.5040	-0.4334
2297100	35	1	0.1500	0.5370	-2.3471
2297310	35	0	0.0030	0.7770	-0.8985
2298123	13	0	0.3050	0.6690	-1.9044
2298608	11	3	-0.1260	0.7090	-1.9565
2298830	49	28	0.3530	0.8270	-1.4799
2299750	7	0	0.2030	0.3530	-1.1310
2299950	19	0	0.1140	0.4150	-0.4855
2300000	26	0	0.5910	0.2000	-0.4972
2300100	22	11	-0.7710	0.6070	-0.0587
2300500	46	0	0.9610	0.3290	-0.5686
2301000	35	0	1.5150	0.3110	-1.4139
2301300	22	0	1.0180	0.4180	-0.6427
2301350	13	9	-1.4120	0.2230	.
2301500	53	0	1.6200	0.3310	-0.5110
2301800	16	0	1.1740	0.3410	-2.6029
2301900	19	6	-0.6290	0.6140	-1.4849
2302500	34	3	0.6630	0.3270	-0.7925
2303100	10	8	.	.	.
2303350	11	11	.	.	.
2303400	22	11	-0.3350	0.5240	0.0541
2303420	11	9	.	.	.
2303800	21	18	.	.	.
2307000	32	1	0.1680	0.4450	-0.6863
2307243	9	8	.	.	.
2307323	15	14	.	.	.
2307359	35	30	-0.8820	1.0600	.
2307697	23	16	-0.7260	0.3460	-0.5943
2309848	15	15	.	.	.
2310000	39	0	0.4150	0.2100	1.1342
2310240	21	21	.	.	.
2310300	22	3	-0.1560	0.3850	-3.5921
2310352	7	3	-1.2330	0.5630	.
2310800	25	11	-0.4220	0.6750	0.3200
2310947	16	7	-0.7390	0.7760	-0.6117
2312180	17	15	.	.	.
2312200	25	3	0.2880	0.6790	-2.5098
2312640	20	0	0.8830	0.6140	-3.0316
2314200	22	5	-0.9780	0.3920	-1.1784
2321000	21	0	0.2870	0.3460	-0.4757

NOTE: A "." indicates that not enough nonzero values were observed to compute indicated flow statistic.

Table 1b. Basin characteristics used in regional regressions in central Florida.

Station ID	Log of Area <LAREA>	Log of Slope <LSLOPE>	Soils Index <SOILS>	Storage Index <STOR>	Shape Factor <SHAPE>
2236500	1.8325	-0.1249	2.52	41.9	0.5351
2237000	2.2553	-0.0315	2.66	32.4	0.5664
2256000	2.2742	0.1206	2.22	15.9	0.3139
2256500	2.4928	0.1239	2.15	15.8	0.4547
2262900	1.9222	0.3096	3.08	15.3	0.1912
2263800	1.9504	0.2504	2.28	16.5	0.3431
2264000	1.4814	-0.3872	5.31	35	0.3392
2267000	1.7701	0.2201	4.88	18.7	0.0977
2269500	1.7846	0.4942	4.21	30.6	0.3591
2270000	1.5888	0.6902	4.38	15.6	0.3087
2271500	2.0374	0.5809	5.14	23.2	0.1424
2293000	1.7782	0.2148	2.01	8.57	0.1761
2293986	2.2041	0.1004	5.20	23.5	0.3220
2295420	2.0828	0.5888	2.05	12.7	0.3075
2296223	1.6222	0.0969	2.22	24.9	0.1954
2296500	2.5185	0.2253	2.08	14.2	0.2852
2297100	2.1206	0.6085	2.03	6.93	0.2644
2297310	2.3385	0.4456	2.05	9.70	0.4831
2298123	2.3674	0.4048	2.00	21.8	0.2482
2298608	2.0969	0.7332	2.05	13.3	0.1325
2298830	2.3598	0.3304	1.90	14.2	0.3732
2299750	1.3802	0.5635	2.05	4.01	0.2689
2299950	1.8149	0.7657	2.05	6.19	0.3255
2300000	1.9031	0.6928	2.05	4.46	0.4316
2300100	1.4969	0.7559	2.05	7.63	0.2593
2300500	2.1732	0.7016	2.55	4.41	0.3589
2301000	2.1303	0.6955	2.70	10.6	0.2282
2301300	2.0294	0.5563	2.05	22.3	0.2132
2301350	0.9370	1.2256	2.05	3.68	0.1547
2301500	2.5250	0.5378	2.14	8.55	0.2456
2301800	1.4472	0.3160	5.38	13.4	0.1121
2301900	0.9777	0.8261	2.05	12.6	0.2435
2302500	2.0414	0.5465	2.72	16.4	0.2671
2303100	1.1761	1.1106	3.69	15.1	0.4076
2303350	1.3617	0.6911	2.64	20.8	0.3032
2303400	1.7482	0.6628	2.47	12	0.1104
2303420	2.0682	0.4232	2.73	25	0.2798
2303800	2.2041	0.3222	2.64	27	0.3865
2307000	1.5441	0.6149	2.23	13.7	0.4902
2307243	1.0000	0.3962	2.05	23.2	0.3982
2307323	1.2304	0.4133	2.05	26.5	0.4089
2307359	1.4771	0.4487	2.05	28.7	0.5288
2307697	0.9542	0.9841	2.05	8.23	0.3167
2309848	1.2330	0.4900	2.30	23.7	0.3878
2310000	1.8603	0.5490	2.42	19.9	0.3511
2310240	1.6335	0.1761	4.11	8.55	0.4620
2310300	2.2601	0.2945	4.11	10.9	0.0795
2310352	1.4654	0.4742	5.38	28	0.2889
2310800	2.1139	0.0374	2.52	43.1	0.2558
2310947	2.4472	0.1239	2.00	46.7	0.3450
2312180	1.9294	0.1271	2.00	50.2	0.3271
2312200	2.1614	0.2430	2.71	37.2	0.3171
2312640	1.6021	0.3222	5.38	3.49	0.3722
2314200	1.4150	0.7709	2.00	38.2	0.3219
2321000	2.2856	0.4698	3.65	37.8	0.2075

One means of making an estimate of the 7- day 10- year low flow at a site is to assume a distribution of 7-day annual minimums and estimate the parameters of the distribution. A popular and often useful distribution is the Log-Pearson III (LPIII) distribution which requires estimates of the mean, standard deviation, and coefficient of skewness of the logs of the nonzero 7-day minimum flows. At sites without flow records one may make the estimates using a set of regional regression equations.

The data used to estimate the regional regression coefficients is data collected at streamgauges in the region or at partial-record sites in the region where a number of base flow measurements have been made and correlated with a long record (Stedinger and Thomas, 1985). The accuracy with which one can make an estimate of a low flow index at a site based on recorded or measured flows may vary greatly from site to site. Factors that affect the accuracy of an estimate of the low flow index include length of record at a streamgauge, number of base flow measurements at a partial-record station, natural variability of low flows at the site, and the degree of autocorrelation of annual low flow data. Ordinary least squares regression is not an appropriate technique when the response variable--for example, the 7-day, 10 year low flow is not observed with equal accuracy at all sites used in the regional regression (Weisberg, 1980, p. 73).

Recently Stedinger and Tasker (1985, 1986) documented the value of EGLS procedures for regional regression of streamflow statistics. The EGLS method can be applied in separate regional regressions to develop models for the coefficient of skewness, standard deviation of annual minimums, and mean of annual minimums. After suitable transformation of variables, the EGLS model may be written in as

$$\tilde{Y} = \underline{X} \underline{B} + \underline{e} \quad (4)$$

where \tilde{Y} is a $n \times 1$ vector of flow characteristic at n sites, \underline{X} is an $(n \times p)$ matrix of $(p-1)$ basin characteristics augmented by a column of one's, \underline{B} is a $(p \times 1)$ vector of regression parameters and \underline{e} is an $(n \times 1)$ vector of random errors. The dependent variable, \tilde{Y} , is a flow characteristic, such as the mean of the logs of annual minimums, that is derived from a sample of observed flows. The generalized least squares (GLS) estimator of \underline{B} is

$$\underline{B} = (\underline{X}^T \underline{W} \underline{X})^{-1} \underline{X}^T \underline{W} \tilde{Y} \quad (5)$$

where it is assumed that the errors have zero mean $E[e] = 0$, and covariance $E[ee^T] = \underline{W}$.

The operational difficulty with (5) is that \underline{W} is unknown and must be estimated from the data, thus the name EGLS for estimated GLS. Several assumptions are made to facilitate the estimation of \underline{W} . In OLS, \underline{W} is estimated as

$$\underline{\hat{W}}_{OLS} = \hat{s}_r^2 \underline{I} \quad (6)$$

where \hat{s}_r^2 is the variance of sample residuals and \underline{I} is an $(n \times n)$ identity matrix.

Stedinger and Tasker (1985) proposed that \underline{W} be estimated as

$$\underline{\hat{W}} = u^2 \underline{I} + \underline{V} \quad (7)$$

where u^2 is an estimate of the model error variance due to an imperfect model and \underline{V} is an $(n \times n)$ matrix of sampling covariances.

To apply the EGLS method one must first estimate the elements of \underline{V} . For the regional skew coefficient regression, the elements of \underline{V} can be approximated by

$$(\text{skew})v_{ij} = \begin{cases} \frac{6n_i(n_i-1)(1+6/n_i)^2}{(n_i-2)(n_i+1)(n_i+3)} & \text{for } i=j \\ 0 & \text{for } i \neq j \end{cases} \quad (8)$$

where n_i is the length of record at site i . (Tasker and Stedinger, 1986).

For the regional standard deviation regression, the elements of \underline{V} can be approximated by

$$(\text{std dev})v_{ij} = \begin{cases} \frac{\hat{s}_i^2(1+.75\hat{g}_i^2)}{2n_i} & \text{for } i=j \\ \frac{\hat{r}_{ij}m_{ij}\hat{s}_i\hat{s}_i}{2n_in_i} \left[\hat{r}_{ij}+.75\hat{g}_i\hat{g}_i \right] & \text{for } i \neq j \end{cases} \quad (9)$$

where g_i and \hat{s}_i are regional estimates of coefficients of skewness and standard deviation, m_{ij} is the concurrent record length between site i and site j and \hat{r}_{ij} is the correlation coefficient between the series of annual minimums between sites i and j .

For the regional mean regression, \underline{V} can be approximated by

$$(\text{mean}) v_{ij} = \begin{cases} \hat{s}_i^2/n_i & \text{for } i=j \\ \hat{r}_{ij} m_{ij} \hat{s}_i \hat{s}_j / n_i n_j & \text{for } i \neq j \end{cases} \quad (10)$$

The regression coefficients, \underline{B} , are found by using an iterative search to solve the equation

$$(\underline{Y} - \underline{X} \hat{\underline{B}})^T \hat{\underline{W}}^{-1} (\underline{Y} - \underline{X} \hat{\underline{B}}) = n - p \quad (11)$$

for the model error u^2 and $\hat{\underline{B}}$.

Equations (9) and (10) apply only to those sites that have continuous records. At base flow measurement sites one may substitute the formulas given in Stedinger and Thomas, 1985, for the diagonal elements in (9) and (10).

Annual low flows sometimes exhibit autocorrelation. If this is the case one may use effective record length (Tasker, 1983) in place of n_i .

Application to Florida low flows

Standard computer packages do not allow one to easily use the EGLS method described. However, a set of FORTRAN programs have been developed at the U.S. Geological Survey. These programs were used to estimate regional regressions for skew coefficient, standard deviation, and mean of logs of nonzero low flows in central Florida. The final equations involve only the explanatory variables found to be significantly different from zero at the 5% level. These equations are:

$$\text{skew} = 1.7974 + 3.06812 (\text{SHAPE}) \quad (12)$$

$$\text{standard deviation} = 0.81278 - .39015(\text{LSLOPE}) - .05489 (\text{SOILS}) \quad (13)$$

$$\text{mean} = -2.09244 + 1.03028 (\text{LAREA}) + .22485 (\text{SOILS}) - .01703 (\text{STOR}) \quad (14)$$

CONDITIONAL PROBABILITY ADJUSTMENT

When zero flows occur in a sample, a conditional probability adjustment is made in order to estimate the T-year recurrence interval low flow. The adjustment is discussed in detail in Haan (1977, p. 146). A brief explanation follows: Suppose $F(x)$ is the unconditional probability that flow X does not exceed x . For the 10-year recurrence interval, $F(x) = .1$. Let h_i be the probability of non-zero flow. The value of h_i is estimated at an ungauged site by the regional logistic regression described earlier. Let $F^*(x)$ be the probability that X does not exceed x , conditioned on non-zero values of X . The relationship between $F(x)$ and $F^*(x)$ is $F(x) = 1 - h_i + h_i F^*(x)$ (Haan, 1977, p. 147). The magnitude of an event with return period T is determined by solving for $F^*(x)$ and using the inverse transformation of $F^*(x)$ to obtain x . If the value of $F^*(x)$ is negative, then the estimated T-year flow is zero.

Example--Suppose one would like an estimate of the 7-day, 10-year low flow, X , at a site in central Florida with drainage area of 100, channel slope of 10, soils index of 3.0 shape factor of .5, and storage index of 10. The mean, standard deviation, and coefficient of skewness of the logs of the nonzero flows are estimated from equations 14, 13, and 12, as

$$\text{mean} = -2.09244 + 1.03028(\log 100) + .22485(3.0) - .01703(10) = .4723$$

$$\text{standard deviation} = .81278 - .39015(\log 10) - .05489(3.0) = .2580$$

$$\text{skew coeff.} = -1.7974 + 3.06812(.5) = -.2633$$

The probability of nonzero flow, h_i , at the site is estimated by equation 3, as

$$\hat{h}_i = \frac{\exp \hat{y}}{1 + \exp \hat{y}} = \frac{\exp(3.544)}{1 + \exp(3.544)} = \frac{34.6}{35.6} = .972 \quad (15)$$

$$\begin{aligned} \text{where } \hat{y} &= 7.613 + 2.6984(\log 100) + 3.5281(\log 10) + \\ &.74402 \quad (3.0) \\ &= 3.544 \end{aligned}$$

For the 10-year event $F(x) = .1$

and

$$F^*(x) = \frac{.1 - 1 + .972}{.972} = .074 \quad (16)$$

Therefore,

$$\begin{aligned} \log(x) &= \text{mean} + K^* (\text{standard deviation}) \\ \log(x) &= .4723 - (1.50) (.2580) \\ \log(x) &= .0853 \\ (x) &= 10^{.0853} = 1.2 \end{aligned}$$

where K^* is the Pearson III percentage point (Harter, 1969) associated with a skew coefficient of $-.2633$ and exceedence probability of $.074$.

REFERENCES

- Cox, D.R. (1970) The Analysis of Binary Data, Chapman and Hall, London.
- Dixon, W.J. (1985) BMDP Statistical Software, University of California, Press, Berkely.
- Haan, C.T. (1977) Statistical Methods in Hydrology, The Iowa State University Press, Ames, Iowa.
- Harter, H.L. (1969) A new table of percentage points of the Pearson type III distribution. *Technometrics*, 11(1), 177-186.
- Riggs, H.C. (1980) Characteristics of low flow, J. Hydraulics Div. ASCE 106(HY5), 717-731.
- Stedinger, J.R. and Tasker, G.D. (1985) Regional hydrologic analysis 1. *Wat. Resour. Res.* 21(9), 1421-1432.
- Stedinger, J.R. and Tasker, G.D. (1986) Regional hydrologic analysis 2. *Wat. Resour. Res.* 22(10), 1487-1499.
- Stedinger, J.R. & Thomas, W.O. (1985) Low-flow frequency estimation using base-flow measurements. U.S. Geological Survey open-file report 85-95.
- Tasker, G.D. (1983) Effective record length for the T-year event. *J. of Hydrol.* 64, 39-47.
- Tasker, G.D. & Stedinger, J.R. (1986) Regional skew with weighted LS regression. *J. Wat. Resources Planning and Management ASCE*, 112(2), 225-237.
- Weisburg, S. (1980) Applied Linear Regression, John Wiley and Sons, New York.