USING LINEAR REGRESSION IN HYDROLOGICAL DESIGN

by

G. D. Peterson, D. R. Davis and J. Weber

INTRODUCTION

In hydrological design, the problems encountered often have substantial political, social, and economic effects, thus the importance of making a good decision is great. Relevant data are often used by the decision maker to obtain a more knowledgable decision. However, sufficient data immediately relevant to the problem is not always available, but often other related data is available and is used to improve the design. For example, in flood levee and bridge design the most relevant data is annual peak flow. If only a short record exists, other data sources could be looked at, such as, annual peak flow in a tributary, rainfall, or tree ring width. The second data source can be used to augment the first so that a better design results. Use of the second source of data has been investigated by others (Fiering, 1963; Matalas and Jacobs, 1964; and Gilroy, 1970) for the specific case of multivariate normally distributed data and a linear model relating the two sources. The conclusions are that the correlation coefficient must exceed a critical value if the use of the second source of data is to yield better estimates of the state variables. The critical values were determined by requiring the variances of the estimated mean and variance using the augmented data be reduced. Rather than be concerned with the estimates of the state variables, this study focuses on the decision itself and the possible decisions are ranked using the associated expected losses or gains. This technique, Bayesian decision theory, has been used by Davis, Kisiel and Duckstein (1972) using only the data from the first source. The technique is extended by this study to include the use of data from the second source. The method is valid for values of the correlation coefficient which do not exceed the critical value as well as those which do.

THE MATHEMATICS

There are a number of initial requirements before the problem can be tackled. Two definitions are given for convenience. A data source is primary if the probability density function, pdf, is expressible in the state variables of the problem. A data source is secondary if the functional relation that would permit its pdf to be expressed in terms of the state variables is unknown, but a relation to the primary data exists. Note that these definitions can depend on the problem definition and the knowledge of the decision makers.

Let y denote the primary data and x denote the secondary data. The observed primary data are denoted $y_1, y_2, \ldots, y_N$ and the corresponding observed secondary data are denoted $x_1, x_2, \ldots, x_N$, N an integer. The x_i

---

The authors are Graduate Research Assistant and Assistant Professor, Dept. of Systems and Industrial Engineering and Professor, Management Dept., University of Arizona, Tucson, respectively.

are vectors of observations, such as rainfall for the ith year on the day of peak flow and the day prior to peak flow, two days prior, etc. Thus, $x_i^t = [1, x_{i1}, x_{i2}, \ldots, x_{im}]$, m an integer. Denote the state variables of the problem Θ, where Θ is a vector. The initial information needed is the pdf of the primary data, $\ell(y|\Theta)$, given the state variables; the loss or gain function associated with the decision d and state variables, $g(d|\Theta)$; the prior distribution of the state variables, $P(\Theta)$; the N sets of observed data; and the additional observation of the secondary data to be used in the decision making process, x_{N+1}.

THE LINEAR MODEL

Since the relation between the secondary data and the state variables is unknown, a linear model is assumed to be the relation between the primary and secondary data. The noise term, e, of the linear model represents the portion of the relation between primary and secondary data which is not linear.

Let X be a matrix, $X^t = [x_1, x_2, \ldots, x_N]$, and Y be a vector, $Y^t = [y_1, y_2, \ldots, y_N]$, then the linear model may be written as

$$Y = X\beta + e \tag{1}$$

where β is an m by 1 vector of unknown coefficients and $e^t = [e_1, e_2, \ldots, e_N]$ with the e_i's being random variables which are independent identically distributed with mean zero and variance σ_e^2. Making the assumption that the conditional distribution of y given x is normal, the maximum likelihood estimate of β is found and is B; $B = [X^tX]^{-1} X^tY$. Since $e = Y - X\beta$, the unbiased estimate of $\sigma^2_{e \cdot x}$, s^2, is $(Y - XB)^t (Y - XB)/(N-m-1)$ (Graybill, 1961). Thus,

$$\frac{y - B^t x_{N+1}}{s\sqrt{x_{N+1}^t [X^tX]^{-1} x_{N+1} + 1}}$$

is distributed students-t with N-m-1 degree of freedom (Draper and Smith, 1966). Denote this distribution $t_r(y|x_{N+1})$.

BAYESIAN METHODS

The use of primary data in Bayesian procedures will be given first so that a background will be layed for the discussion of the use of secondary data.

Primary data. In Bayesian procedures where only the primary source is used, a decision d is chosen which minimizes (maximizes) the expected risk (gain),

$$\min \text{ Bayes Risk} = \min_{d} \int_{A} g(d|\Theta)\ P(\Theta|y_{N+1})d\Theta \tag{2}$$

where $P(\Theta|y_{N+1})$ is the pdf of the state variables Θ given the new piece of data y_{N+1} and A is the domain of $P(\Theta|y_{N+1})$. The pdf $P(\Theta|y_{N+1})$ (the posterior distribution of the state variables) is found using Bayes rule,

$$P(\Theta|y_{N+1}) = \frac{P(\Theta)\ \ell(y_{N+1}|\Theta)}{\int_{A} P(\Theta)\ \ell(y_{N+1}|\Theta)d\Theta}\ . \tag{3}$$

This calculation is often simplified by using the natural conjugage distribution of $\ell(y|\Theta)$ (Raiffa and Schlaifer, 1961).

Having made the Bayesian decision, d*, the uncertainty of the decision is measured by the expected opportunity loss, XOL:

$$XOL_y = \int_{A} [g(d^*|\Theta) - g(d_t|\Theta)]\ P(\Theta|y_{N+1})\ d\Theta \tag{4}$$

where d_t is the decision which minimizes the loss function for a fixed vector Θ. A more informative form of the expected opportunity loss is given:

$$XOL_y = \text{min Bayes Risk} - \int_{A} \min_{d} [g(d|\Theta)]P(\Theta|y_{N+1})\ d\Theta.$$

When comparing two decisions based on different sets of data, the decision with the lower XOL has less uncertainty associated with it. This property leads to ascertaining the worth of an additional piece of data.

Let h(y) be the predictive distribution of y, thus:

$$h(y) = \int_{A} \ell(y|\Theta)\ P(\Theta)\ d\Theta. \tag{5}$$

Define the expected, expected opportunity loss as:

$$XXOL_y = \int_{B}\int_{A} [g(d^*|\Theta) - g(d_t|\Theta)]\ P(\Theta|y)\ d\Theta\ h(y)\ dy \tag{6}$$

where B is the domain of h(y), and thus, if the decision based on N pieces of data has associated with it XOL_N, the worth of obtaining an additional piece of information is the expected reduction in XOL and has been labeled the expected value of sample information: $XVSI_y = XOL_N - XXOL_y$. Thus, the XVSI is a measure of the worth of data.

A more complete discussion of the use of Bayesian techniques with primary data is given by Davis, Kisiel and Duckstein (1972).

Secondary data. When using secondary data in Bayesian procedures, the distribution of the secondary data in terms of the state variables is needed. If it were known, the posterior distribution of the state variables given the secondary data could be obtained by Bayes rule:

$$P(\Theta|x_{N+1}) = \frac{P(\Theta)\ f(x_{N+1}|\Theta)}{\int_A P(\Theta)\ f(x_{N+1}|\Theta)d\Theta}. \tag{8}$$

The distribution $f(x_{N+1}|\Theta)$ could be found by the equation

$$f(x_{N+1}|\Theta) = \int_B \ell(y|\Theta)\ t(x_{N+1}|y)\ dy. \tag{9}$$

However, the distribution $t(x_{N+1}|y)$ is not known, but from the development of the linear model the distribution of the primary data given the secondary data was shown to be $t_r(y|x_{N+1})$, and thus by Bayes rule,

$$t(x_{N+1}|y) = \frac{t_r(y|x_{N+1})\ D(x_{N+1})}{\int_C t_r(y|x)\ D(x)\ dx} \tag{10}$$

where D(x) is the multivariate distribution of the secondary data and C is the domain of D(x).

Now Equation (10) may be substituted into Equation (9) and the result may be substituted into Equation (8). Notice that $D(x_{N+1})$ will cancel in the numerator and denominator of Equation (8).

The Bayesian decision for the use of secondary data is the d which minimizes the Bayes Risk

$$\int_A g(d|\Theta)\ P(\Theta|x_{N+1})\ d\Theta, \tag{11}$$

and the expected opportunity loss is:

$$XOL_x = \int [g(d^*|\Theta) - g(d_t|\Theta)]\ P(\Theta|x_{N+1})\ d\Theta \tag{12}$$

and the expected, expected opportunity loss is

$$XXOL_x = \int_C \int_A [g(d^*|\Theta) - g(d_t|\Theta)]\ P(\Theta|x)\ d\Theta\ D(x)\ dx \tag{13}$$

and the expected value of secondary sample information is

$$XVSI_x = XOL_N - XXOL_x. \tag{14}$$

When the point is reached such that the $XXOL_x$ and the $XXOL_y$ are known, a comparison of the worth of primary data to the worth of secondary data can be made.

EXAMPLE PROBLEM

The problem considered is the determination of the depth to which piles are to be driven in bridge construction. The bridge to be constructed is on the Rillito Creek on the north edge of Tucson, Arizona. The bridge spans 500 feet and rests on 100 piles placed in four piers of 25 piles each. The cost incurred if the piers are washed out is \$150,000, and the cost of sinking one pile one foot is four dollars (Laursen, 1969). The useful life of the bridge is assumed to be 25 years. If a flow in the river occurs such that the river bed is scoured to a depth greater than the pile depth, then the bridge is damaged. Thus the loss function $g(d|\Theta)$ is the cost of the damage to the bridge times the probability of the damage occurring once in 25 years plus the cost of driving the piles to a depth d (Davis and Dvoranchik, 1971).

The distribution of the primary data is assumed to be a log-normal distribution; thus, the state variables, Θ, are the mean μ and the variance σ^2 of y, the log of the peak annual flow. The sample mean and variance are joint sufficient statistics for μ and σ^2 (Hogg and Craig, 1970). The distribution of the state variables is an independent normal-gamma, $P(\mu,\sigma^2)$, because the distribution of the mean of a sample of size N is normal with mean μ and variance σ^2/N, while NS^2/σ^2 is independent of the sample mean and is distributed Chi-square with N-1 degrees of freedom (Hogg and Craig, 1970). The normal-gamma is the conjugate distribution of the normal distribution, $\ell(y|\mu,\sigma^2)$, (Raiffa and Schlaifer, 1961).

The secondary data used is the rainfall measured at the N Lazy H Ranch on the day of peak flow and the day prior to peak flow, thus m = 2. The distribution of the rainfall which was used is empirical and was obtained from 31 years of record.

The distribution of the log of the peak annual flow given the rainfall $t_r(y|x)$ was obtained using 10 years of paired data, N = 10.

Using the Bayesian techniques presented, the decision depth of the piles is the depth which yields the minimum Bayes Risk:

$$\text{Bayes Risk} = \int\int g(d|\mu,\sigma^2)\ P(\mu,\sigma^2|x_{N+1})\ d\mu d\sigma^2. \qquad (15)$$

The limits of the outer integration are zero to infinity, and the limits of the inner integration are minus infinity to plus infinity. Equation (8) for the example becomes

$$P(\mu,\sigma^2|x_{N+1}) = \frac{P(\mu,\sigma^2)\ \ell(x|\mu,\sigma^2)}{\int\int P(\mu,\sigma^2)\ \ell(x|\mu,\sigma^2)\ d\mu d\sigma^2}, \tag{16}$$

where the limits of integration are as in Equation (15). Equation (9) becomes

$$f(x_{N+1}|\mu,\sigma^2) = \int \ell(y|\mu,\sigma^2)\ t(x_{N+1}|y)\ dy, \tag{17}$$

where the limits of integration are zero to infinity and, since the distribution of the rainfall is empirical, Equation (10) becomes

$$t(x_{N+1}|\mu,\sigma^2) = \frac{t_r(y|x_{N+1})\ D(x_{N+1})}{\sum_i t_r(y|x_i)\ D(x_i)} \tag{18}$$

The prior distribution of the state variables, $P(\mu,\sigma^2)$, is found through Bayes rule using the N primary data points of the log of peak annual flow; the conjugacy of the normal-gamma distribution to the normal distribution greatly simplifies this procedure. Equation (16) may be used in Equation (15) to find the pile depth which minimizes the Bayes Risk. The measure of regret for perhaps having made an incorrect decision is

$$XOL_x = \int\int [g(d^*|\mu,\sigma^2) - g(d_t|\mu,\sigma^2)]\ P(\mu,\sigma^2|x_{N+1})\ d\mu d\sigma^2, \tag{19}$$

where d^* is the Bayes decision and d_t is the decision made when the state variables are known; the limits of integration are as in Equation (15). Again, since the distribution of the secondary data is empirical, the expected, expected opportunity loss due to secondary data is

$$XXOL_x = \sum_i \int\int [g(d^*|\mu,\sigma^2) - g(d_t|\mu,\sigma^2)]\ P(\mu,\sigma^2|x_i)\ d\mu d\sigma^2\ D(x_i) \tag{20}$$

where the limits of integration are as in Equation (15). The expected, expected opportunity loss due to primary data is

$$XXOL_y = \int\int\int [g(d^*|\mu,\sigma^2) - g(d_t|\mu,\sigma^2)]\ P(\mu,\sigma^2|y) d\mu d\sigma^2\ T(y) dy \tag{21}$$

where the limits of the outter and middle integrations are zero to infinity, and the limits of the inner integration are minus infinite to plus infinity and

$$T(y) = \int\int P(\mu,\sigma^2)\ \ell(y|\mu,\sigma^2)\ d\mu d\sigma^2, \tag{22}$$

where the limits of integration are as in Equation (15) and is students-t with N-1 degrees of freedom.

Given the XOL of the 10-year decision, the $XVSI_x$ and $XVSI_y$ can be found and the comparative worth of additional data, primary and secondary, can be established.

Implementation. The procedure was implemented on a CDC 6400 computer. The integrations were done with Gaussian quadrature schemes and a quadratic search was used for the minimizations. The time used for one calculation of the minimum Bayes Risk and the XOL using secondary data is less than 15 seconds, including six seconds compiling time.

RESULTS

In Tables 1 and 2, the first line is the decision made for the 10 base years of data without using an additional data point. The second line is the decision made using the eleventh year of historical information of the peak annual flow. The third line is the decision made using the methodology presented in this paper for the rainfall on the day of and day prior to peak flow. The fourth line is the decision made when using the prediction from the linear regression model as certain information (ignoring the noise term, e, in the linear model). Since the day of peak annual flow for the additional year is not usually known so that the appropriate historical rainfall can be found, lines five and six correspond to lines three and four for the two consecutive days of rainfall that yield the annual maximum predicted flow.

Table 1

RESULTS OF SECONDARY DATA WITH AND WITHOUT UNCERTAINTY FOR RILLITO CREEK

Additional rainfall observation $x_1;x_2$ (in.)	Flow observation (cfs) Regression prediction	Historical record	Bayes risk ($)	XOL ($)	Pile depth (ft.)
Base primary data only	--	--	6076	1650	13.19
11 years primary data	--	3610	5600	1366	12.20
0.10;0.79[1]	Bayes procedure		5905	1516	12.79
0.10;0.79[1]	4856	--	5540	1344	12.11
1.15;0.0[2]	Bayes procedure		5814	1474	12.74
1.15;0.0[2]	5659	--	5549	1345	12.15

Base data years: 1950-59 Observations from: 1960
Residual variance: .091 Correlation coefficient: .26
Rain Gauge Site: N Lazy H Ranch

1. The historical rainfall for day of and day prior to peak annual flow.
2. The two consecutive days of rain giving maximum regression prediction.

Table 2

RESULTS OF SECONDARY DATA WITH AND WITHOUT UNCERTAINTY FOR RILLITO CREEK

Additional rainfall observation $x_1;x_2$ (in.)	Flow observation (cfs) Regression prediction	Historical record	Bayes risk ($)	XOL ($)	Pile depth (ft.)
Base primary data only	--	--	5850	1600	12.71
11 years primary data	--	8930	5600	1366	12.20
0.46;0.03[1]	Bayes procedure		5609	1405	12.19
0.46;0.03[1]	3936	--	5351	1311	11.74
0.96;0.0[2]	Bayes procedure		5589	1395	12.15
0.96;0.0[2]	4812	--	5340	1308	11.72

Base data years: 1950-57; 59-60 Observations from: 1958
Residual variance: .078 Correlation coefficient: .42
Rain Gauge Site: N Lazy H Ranch

In Tables 1 and 2, the use of secondary data is compared to the use of an additional piece of primary data. When the eleven year historical log flow is not very different from the ten year mean, as in Table 1, the use of the regression prediction as if it were actual log flow data is close to the eleven year primary decision. Although the use of secondary information in Bayesian procedures is an improvement over the use of the ten year primary information, it does not improve the design as much as primary information does. In Table 2, where the eleventh year of log flow is more extremal, the use of the regression prediction as though it were actual log flow results in an optimistic design; its decision depth is less than the historical decision depth, as are its Bayes Risk and XOL. The decision obtained when using the Bayesian procedures is closer to the eleven year primary data decision, although the XOL's of the secondary data situation are greater than those of the eleven year primary data situation. The higher XOL's of the Bayesian procedures in both Tables 1 and 2 are due to the consideration of the uncertainty in the linear model. In the case where the eleventh year of primary data is significantly different from the ten year mean (.95σ), the Bayes Risks and XOL's for both rainfall sets are lower than the Bayes Risk and XOL of the eleven year primary situation. This supports the assertion that using the regression prediction as certain information is assuming more knowledge of the relation between the primary and secondary data than is warranted. Since the eleventh year of primary data is not known in an actual decision problem for which the methodology of this paper is used, these extremal events are the ones of most interest in evaluating the design.

The ten year decision of the two cases presented in Tables 1 and 2 are again given, along with the statistics of primary data for the ten base years, residual variances and correlation coefficients, in Table 3. The values of interest are the relative worths of one additional piece of data,

primary and secondary. It can be seen that the rainfall data improves the bridge design by less than half the improvement due to the log of peak flow data, but the use of the secondary data contributes substantially to an improved bridge design. The improvement due to the secondary data for case 2, which has the lower residual variance, is greater than the improvement due to secondary data in case 1. Also, it is noted that the improved bridge design results even though the correlation values do not meet the critical values as given by previous investigators (Gilroy, 1970).

Table 3

Relative Worth of Primary and Secondary Data

	Case 1	Case 2
Base years	1950-59	1950-57 1959-60
Bayes Risk ($)	6076	5850
XOL ($)	1650	1600
d*(ft)	13.19	12.71
Mean of log of peak flow	3.7470	3.7077
Variance of log of peak flow	.0683	.0662
Residual variance, s^2	.091	.078
Correlation coefficient, R	.2635	.4217
$XVSI_y$ ($)	408	387
$XVSI_x$ ($)	174	184
$\frac{XVSI_x}{XVSI_y}$	.426	.475

The decision resulting from this methodology is not only dependent on the predicted value of flow, but also on the values of the rainfall which yield the prediction, i.e., a predicted value of 4400 cubic feet per second might be obtained by rainfalls of $(1.0\ .25)^t$ and $(.75\ .60)^t$ but these different rainfalls will give different decisions because their associated uncertainties as given by the distribution t_r are different.

DISCUSSION AND CONCLUSIONS

It has been demonstrated that the use of secondary data when insufficient primary data is available yields an improved design. Although the Bayes Risk, XOL and XXOL depend on the correlation coefficient (the dependence is indirect through the residual variance), the use of secondary data is not restricted to certain values of the correlation coefficient, nor is there a decision necessary whether to use the secondary data in marginal cases where the correlation coefficient is close to the critical value.

The methodology presented has the appealing virtue of giving a measure of the expected loss due to the decision made (Bayes Risk) and a measure of the uncertainty of the decision, XOL. The previous investigators (Fiering, 1963; Matalas and Jacobs, 1964; Gilroy, 1970) have presented methods which augment the statistics of the primary data. The design would be done with these fixed statistics with, perhaps, confidence intervals being used to measure the uncertainty. With the procedures of the paper, no fixed statistics are needed; all possible values are considered, the loss function for each value of the state variable and a single decision is weighted by the probability of occurrance. The uncertainty in the state variables is considered and a measure of the uncertainty is given.

The numerical procedures of the methodology presented here are non-trivial. When augmenting the primary data with one additional piece of secondary data, the technique requires triple integration. Each additional piece of secondary data requires that at least one additional integration be added; with each additional integration, the numerical procedures become more difficult and greatly increase the time required to implement the procedure. The procedures devised by Matalas and Jacobs (1964) and Gilroy (1970), on the other hand, augment the statistics of the primary data with more than one additional piece of secondary data with comparative ease.

When using Bayesian procedures, the decision maker is willing to take the risk of damage occurring. If the decision maker is unwilling to take a risk then other methods should be used. For instance, in the example, if the bridge were necessary for a defence route or the only route to a highly populated area, the bridge might be designed to withstand the 500-year flood.

ACKNOWLEDGMENT

This study was partially funded by the National Science Foundation Grant Number GK-35791 on Bayesian Decision Theory and is gratefully acknowledged.

REFERENCES CITED

Davis, D. R. and W. M. Dvoranchik, Evaluation of the worth of additional data, Water Resources Bulletin, 7(4), 1971.

Davis, D. R., C. C. Kisiel and L. Duckstein, Bayesian decision theory applied to design in hydrology, Water Resources Research, 8(1), 1972.

Draper, N. R. and H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1956.

Fiering, M. B., Use of correlation to improve estimates of the mean and variance, U. S. Geol. Survey Prof. Paper 434-C, 1963.

Gilroy, E. J., Reliability of a variance estimate obtained from a sample augmented by multivariate regression, Water Resources Research 6(6), 1970.

Graybill, F. A., An Introduction to Linear Statistical Models, Vol. I, McGraw-Hill, New York, 1961.

Hogg, R. V. and A. T. Craig, Introduction to Mathematical Statistics, 3rd Ed., MacMillan, New York, 1970.

Laursen, E. M., Bridge design considering risk and scout, presented at ASCE Annual Meeting, Louisville, Kentucky, April 1969.

Matalas, N. C. and B. Jacobs, A correlation method for augmenting hydrologic data, U. S. Geol. Survey Prof. Paper 434-E, 1964.

Raiffa, H. and R. Schlaifer, Applied Statistical Decision Theory, Harvard University Press, Cambridge, Mass., 1961.