A SOLUTION TO SMALL SAMPLE BIAS IN FLOOD ESTIMATION

by

William Metler*

ABSTRACT

In order to design culverts and bridges, it is necessary to compute an estimate of the design flood. Regionalization of flows by regression analysis is currently the method advocated by the U.S. Geological Survey to provide an estimate of the culvert and bridge design floods. In the regression analysis a set of simultaneous equations is solved for the regression coefficients which will be used to compute a design flood prediction for a construction site. The dependent variables in the set of simultaneous equations are the historical estimates of the design flood computed from the historical records of gaged sites in a region. If a log normal distribution of the annual peak flows is assumed, then the historical estimate of the design flood for site i may be computed by the normal as

$$\log Q_{d,i} = \bar{x}_i + k_d \hat{s}_i .$$

However because of the relatively small samples of peak flows commonly used in this problem, this paper shows that the historical estimate should be computed by

$$\log Q_{d,i} = \bar{x}_i + t_{d,n-1}\sqrt{\frac{n+1}{n}}\ \hat{s}_i$$

where $t_{d,n-1}$ is obtained from tables of the Student's t. This t-estimate when used as input to the regression analysis provides a more realistic prediction in light of the small sample size, than the estimate yielded by the normal.

* Graduate, Systems & Industrial Engineering Department, University of Arizona, Tucson, Arizona 85721

INTRODUCTION

In the construction of culverts and bridges, the design is specified by the design flood. If the design flood were the 25 year flood, then the culvert or bridge is to pass that flood which is equaled or exceeded once in 25 years. The prediction of the design flood for a construction site may be found by some method which considers the estimates of the design flood computed from the records at the gaged sites within the region.

Regression analysis is one such method of regionalization which may be used to predict the design flood at a construction site. The USGS (Thomas and Benson, 1970) has assumed that the design flood, a streamflow characteristic, is related to the basin and climatic characteristics for the drainage site by the formula

$$Q_d = aA^{b_1} E^{b_2} Si^{b_3} \qquad (1)$$

for example. In (1) Q_d represents the design flood, A represents the drainage in square miles, E represents the elevation in thousands of feet, and Si represents a soil index figure. By taking logrithms in (1) we have a linear relationship of topographical and climatic characteristics with a streamflow characteristic

$$\log Q_d = b_0 + b_1 \log A + b_2 \log E + b_3 \log Si \quad . \qquad (2)$$

Consider for example that there are m gaged sites in a region where the predictor variables are area, elevation, and soil index. Then

the set of simultaneous equations for the 25 year flood would be

$$\begin{aligned}
\log Q_{25,1} &= b_0 + b_1 \log A_1 + b_2 \log E_1 + b_3 \log Si_1 \\
\log Q_{25,2} &= b_0 + b_1 \log A_2 + b_2 \log E_2 + b_3 \log Si_2 \qquad (3) \\
&\cdot \\
&\cdot \\
&\cdot \\
\log Q_{25,m} &= b_0 + b_1 \log A_m + b_2 \log E_m + b_3 \log Si_m.
\end{aligned}$$

The equations in (3) could then be solved for the b_i's. For a construction site the topographical and climatic characteristics for that site along with the regression coefficients, the b_i's, would provide an estimate of the design flood for the ungaged construction site. The procedure could be repeated and a set of regression coefficients for a region could be computed for the 2,5,10,25,50, and 100 year floods or any other streamflow characteristic. This method of regionalization permits one from the data already on hand to infer a streamflow characteristic, e.g., the 25 year design flood given the readily measurable topographical and climatic characteristics of a particular site.

It is obvious here that the determination of the prediction of the design flood for the construction site depends on the regression coefficients. The regression coefficients depend on the estimates of the design flood computed from the historical record. This paper is concerned with the computation of the historical estimate which must serve as input to the regression analysis.

THE HISTORICAL ESTIMATE

If a log normal distribution of the annual peak flows is assumed for a gaged site i, then the logs of the peak flows are distributed normally. Hence the sample mean for site i is

$$\bar{x}_i = \sum_j \frac{x_{ij}}{n_i} \tag{4}$$

and the sample variance is

$$\hat{s}_i = \sum_j \frac{(x_{ij} - \bar{x}_i)^2}{n_i - 1} \tag{5}$$

where x_{ij} is the log of the discharge for the j^{th} year of record for the i^{th} site and n_i is the length in years of record for the i^{th} site. Thus the historical estimate of the streamflow characteristic, say the 25 year flood, might be computed by

$$\log Q_{25,i} = \bar{x}_i + k_{25} \hat{s}_i \tag{6}$$

where k_{25} is the number of standard normal deviates from the mean and may be obtained from tables of the standard normal distribution for the probability level 1/25 = .04 (k_{25} = 1.75). However, the small sample sizes commonly encountered in streamflow estimation introduces a bias into the estimate (6). It is shown in Appendix A that

$$\log Q_{d,i} = \bar{x}_i + t_{d,n_i-1} \sqrt{\frac{n_i+1}{n_i}} \hat{s}_i \tag{7}$$

provides an estimate from the historical record which removes the bias introduced by the small sample size. In (7) $Q_{d,i}$, $\bar{x}_i$, n_i, and $\hat{s}_i$ are

as previously defined and t_{d,n_i-1} is found in tables of the Student's t with $n_i - 1$ degrees of freedom and the cumulative probability level equal to 1/d where d is the recurrence interval of the design flood. It will be shown in this paper that the estimate as computed by (6) will always underestimate the value computed by (7) for the design floods commonly used in construction specifications. Because the historical estimate must serve as input to the regression analysis for defining the regression coefficients, the regression predictions based on the inputs by (6) will always be less than the regression predictions based on the inputs from (7). The amount of underestimation for the design floods is defined as being significant in terms of percentage error.

EXAMPLE AND RESULTS

The example for this paper is the plains region in Missouri and the sites of record are restricted to those areas of less than 30 square miles. The topographical and climatic characteristics are listed in Table 1 and further details can be found in Homyk (1971). The historical records were provided by the U.S. Geological Survey.

Table 2 gives the coefficient for the sample variance when using (7). This table is used to compute the estimate of a design flood by (7). For example suppose the sample mean is 5.85 in log units and the sample standard deviation is .69 in log units. Then the estimate of the 25 year flood (1/d=1/25=.04) for a record length of 16 years is from Table 2

$$\log_e Q_{25} = 5.85 + (2.099)(.69) = 7.3$$

$$Q_{25} = 1480 \text{ ft}^3/\text{sec.}$$

The normal estimate would be

$$\log_e Q_{25} = 5.85 + (1.75)(.69) = 7.05$$

$$Q_{25} = 1153 \text{ ft}^3/\text{sec.}$$

Thus the percentage error based on the normal is $\frac{1480-1153}{1153} \cdot 100 = 28\%$. That is, the normal because of the small sample size underestimates the 25 year design flood by as much as 28%.

Tables 3,4,5,6,7, and 8 were computed in the following manner. The regression coefficients for a given design flood were computed using the inputs from 30 sites by (6). These coefficients were used to compute a prediction for each of the 30 sites. This was repested using the t-estimates as input and the percentage error of the normal versus the t was computed as listed in Table 9. Table 9 also includes the average percentage error of the normal versus the t estimates computed from the historical record. The regression analysis gives a smoothing effect as the percentage error for a given flood is less for the regression prediction than for the historical estimation.

Table 9 also lists the R^2 for each analysis. R^2 is a number between zero and one; the closer R^2 is to one, the better is the fit of the regression line. It will be noticed that the R^2 for the normal case is slightly better than for the t inputs. It will be also noticed that as the recurrence interval of the flood increases the R^2 decreases. These results indicate that although one measure of our confidence, R^2, in the predictions computed with the t inputs is slightly poorer than the predictions computed by the normal inputs, the t inputs provide a significantly more conservative, i.e.,

realistic, prediction than the normal. The conclusion is that in order to compute an estimate of the design flood for construction projects the estimate should include the effect of the small sample size, and the use of the Student's t accomplishes this task.

TABLE 1

Site index with associated topographical and climatic characteristics

Site Index	Area	Elevation	Forest Cover	Precipitation	2 yrs/24 hrs. Precipitation	Soil Index
54977.0	2.4	.8	4.3	11.0	3.3	2.4
55020.0	31.0	.7	7.9	11.0	3.4	2.5
55136.5	3.1	.6	23.5	11.0	3.4	2.6
68200.0	6.0	1.1	2.0	9.0	3.3	3.2
68945.0	20.0	1.0	11.0	11.0	3.5	3.5
69013.0	.1	.9	14.8	11.0	3.3	2.4
69075.0	16.6	.9	7.1	13.0	3.5	2.8
69102.0	1.0	.8	1.0	13.0	3.4	3.5
54951.0	.7	.6	23.3	11.0	3.3	2.6
55030.0	2.6	.8	7.8	11.0	3.4	2.2
55136.0	1.5	.6	36.3	11.0	3.4	2.6
55142.0	.5	.9	1.0	11.0	3.4	2.2
68160.0	4.9	1.1	6.0	9.0	3.4	3.5
68203.0	1.3	1.1	1.0	9.0	3.3	3.2
68210.0	2.7	1.0	1.0	9.0	3.4	3.1
68961.8	.4	.9	1.0	9.0	3.3	2.4
68965.0	5.6	1.0	6.8	9.0	3.3	2.4
68967.0	.8	1.0	2.5	9.0	3.4	2.4
68972.0	4.7	1.0	17.4	9.0	3.3	2.4
68996.0	.2	.8	1.0	9.0	3.3	2.4
69025.0	2.5	.9	5.4	11.0	3.4	2.4
69028.0	1.0	.8	2.0	11.0	3.4	2.4
69047.0	1.0	.9	21.4	11.0	3.3	2.8
69057.0	.8	.7	5.6	11.0	3.4	2.8
69072.0	1.6	.8	1.2	13.0	3.5	2.3
69083.0	1.0	.8	9.0	13.0	3.5	2.8
69085.0	2.9	.8	2.3	13.0	3.5	2.8
69094.0	.3	.7	6.7	13.0	3.5	3.5
69097.0	.5	.8	4.0	13.0	3.5	3.5
69102.5	.6	.8	11.5	13.0	3.5	3.5

TABLE 2

Values of $t_{d,n-1}\sqrt{\frac{n+1}{n}}$ for use in (7).

Record length	2 yr. flood	5 yr.	10 yr.	25 yr.	50 yr.	100 yr.
6	0.000	1.107	1.658	2.343	2.842	3.354
7	0.000	1.096	1.632	2.282	2.747	3.217
8	0.000	1.088	1.612	2.238	2.677	3.117
9	0.000	1.081	1.596	2.204	2.624	3.041
10	0.000	1.076	1.584	2.177	2.582	2.981
11	0.000	1.072	1.574	2.155	2.549	2.933
12	0.000	1.069	1.566	2.138	2.521	2.894
13	0.000	1.066	1.559	2.123	2.499	2.862
14	0.000	1.063	1.553	2.110	2.479	2.834
15	0.000	1.061	1.548	2.099	2.463	2.810
16	0.000	1.059	1.543	2.090	2.449	2.790
17	0.000	1.057	1.540	2.082	2.436	2.773
18	0.000	1.056	1.536	2.074	2.425	2.756
19	0.000	1.054	1.533	2.068	2.415	2.743
20	0.000	1.053	1.530	2.062	2.406	2.730
21	0.000	1.052	1.528	2.057	2.398	2.719
22	0.000	1.051	1.526	2.052	2.391	2.709
23	0.000	1.050	1.524	2.048	2.385	2.700
24	0.000	1.049	1.522	2.044	2.379	2.692
25	0.000	1.049	1.520	2.040	2.373	2.684
26	0.000	1.048	1.518	2.037	2.368	2.677
27	0.000	1.047	1.517	2.034	2.364	2.671
28	0.000	1.047	1.516	2.031	2.359	2.664
29	0.000	1.046	1.514	2.028	2.356	2.659
30	0.000	1.046	1.513	2.026	2.352	2.654
31	0.000	1.045	1.512	2.024	2.349	2.650
32	0.000	1.045	1.511	2.022	2.345	2.644
33	0.000	1.044	1.510	2.020	2.342	2.641
34	0.000	1.044	1.509	2.018	2.340	2.636
35	0.000	1.044	1.508	2.016	2.337	2.633
36	0.000	1.043	1.508	2.014	2.334	2.629
37	0.000	1.043	1.507	2.013	2.332	2.626
38	0.000	1.042	1.506	2.011	2.329	2.623
39	0.000	1.042	1.505	2.010	2.327	2.620
40	0.000	1.042	1.505	2.008	2.326	2.617
41	0.000	1.042	1.504	2.007	2.324	2.615
42	0.000	1.041	1.503	2.006	2.322	2.612
43	0.000	1.041	1.503	2.005	2.320	2.610
44	0.000	1.041	1.502	2.004	2.318	2.607
45	0.000	1.041	1.502	2.002	2.317	2.605
46	0.000	1.040	1.501	2.002	2.316	2.603
47	0.000	1.040	1.501	2.001	2.314	2.601
48	0.000	1.040	1.500	2.000	2.313	2.599
49	0.000	1.040	1.500	1.999	2.311	2.598
50	0.000	1.040	1.500	1.998	2.310	2.596
⋮						
∞	0.000	0.849	1.282	1.750	2.055	2.326

TABLE 3

Comparison of 2 year flood by (6) and (7)

Site Index	Record Length	Normal Regression Prediction	Student's t Regression Prediction
549770	13	475	475
550200	25	1616	1616
551365	13	343	343
682000	19	866	866
689450	17	2113	2113
690130	13	133	133
690750	14	2190	2190
691020	13	233	233
549510	12	148	148
550300	12	578	578
551360	13	240	240
551420	11	236	236
681600	18	811	811
682030	15	333	333
682100	18	426	426
689618	13	136	136
689650	12	976	976
689670	13	299	299
689720	13	987	987
689960	13	76	76
690250	12	643	643
690280	13	272	272
690470	13	385	385
690570	13	176	176
690720	13	437	437
690830	13	348	348
690850	13	553	553
690940	10	109	109
690970	13	180	180
691025	10	217	217

TABLE 4

Comparison of 5 year flood by (6) and (7)

Site Index	Record Length	Normal Regression Prediction	Student's t Regression Prediction
549770	13	933	929
550200	25	3390	3362
551365	13	812	811
682000	19	2280	2262
689450	17	4325	4288
690130	13	232	232
690750	14	3309	3282
691020	13	531	529
549510	12	357	357
550300	12	1023	1019
551360	13	560	560
551420	11	437	436
681600	18	2161	2146
682030	15	929	924
682100	18	1236	1228
689618	13	359	359
689650	12	2065	2051
689670	13	673	671
689720	13	1987	1977
689960	13	221	221
690250	12	1136	1131
690280	13	562	560
690470	13	721	720
690570	13	433	433
690720	13	706	702
690830	13	590	590
690850	13	998	993
690940	10	253	254
690970	13	385	385
691025	10	437	437

TABLE 5

Comparison of 10 year flood by (6) and (7)

Site Index	Record Length	Normal Regression Prediction	Student's t Regression Prediction
549770	13	1343	1427
550200	25	4810	5024
551365	13	1095	1164
682000	19	3314	3491
689450	17	5767	6027
690130	13	304	330
690750	14	4581	4798
691020	13	831	889
549510	12	480	516
550300	12	1423	1511
551360	13	735	786
551420	11	674	724
681600	18	2943	3105
682030	15	1401	1492
682100	18	1889	2003
689618	13	554	596
689650	12	2824	2978
689670	13	953	1019
689720	13	2571	2715
689960	13	345	373
690250	12	1592	1690
690280	13	845	903
690470	13	930	994
690570	13	622	667
690720	13	1094	1166
690830	13	813	869
690850	13	1491	1582
690940	10	359	388
690970	13	555	596
691025	10	592	635

TABLE 6

Comparison of 25 year flood by (6) and (7)

Site Index	Record Length	Normal Regression Prediction	Student's t Regression Prediction
549770	13	1645	1848
550200	25	8206	8882
551365	13	2377	2685
682000	19	4470	5065
689450	17	8143	8976
690130	13	347	411
690750	14	4863	5263
691020	13	867	978
549510	12	855	982
550300	12	2197	2488
551360	13	1623	1858
551420	11	940	1096
681600	18	5040	5793
682030	15	1962	2287
682100	18	3678	4274
689618	13	1016	1211
689650	12	4277	4854
689670	13	1911	2271
689720	13	3911	4452
689960	13	740	892
690250	12	2138	2424
690280	13	1335	1538
690470	13	1012	1156
690570	13	1160	1343
690720	13	1414	1597
690830	13	1064	1214
690850	13	1902	2125
690940	10	568	662
690970	13	739	853
691025	10	786	905

TABLE 7

Comparison of 50 year flood by (6) and (7)

Site Index	Record Length	Normal Regression Prediction	Student's t Regression Prediction
549770	13	2060	2894
550200	25	10196	11361
551365	13	3054	3547
682000	19	5987	9458
689450	17	10477	12306
690130	13	453	642
690750	14	5865	6334
691020	13	1058	1820
549510	12	1089	1651
550300	12	2828	2821
551360	13	2107	2451
551420	11	1238	1240
681600	18	6972	9193
682030	15	2686	4259
682100	18	5132	6066
689618	13	1416	1732
689650	12	5735	6970
689670	13	2713	2547
689720	13	5257	6391
689960	13	1040	1274
690250	12	2754	2975
690280	13	1741	1885
690470	13	1284	2081
690570	13	1519	1894
690720	13	1763	1600
690830	13	1337	1454
690850	13	2352	2551
690940	10	726	970
690970	13	937	1251
691025	10	995	1328

TABLE 8

Comparison of 100 year flood by (6) and (7)

Site Index	Record Length	Normal Regression Prediction	Student's t Regression Prediction
549770	13	2959	3860
550200	25	12147	14382
551365	13	3666	4765
682000	19	10356	13883
689450	17	13452	16629
690130	13	626	915
690750	14	6611	7817
691020	13	1852	2485
549510	12	1667	2295
550300	12	2868	3700
551360	13	2505	3348
551420	11	1228	1686
681600	18	10111	13762
682030	15	4551	6477
682100	18	6540	9023
689618	13	1770	2584
689650	12	7434	9777
689670	13	2634	3736
689720	13	6798	9000
689960	13	1289	1927
690250	12	3045	3963
690280	13	1902	2562
690470	13	2125	2907
690570	13	1928	2656
690720	13	1582	2016
690830	13	1450	1915
690850	13	2589	3278
690940	10	968	1362
690970	13	1258	1737
691025	10	1338	1839

TABLE 9

A Summary of the Regression Analyses

Year Flood	R^2 for Normal	R^2 for t	Average Percent Error based on Normal	
			Historical	Regression
2	.92	.92	0.0	0.0
5	.889	.889	2.5	- .3
10	.825	.809	11.3	6.5
25	.739	.705	21.9	14.4
50	.683	.627	33.0	24.5
100	.635	.556	47.8	34.4

ACKNOWLEDGEMENT

The U.S. Geological Survey is gratefully acknowledged for the financial support provided to this study. The advice and experience of Professors Duckstein, Kisiel, and Davis of the University of Arizona were greatly appreciated.

APPENDIX A

The t-estimate

It can be shown (Hogg & Craig 1971) that

$$\frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

where

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

and $x_i = \log y_i$ for y_i the ith year maximum annual flow. Now since

$$\hat{s}^2 = \frac{\Sigma(x_i - \bar{x})}{n - 1}$$

then

$$\frac{(n - 1)\hat{s}^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Letting r = n-1 (the degrees of freedom) then

$$\frac{r\hat{s}^2}{\sigma^2} \sim \chi^2_r$$

Since $X \sim N(\mu, \sigma^2)$

and $X \sim N(\mu \frac{\sigma^2}{n})$

then

$$\bar{U} = X - \bar{X} \sim N(0, \sigma^2 + \frac{\sigma^2}{n}) \qquad (A.1)$$

and the standard normalizing $\bar{U}$ to Z

$$Z \sim N(0,1).$$

Thus from (1) $Z = z_i = \frac{x_i - \bar{x} - 0}{\sqrt{\sigma^2 + \frac{\sigma^2}{n}}} = \frac{x_i - \bar{x}}{\sqrt{\sigma^2 + \frac{\sigma^2}{n}}}$ (A.2)

Since the Student's t distribution is defined by

$$\frac{Z}{\sqrt{V/r}} \sim t_r \tag{A.3}$$

where $Z \sim N(0,1)$ and $V \sim \chi^2_r$ then from (A.2) and (A.3)

$$\frac{\dfrac{x_i - \bar{x}}{\sqrt{\sigma^2 + \dfrac{\sigma^2}{n}}}}{\sqrt{\dfrac{(n-1)\,\hat{s}^2}{\sigma^2}\Big/ r}} \sim t_r \tag{A.4}$$

Simplifying (A.4) we get

$$\frac{x_i - \bar{x}}{\hat{s}} \sqrt{\frac{r+1}{r+2}} \sim t_r$$

and the random variables

$$T = \frac{Z}{\sqrt{V/r}} = \frac{x_i - \bar{x}}{\hat{s}} \sqrt{\frac{r+1}{r+2}} \tag{A.5}$$

Thus for $P(T \geq t) = P\left(T \geq \dfrac{x_{50} - \bar{x}}{s} \sqrt{\dfrac{r+1}{r+2}}\right) = .02$,

The exact 50-year flow for small samples n is

$$\frac{x_{50} - \bar{x}}{\hat{s}} \sqrt{\frac{r+1}{r+2}} = t_{.02,r}$$

$$x_{50} = \bar{x} + t_{.02,r} \sqrt{\frac{r+2}{r+1}}\ \hat{s},$$

or

$$x_d = \bar{x} + t_{d,n-1} \sqrt{\frac{n+1}{n}}\ \hat{s} \tag{A.6}$$

REFERENCES

Hogg, R.B., and Craig, A.T., Introduction to Mathematical Statistics, MacMillan, London, 1971.

Skelton, J. and Homyk, A., "A Proposed Streamflow Data Program for Missouri", Open file report, U.S. Geological Survey Water Resources Division, Rolla, Missouri, 1970.

Thomas, D.M., and Benson, M.A., "Generalization of Streamflow Characteristics from Drainage-Basin Characteristics", U.S. Geological Survey Water Supply Paper, 1975, 1970.