Archive

THE INCOMPLETE MEANS ESTIMATION PROCEDURE APPLIED TO FLOOD FREQUENCY ANALYSIS

John C. Houghton

Working Paper No. MIT-EL77-034WP

October 15, 1977

Acknowledgements

This research was done in the Environmental Systems Program in the Division of Applied Science at Harvard University while the author was a Ph.D. candidate. Many of the ideas originated with Professor Harold A. Thomas, Jr. The author wishes to thank the U.S. Geological Survey which sponsored the majority of this research.

Introduction

In a preceding paper by this author (Houghton 1977b), the new Wakeby distribution was shown to have certain advantages over various traditional distributions in modeling U.S. flood records. The Wakeby is defined in an inverse way:

$$x = -a(1-F)^{b} + c(1-F)^{-d} + e \quad , \tag{1}$$

where F is a uniform (0,1) variable. In this paper, we are not searching for a parent distribution. Rather, we are concerned with the problem of estimating the design event once the sample is given. The incomplete means estimation procedure presented in this paper is compared with other estimation procedures using a relative regret model developed in the Appendix. Although we apply a two-parameter version to a variant of the Wakeby distribution, the incomplete means procedure can employ a greater number of parameters, and may also be applied to other distributions than the Wakeby.

Fitting Procedure

The incomplete means method of estimation is applied to the right-hand side of a sample using the following model:

$$x = c(1-F)^{-d} + e \quad . \tag{2}$$

We can neglect the left-hand side of the sample for this application, because in flood frequency analysis only the higher quantiles are predicted. However, this estimation procedure could be easily extended to include the left-hand side for such applications as drought analysis.

Note that Equation (2) is identical to the right-hand side of the Wakeby distribution defined in Equation (1). To date, we have only applied this estimation procedure to small samples ($n = 20$). For these samples, we assume a value of d and estimate the two parameters c and e by linear regression analysis. Analyzing the forty-six flood records (see Houghton 1977b), we found that the d observed in nature is in the neighborhood of 0.2 . Fortunately, the estimate of the T-year flood ($T \leq 100$) is not very sensitive to d, at least for a small neighborhood around d.

The incomplete means estimation procedure uses a combination of means calculated over only part of the range. This new procedure yields fairly stable estimates with little bias because it uses no moments higher than the first. Consider a sample of n ranked observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Calculate the mean $\bar{x}$; it will fall between the two adjacent observations in the sample. It effectively divides the sample into two disjoint sets. Calculate the mean of the upper set (as shown in Figure 1), and call it $\bar{x}_1$. Similarly, calculate the mean of all observations above $\bar{x}_1$ and label that incomplete mean $\bar{x}_2$. The incomplete means and corresponding n_i's can be related to the parameters in the Wakeby distribution in a straightforward way. This is due to the inverse definition of the Wakeby; the relation of incomplete means to parameters of a lognormal distribution, for example, would not be as simple.

To derive the relation between the parameters and the incomplete means, recall that

$$\bar{x} = \int_{-\infty}^{\infty} x f(x) dx \quad . \tag{3}$$

Let $y = F(x)$, $x = F^{-1}(y)$, and $dy = f(x)dx$. Then

$$\bar{x} = \int_0^1 F^{-1}(y)dy \quad . \tag{4}$$

Define $\bar{x}_{(a,b)}$ to be the mean of the interval (x_a, x_b) in x-space, or of the interval (a,b) in F-space. Then

$$\bar{x}_{(a,b)} = \frac{1}{b-a} \int_a^b F^{-1}(y)dy \quad . \tag{5}$$

In the incomplete means method, endpoints are determined by functions of n_i. For example,

$$\bar{x}_2 = \frac{1}{1 - \frac{n_2}{n}} \int_{\frac{n_2}{n}}^1 \left(c(1-F)^{-d} + e \right) dF \quad . \tag{6}$$

After integration,

$$\bar{x}_i = \frac{c}{1-d} \left(1 - \frac{n_i}{n} \right) + e \quad . \tag{7}$$

The first two incomplete means, $\bar{x}_1$ and $\bar{x}_2$, are then used to calculate c and e. If d is given, then the relationship is linear in c and e. This probably implies less bias, and will simplify research on theoretical properties in the future.

Framework For Testing

Using an approach similar to Slack et al. (1975), we use a Monte Carlo experiment to compare various fitting procedures. We assume the "real world" follows a given distribution, and the simulation of syn-

thetic records allows us to evaluate procedures according to a specific criteria. However, this analysis differs from the earlier study in two important ways. First, rather than using a set of equally likely background distributions, with several alternative coefficients of skew, we have used a single distribution and a corresponding set of parameter values. This is the "Central Wakeby", which is defined in Houghton (1977a), and can in some sense be regarded as the most typical or average flood for the set of long records of high quality derived from the U.S.G.S. gaging stations.

Second, an alternative economic criterion is used in this analysis. In general, the design event for flood frequency analysis is $x(T)$, that flow which has an expected return period of T years. A simplified economic model, presented in the Appendix, indicates that it is more logical to estimate the design event x as a function of the density rather than of the distribution function. While one would want a full-scale economic model for any particular project, a simplified model such as this is useful in comparing fitting strategies. Results of the marginal analysis define the design event as

$$x^* = f^{-1}\left(\frac{\beta r}{D}\right), \tag{8}$$

where β, r, and D are economic parameters. The results suggest that for small projects, such as culverts, with large scale economies, the appropriate size is in the range of $f(x^*) = .015$ to $.035$. In addition, the analysis results in a loss function for over- and under-design. Expected regret is assigned to each of the $\hat{x}^*$, and then averaged over the replications, which results in a ranking of the alternative estimation procedures.

The $\hat{x}^*$ is calculated for each synthetic sample by estimating parameters of the distribution, then solving for $\hat{x}^*$. For example, if we are fitting the normal distribution using the method of moments, we would find $\hat{\mu}$ and $\hat{\sigma}$ from the first two moments of the sample. Since

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} , \tag{9}$$

and we define $f(x^*) = \frac{\beta r}{D}$, then

$$\hat{x}^* = \hat{\mu} + \hat{\sigma}\sqrt{2 \ln\left(\frac{D}{\hat{\sigma}\ \beta\ r\sqrt{2\pi}}\right)} , \tag{10}$$

where β, r, and D are given economic parameters. Similarly, the formula for the incomplete means estimator is

$$\hat{x}^* = \hat{e} + \left[\frac{\hat{c}\left(\frac{1}{1+\tilde{d}}\right)}{\tilde{d}\left(\frac{\tilde{d}}{1+\tilde{d}}\right)}\right]\left(\frac{D}{\beta r}\right)^{\left(\frac{\tilde{d}}{1+\tilde{d}}\right)} , \tag{11}$$

where $\tilde{d}$ is an assumed value. For the two-parameter lognormal distribution, the formula is

$$\hat{x}^* = \hat{\mu}_y - \hat{\sigma}_y^2 + \hat{\sigma}_y\sqrt{\hat{\sigma}_y^2 - 2\hat{\mu}_y + 2 \ln\left(\frac{D}{\sqrt{2\pi}\left(\beta r\hat{\sigma}_y\right)}\right)} , \tag{12}$$

where $\hat{\mu}_y$, $\hat{\sigma}_y^2$ are the first two moments in log-space. We also used incomplete means to fit Equation (2) to the logs of the observations.

In this case, the assumed value of $\tilde{d}$ is less than the value of $\tilde{d}$ in arithmetic space. Using $\tilde{c}_y$ and $\tilde{e}_y$ calculated in log-space, the following implicit function is solved for $\hat{x}^*$:

$$\frac{1}{\hat{c}_y \tilde{d}_y} \left(\frac{1}{\hat{x}^*}\right) \left(\frac{ln\ \hat{x}^* - \hat{e}_y}{\hat{c}_y}\right)^{-\left(\frac{1+\tilde{d}}{\tilde{d}}\right)} = \frac{\beta r}{D} \quad . \tag{13}$$

Results

Five hundred replications of $n = 20$ samples were generated from a "Central Wakeby". Each sample was fit six ways corresponding to the six columns in Table 1. The columns represent: (a) the normal distribution using method of moments; (b) the incomplete means assuming a value of $\tilde{d} = .15$; (c) the incomplete means estimator using a value of $\tilde{d} = .25$; (d) the two-parameter lognormal distribution using method of moments applied to the logs; (e) the incomplete means estimator applied to the logs using a value of $\tilde{d} = .04$; and (f) the incomplete means estimator applied to the logs using a value of $\tilde{d} = .07$.

Table 1 is divided into four sets of five rows, each set of rows corresponding to one of four sizes of design events: $f(x^*) = .04$, .021, .008, and .004. These would imply return periods of 30, 50, 100, and 175 years respectively. The second set of numbers, for which $f(x^*) = .021$, corresponds to the data on culvert design presented in the Appendix. The other design events in this Table offer comparison over a wider range. The difference between the estimated event, $\hat{x}^*$, and the true event, x^*, $\Delta x^* = \hat{x}^* - x^*$, is calculated for each combination of a design event and estimation procedure, over all 500 samples. The average of the differences

is labeled "mean" in Table 1, and is a measure of bias. The standard deviations of the differences, labeled "s.d.", measure the dispersion of the estimation procedure. Values of the means square error, "m.s.e.", are proportional to the expected loss assuming a quadratic loss function. "Regret" and "s.d. regret" measure the expected regret and associated standard deviation for the loss function described in the Appendix.

Table 1 shows that the incomplete means estimator, with less bias and lower mean square error, is superior to the normal distribution estimator over most ranges of design events. For all but the smallest design event, $f(x^*) = .04$, the expected regret is half that of the normal estimator. The standard deviations of the mean and regret measures are presented in Table 1 so that any significant differences between values can be calculated. Given the substantially lower expected relative regret associated with the incomplete means method for most values of $f(x^*)$, we did not apply detailed significance testing. However, approximate numbers may be calculated to compare the expected regrets of, for example, the first and second columns in Table 1. Using the central limit theorem for the standard deviation of the sample mean, for $f(x^*) = .021$, we would find that $.18 \pm \frac{.23}{\sqrt{500}}$ (from the first column) is significantly different from $.09 \pm \frac{.15}{\sqrt{500}}$ (from the second column). Note that because the same sets of samples were used for each of these evaluations, the expected regret numbers are positively correlated, which enhances the discriminating power of the test.

However, continuing to the fourth, fifth, and sixth columns, we discover that other estimators do better than the normal. The two-parameter lognormal distribution, fit by taking moments in log-space rather than arithmetic space, was not tested by Slack et al. (1975). However, studies

sponsored by the U.S. Water Resources Council (1976) indicate that, for a wide variety of assumptions and according to certain criteria, the log-Pearson and the two-parameter lognormal were superior to a number of other distributions. Although they recommended the log-Pearson, the two-parameter lognormal performed nearly as well. We attributed the success of both methods to the fact that logs are taken first, and the moments then calculated in log-space. In this manner, sensitivity to high outliers is decreased.

The results of applying the two-parameter lognormal are shown in the fourth column. Although there is substantial bias in the lognormal estimator for large design events, the dispersion is small and the expected regret less than that of the estimators displayed in the first three columns. The incomplete means estimator is then applied to samples in log-space. The fifth and sixth columns show that the incomplete means estimator applied in log-space does much better than even the two-parameter lognormal. It has less bias for some design events, but less dispersion for all design events. And the expected regret is considerably less than that of the two-parameter lognormal.

The transformation to logarithms is undoubtedly an improvement for this choice of parent distribution. But we do not know whether the log transformation is always better than other transformations, such as the square root or even the reciprocal. Application of the incomplete means estimator to parent distributions with thinner tails will probably do better with less radical transformations than the log, such as the square root. It is difficult to determine a priori which transformation to make. Adaptive techniques, that is procedures which rely on parameters which are determined by the observations in the sample, are likely to be

successful. Suggestions about potential transformations are presented in Houghton (1977a).

Conclusions

More extensive Monte Carlo testing is needed to indicate definitively the superiority of the incomplete means estimator. It should be tested using several different background distributions, and be compared with other accepted fitting procedures, such as the Pearson Type III and the three-parameter lognormal. The incomplete means estimator could also be tested using traditional economic criteria such as the T-year event, and new regret models could be developed. But the superiority of the incomplete means estimator demonstrated in this study suggests that future research should consider the use of this procedure.

In addition, the framework for testing alternative fitting procedures, as outlined in this paper, should be useful for future flood frequency research. This analysis differs from earlier formulations of the problem. The Benson (1968) approach emphasizes the ability of a certain procedure to fit the given data points, although goodness-of-fit criteria are inferior to economic criteria from the viewpoint of project design. In addition, in the absence of computer simulation, the results are tied to only a few samples. The Slack et al. (1975) approach answers a very different question: How well do certain fitting procedures estimate the T-year event? Their approach could be improved by incorporating sets of Wakeby distributions as parents on the assumption that they generate patterns of flow which are closer approximations of the real world. Furthermore, loss functions typical of most real situations are asymmetrical and penalize underestimates more than overestimates. A very few

selected loss functions probably better represent the real world than a broad spectrum of all possible shapes.

In future flood frequency research, certain groundrules might be considered in the evaluation of alternative fitting procedures. Initially, a set of distributions, perhaps Wakebys, should be derived from national flood records to represent real flood flows. A relatively small set should be able to "explain" most of the variance found in nature. A method for testing this is presented in Houghton (1977a). Finally, the development of national regret measures would provide a sound basis for comparison of alternative fitting strategies.

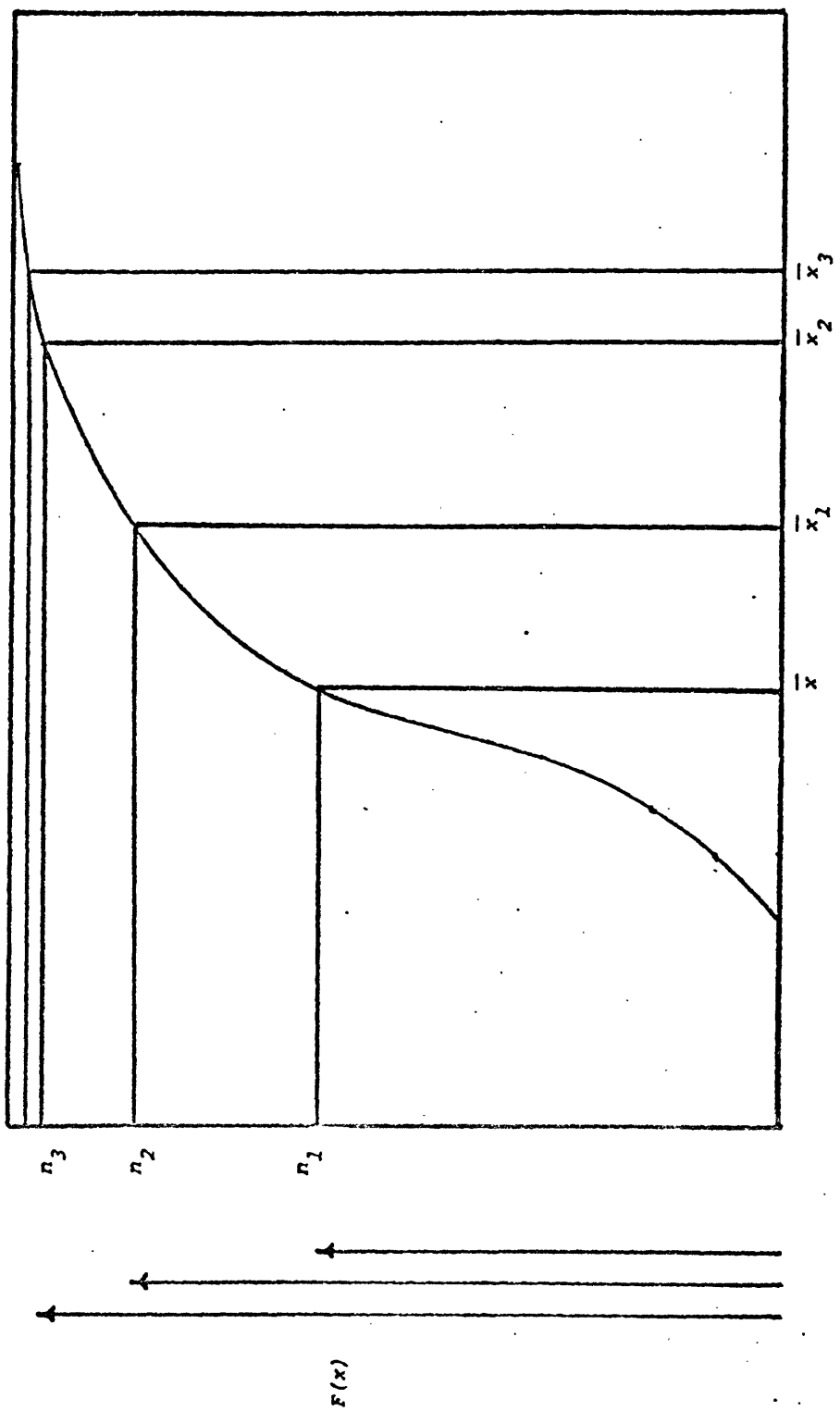

Figure 1. INCOMPLETE MEANS APPLIED TO DISTRIBUTION FUNCTION

TABLE 1. Evaluation of Estimation Procedures Applied To the "Central Wakeby" Distribution †

	Normal	Incomplete Means		Two-parameter Lognormal	Incomplete Means Applied to Logs	
		d=.15	*d=.25*		*d=.04*	*d=.07*
	f(x) = .040		*T = 30*	*x = 2.48*		
mean	-.09	-.08	-.13	.01	-.16	-.17
st.dev.	.60	.42	.40	.39	.24	.23
m.s.e.	.61	.43	.42	.39	.28	.29
regret	.10	.42 *	.42 *	.04	.03	.03
s.d.regret	.14	8.13 *	8.13 *	.07	.06	.06
	f(x) = .021		*T = 50*	*x = 2.94*		
mean	-.13	-.08	-.14	-.06	-.13	-.15
st.dev.	.69	.60	.56	.51	.35	.34
m.s.e.	.78	.61	.58	.51	.38	.38
regret	.18	.09	.08	.06	.04	.04
s.d.regret	.23	.15	.14	.10	.09	.09
	f(x) = .008		*T = 100*	*x = 3.72*		
mean	-.87	-.10	-.13	-.24	-.03	-.04
st.dev.	.79	.99	.91	.70	.62	.60
m.s.e.	1.18	.99	.92	.74	.62	.60
regret	.44	.15	.13	.11	.07	.06
s.d.regret	.51	.25	.22	.20	.15	.14
	f(x) = .004		*T = 175*	*x = 4.42*		
mean	-1.34	-.15	-.12	-.44	.13	.14
st.dev.	.86	1.34	1.24	.87	.91	.88
m.s.e.	1.59	1.35	1.24	.97	.92	.89
regret	.81	.21	.17	.17	.09	.09
s.d.regret	.86	.35	.29	.30	.20	.18

* Calculated with upper limit on relative regret

Repetitions = 500
n = 20

† Parameters
a = 0.431
b = 2.627
c = 1.372
d = 0.236
e = -0.650

Appendix: Regret Model

In this Appendix, we present a simplified economic model which optimizes the size of a culvert. The expected total cost is the sum of the cost of building the culvert itself and of the damage associated with floods which exceed the design capacity:

$$E\left[Z(x)\right] = K(x) + \frac{E\left(D(\tilde{q},x)\right)}{r} , \qquad (14)$$

where x = the size of the culvert in units of flow

$Z(x)$ = the present value of total cost

$K(x)$ = the cost of building the culvert

$\tilde{q}$ = a random variable, the flood in any year

$D(\tilde{q},x)$ = damage due to yearly flood

$E[\ ]$ = expected value operator

r = the interest rate (the model assumes infinite time horizon)

The expected damage is simplified to: $E\left[D(\tilde{q},x)\right] = \left[1-F(x)\right] D$, where $F(x)$ is the probability that any flood is $\leq x$. That is, expected damage is the probability of a flood which is greater than the design event times the average value of the damage for any failure. The expression is dependent on x for the probability of a failure, but not for the extent of the damage. Given that a culvert fails (i.e. is too small for flow), for example, the average loss will be independent of the size of the original culvert. Or in the case of a dam overflow, the expected damage is independent of the size of the dam. This simplification is reasonable for a wide variety of applications.

Cost functions for engineering projects are commonly defined as a

a power function: $K(x) = \alpha_0 x^{\beta_0}$, where β_0 is generally in the range of .6 to .8 . Given that

$$Z(x) = \alpha_0 x^{\beta_0} + \frac{(1-F(x))D}{r} , \tag{15}$$

the minimum regret is

$$\alpha_0\beta_0 x^{\beta_0-1} + \frac{D}{r}\left(-f(x)\right) = 0. \tag{16}$$

Because this does not solve implicitly for x, we substitute a linear approximation around the optimum for the power function

$$\alpha + \beta x = \alpha_0 x^{\beta_0} , \tag{17}$$

which will be fit to real data. Dropping the expected value notation, this approximation changes the total cost function to:

$$Z(x) = \alpha + \beta x + \frac{D}{r}\left(1-F(x)\right) . \tag{18}$$

The minimum cost occurs when

$$Z'(x^*) = 0 = \beta + \frac{D}{r}\left(-f(x^*)\right) \tag{19}$$

or

$$f(x^*) = \frac{\beta r}{D} . \tag{20}$$

This can be interpreted as choosing the flow which corresponds to the value of the density function, whereas the traditional design uses a value of the distribution function.

To obtain realistic numbers for these parameters, we have chosen typical values from a recent report by *Meta Systems, Inc.* [1976], in which they derived cost functions for culverts used in highway design. For box culverts,

$$K'(A) = 23.8\, A^{.660}, \tag{21}$$

where A = cross-sectional area in square feet

$K'(A)$ = the dollar cost per lineal foot.

The average length for an interstate highway is 204 feet, so that

$$K(A) = (23.8)(204)\, A^{.660}, \tag{22}$$

where $K(A)$ is the cost in dollars. The relationship between area A and maximum flow x is given by

$$x = 5.12\, A^{5/4}, \tag{23}$$

so that the cost as a function of flow is

$$K(x) = 2049.8\, x^{.528}. \tag{24}$$

The average cross-sectional area of a box culvert is 50 square feet, which corresponds to a flow of 681 cubic feet per second. Fitting a straight line to the cost function gives:

$$K_0(x) = 29{,}710 + 50.16\, x. \tag{25}$$

Unfortunately, there are no damage functions in the *Meta Systems, Inc.* report. But we can impute a damage given assumptions about the rest of the variables and the distribution of the flows. First, we assume the normalized flows follow a "Righteous Wakeby" distribution. The "Righteous Wakeby" parameters, as defined in Houghton (1977a), are typical of certain U.S. flood flows. Culverts are traditionally designed for a nominal (biased) return period of 50 years; calculating the mean of the "Righteous Wakeby" distribution which has that return period for a flow of 681 cubic feet per second gives: $x = 231$. The imputed damage at the optimum is

$$D = \beta r d c \left(\frac{x^* - e}{c}\right)^{\left(\frac{1+d}{d}\right)} = 33910. \qquad (26)$$

Although a T-year event is used to calculate the parameters, the fitting procedures estimate the $\hat{x}^* = f^{-1}\left(\frac{\beta r}{D}\right)$, which is the optimal flow.

The true (unbiased) return period from traditional flood frequency estimations is much less than 50 years, and other projects are designed for longer return periods. Consequently, the estimation procedures presented in Table 1 are evaluated for several different design events. For each of these design events, imputed damage is calculated in the same way.

Expected relative regret is the additional expected damage associated with the difference between the estimated design event and the true value relative to the damage at the minimum:

$$R(x) = \frac{Z(x) - Z(x^*)}{Z(x^*)} . \qquad (27)$$

Relative regret for the return period T = 50 years is shown in Figure 2.

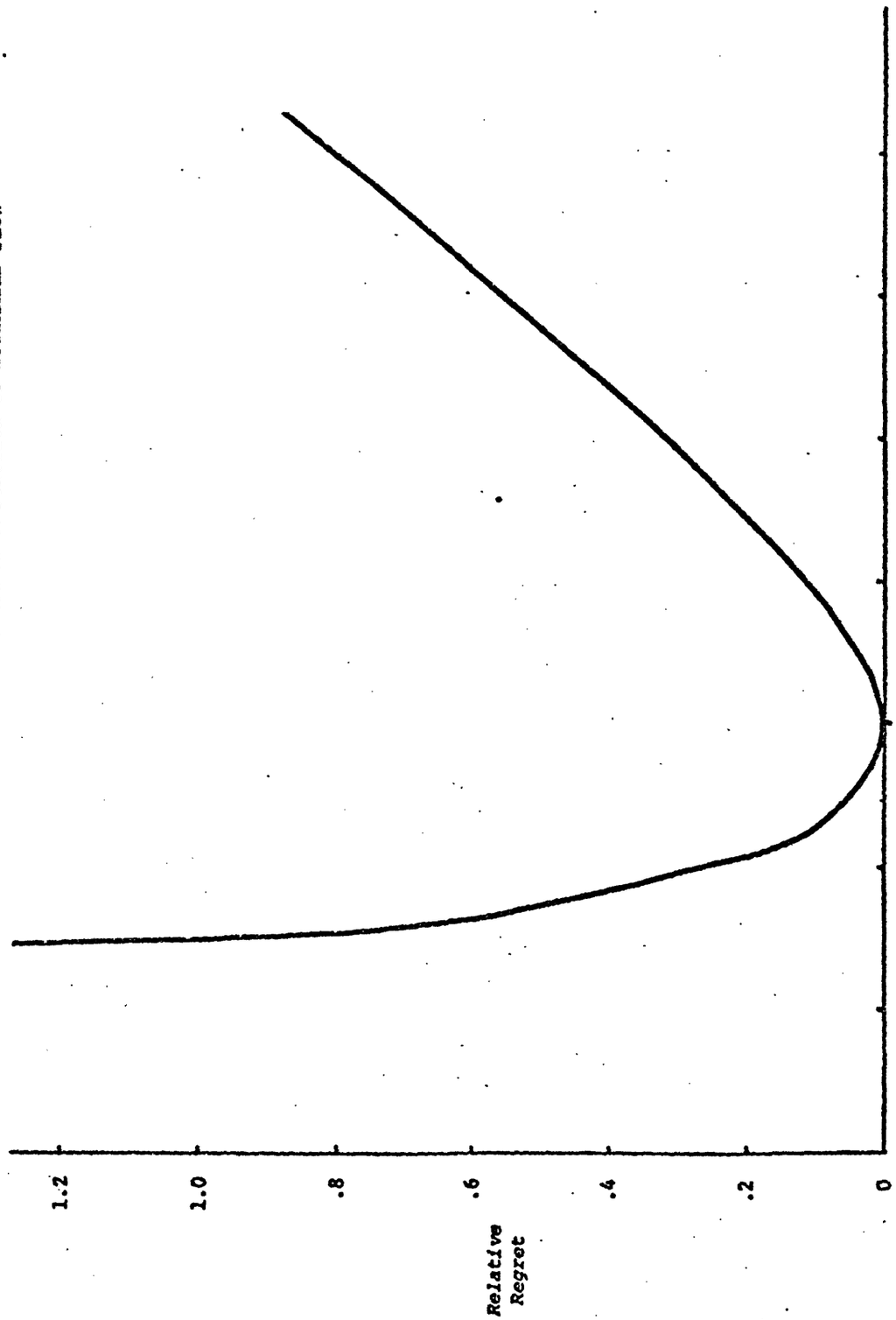

FIGURE 2. RELATIVE REGRET FOR CULVERT DESIGN AS A FUNCTION OF ESTIMATED FLOW

Bibliography

Benson, Manuel A., "Uniform flood-frequency estimating methods for Federal agencies", Water Resources Research, vol. 4, no. 5, pp.891-908, October 1968.

Houghton, John C., "Robust Estimation of the Frequency of Extreme Events in A Flood Frequency Context", Unpublished Ph.D. Dissertation, Division of Engineering and Applied Physics, Harvard University, 1977a.

Houghton, John C., "Birth Of A Parent: The Wakeby Distribution For Modeling Flood Flows", 1977b.

Meta Systems, Inc., "Assessment of National Small Rural Water-Sheds Program", Technical Report prepared under Contract No. DOT-FH-11-8605 for Department of Trnasportation, Federal Highway Administration, March 1976.

Slack, J. R., J. R. Wallis, and N. C. Matalas, "On the value of information to flood frequency analysis", Water Resources Research, vol. 11, no. 5, pp.629-647, 1975.

U. S. Water Resources Council, "Guidelines For Determining Flood Flow Frequency", Bulletin #17 of the Hydrology Committee, March 1976.