Estimating Annual Waterway Tonnage from Lock Data: Methodology

DONALD LEAVITT

To provide more timely data, the Corps of Engineers has developed advance reports on estimates of commodity tonnage on American waterways since 1989. They are based on correlations between lockage data and operator-reported data over a decade. Experimentation and development of better methods provide more accurate estimates a year before the final published reports. Better methods include regression equations and common sense. Graphical analysis was used. Scattergrams pinpoint outliers due to poor data collection. Line graphs of tonnages by commodity of river locks show which give the maximum waterway tonnage. Time-series graphs of actual tonnage and promising regression estimates show the best estimator and unreported data. Analysis of the past flow of commodities shows which locks are more logical predictors. Graphs can show patterns in past tonnage relationships to justify estimation. Unreliable data must be identified and dropped from the analysis. The temptation to be mechanistic in the methodology must be avoided, and common sense must be used to overcome spuriousness and multicollinearity errors. These precautions can overcome the pitfalls of automatic computer-generated methods. Methods that improved accuracy included graphical analysis, adjusted regression equations, commonsense methods, and adjusted R-square. Analysis of errors improves methods. Advance data for 1990 showed improvement. Validity measures were developed that prove the model valid. Reliability can be predicted, and factors that increase reliability are known.

To provide more timely statistics, the Waterborne Commerce Statistics Center (WCSC) decided to develop annual estimates of waterway tonnage based on the data of prior years. The goal was to estimate timely WCSC tonnages for the previous year with reasonable accuracy. Annual lock tonnages on those waterways were used. Estimates were made for seven categories of commodities for 17 American rivers and waterways, by direction. Estimates were computed for 1988, 1989, 1990, and 1991. Data users indicate satisfaction with the more timely figures.

Actual WCSC annual tonnages were determined for each past year for each waterway and transferred into a Lotus 4.0 spreadsheet. The commodity groups are coal, petroleum, chemicals, metals, farm products, nonmetals, and miscellaneous. Commodity tonnages were obtained from the Corps of Engineers' Lock Performance Monitoring System (LPMS) for each lock and direction on the waterway and nearby locks on associated rivers. Each variable included annual data from 1982 to 1989 with a usual N of 8.

A correlation matrix was calculated for the variables for each commodity group in the spreadsheet. WCSC figures were correlated with all lock variables. The correlations showed which locks were most likely to accurately predict the WCSC tonnage for that commodity group and direction. In 1991, the 1990 tonnages were also checked by a tonnage-flow analysis to determine whether locks carried a significant proportion of the WCSC tonnage.

The best predictor locks were selected on the basis of higher correlations and higher tonnages flowing through the locks. Initially the percentage change from last year for the predictor lock was used to estimate the percentage change for the waterway.

Another method was to develop regression equations by using the best predictor lock. The equations were based on annual data from 1982 to the current year. The new lock figures (1990) were substituted in the equation to get the 1990 WCSC estimate. We regarded the estimation methodology as an iterative process. Minor modifications of our methods were tried each succeeding year to improve the accuracy of the estimates.

The percentage change method was found empirically to be less accurate than the regression equation method, so it was dropped. Various criteria were used to select the "best equation." Some were found to be associated with greater accuracy and were thus relied on more strongly later in the study.

The problem of faulty data was dealt with as experience was gained. Some years, some locks, and some commodities were inaccurate without obvious cause or regularity. Thus we relied on the more reliable post-1981 period for our data. (Several technical problems and a new system made the inclusion of pre-1982 data infeasible.) Even then, for some waterways only more recent data are accurate. An attempt to use quarterly data for a larger N was soon scrapped, since it not only seemed to increase the rate of faulty data but also added the problem of correction for seasonal variation.

Several regression equations were tried for each estimate, and the one that best met our criteria was selected. Alternate locks were used; several locks were used as variables in the equation; the variable of year was used; sums or differences of locks were tried; when justified, the constant was dropped. The question was asked, Does the predictive regression equation make sense in terms of the real world and all the information we have about the process?

The basic mathematical model is

$$\text{Estimate of WCSC tonnage} = \text{constant} + \text{coefficient} * (\text{tonnage through the best predictor lock}) \quad (1)$$

Waterborne Commerce Statistics Center, U.S. Army Corps of Engineers, P.O. Box 61280, New Orleans, La. 70161-1280.

Estimate (metal products for upbound Mississippi, 1991)

= 500,000 + 1.24 * (upbound metal in Ohio Lock 52)

+ 1.88 * (upbound metal in Mississippi Lock 27, Chain of Locks)

= 500,000 + (1.24 * 5,300,000) + (1.88 * 1,500,000)

= 9,900,000

See Figure 1 for more examples of actual equations. This may vary by adding more variables, including year.

Use of a dynamic rather than a static methodology and other experiments violated "taboos," for which we may be criticized. However, accuracy seems preferable to orthodoxy.

Various criteria used to select the best equation included correlation of the variables with the dependent variable, logic, a measure of fit called average error (which is a modification of the Klein ex post forecast), and proportion of waterway tonnage actually going through the lock.

A measure of fit was developed to test the equations. It uses the absolute difference between the estimate and the annual WCSC data for each year, subtracted from 1. The average of those hypothetical errors was used as the probable best measure of fit for an estimate. (However, no error was allowed to exceed 100 percent.) This measure is called "average error": it is the average error between the actual data and the equation-estimated data for each year. It is similar to using half the data to predict the other half. Equations were selected that minimized this error. Subsequent analysis showed this to be the best predictor of the reliability of the estimate.

The multiple correlation of the equation with the WCSC variable and the adjusted multiple correlation were additional criteria.

THEORY

The theory behind this study is that the tonnage through each lock may be a "sample" of the whole tonnage for the waterway. The problem is to get the most reliable sample: as close to a 100 percent sample as possible. This is actually an estimate, like estimating a vote from a preelection poll. However, in some respects, it is like a forecast.

A monograph on forecasting (*1*) discusses and supports methods that influenced the shift in emphasis explained below. A forecast (or estimation) is "the attempt to make scientific statements about non-sample situations on the basis of relationships determined from sample observations" (*1*,p.76; *2*,p.10). Thus data from the past (sample observations of lock and WCSC tonnage) are used to estimate the future (non-sample observations). Ostrom suggests leaving out certain sample points to determine whether the equation can predict these with sufficient accuracy. This was tried at the start, but with single-digit sample sizes (a maximum of 8 years), soon abandoned. Instead the average error criterion was developed.

According to Ostrom (*1*), the forecast error will be smaller with larger samples, explanatory variables with a larger dispersion, and smaller distances between the nonsample observation and the mean of past, sample observations. Thus, future estimates should be more accurate with more years of data. A large variation in past LPMS data should give better results (e.g., 1 million tons some years and none in others). For a new LPMS figure that is quite different from past figures, the estimate is likely to be less accurate. If the new lock figures are much larger or smaller than the past, we can expect less accuracy. This is consistent with the "contention that we are better able to forecast within our range of experience than outside of it" (*1*,p.78; *3*,p.250). Along these lines, it is also easier to estimate the past than predict the future (by excessive curve-fitting). Economic change and other unpredictables are

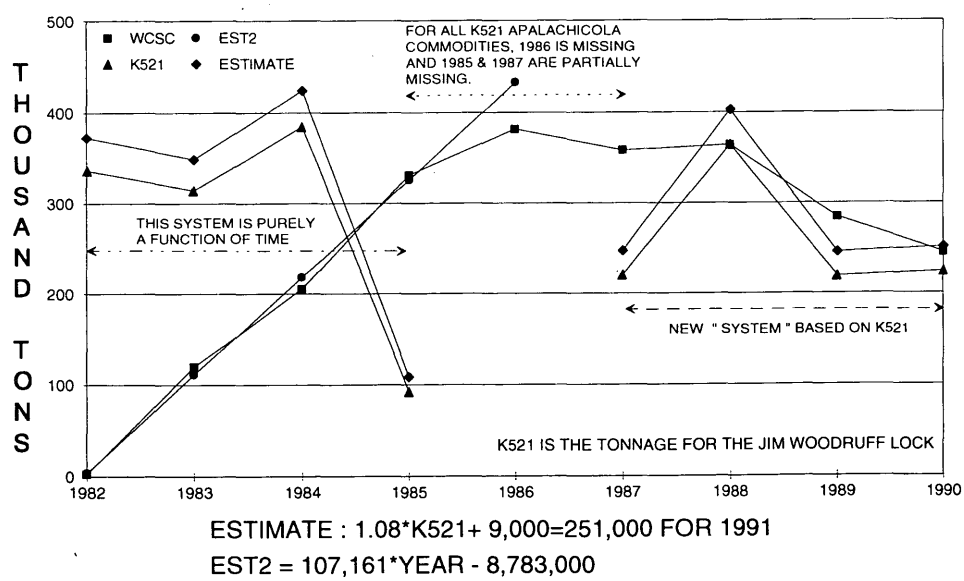

FIGURE 1 Apalachicola River, downbound, nonmetals (1982–1990).

constantly and dynamically changing past relationships. The best solution might be to keep our methods equally dynamic because our past computer models will soon be outmoded.

MIND OVER AUTOMATION

Much emphasis has recently been placed on picking the "best" regression equation by using a particular criterion, formula, strategy, or computer-derived method: stepwise, forward, backward, or chunkwise (*4,5*). However, I have begun to believe that human, nonmechanical techniques are better than automatic, computer-generated ones. The mind can take into consideration available information in a more complex way than do present computer methods.

Ostrom (*1*) says that "there may be times when we wish to be a bit less mechanical in our approach to forecast generation [quoting Klein (*6*,pp.278–279)] . . . (T)here is considerable room for judgment and insight in the generation of forecasts . . . (P)urely numerical methods cannot be used, but must be supplemented by special information and personal judgment . . . (A)ttempts at pure push button mechanistic uses are sure to fail and prove inferior to methods that combine a formal estimated model with a priori information and judgment" (*2*,pp.81–82).

This suggests that when statistical methods result in an estimation that seems suspect, we are allowed to modify it (the fudge factor). Even Einstein added a "universal constant," and it resulted in a model with great predictive value. A similar but computerized method of "fudging" might be to use dummy variables. This, however, brings the disadvantage of adding spuriously to the multiple correlation, just as adding spurious or random variables also appears to give a better prediction by raising the unadjusted multiple correlation. Whether a computer or the human mind should do this "fine tuning" is a question for debate among methodologists. It seems likely that the addition of dummy variables will add misleading spurious variance to a regression equation. The increased multiple correlation may be illusory.

Examples of fallacious results from the overuse of computers in our study include an estimate of negative tonnage, a number that may be quite out of line with the maximum number of tons of a commodity in locks in the waterway, and a negative relationship (lock tonnage goes up but the estimate goes down). Also, when the sum of commodities estimated is out of line with the estimated total, adjustments were made.

We may revise the total equation or the least sound and reliable commodity estimates. Correlation analysis showed that the true figure lies somewhere between the sum of the estimated commodities and the estimate of the total.

Regression equations that multiply a lock tonnage by several times were questioned. Also suspect is a lock that is the sole measure of the WCSC estimate multiplied by a small fractional coefficient. The actual WCSC river tonnage could be more (if some does not go through that lock), but it is unlikely to be significantly less.

Previously, the computer "drew the regression line" and computed the formula. For the 1990 estimates, I sometimes redrew the line that I thought to give the best prediction (judged by a visual scan of the graph). This may involve adding a time factor, combining several locks, or omitting a constant. This was often done when combining several of the larger commodity lock variables to get an accurate estimate of the total. A goal was the closest relationship of the estimation equation with more recent lock data, so the line was drawn to more closely fit the last few data points.

Human judgment, rather than raw computer power or standard programmed formulas or software, seemed most likely to reduce past error levels in our estimates. "Variable selection is a mixture of art and science, and . . . the analyst should be guided by a combination of theory, intuition, and common sense" (*7*).

In sum, we should use all the information available in our estimate, even if it cannot be stated in a precise mathematical form.

In this project, information about lock data is available that can improve the technique beyond a blind adherence to mathematical criteria. Most scientists also have a theory to guide their understanding of a process. A 20-variable correlation matrix without a hypothesis as to which variables should correlate and which should not is a fishing expedition (made easy by computers), not a scientific study. Yet, Bowerman and O'Connell (*4*, chapter 8) recommend putting in all possible variables (like a gumbo) and using statistics to pick out the "best" regression equation. However, I have found that one must discriminate to avoid spurious correlations. A plot or visual scan will reveal outliers and improve judgment for further analysis.

Spuriousness

If we run a 14 by 14 matrix of variables, accepting a .05 error level, we could get five spurious correlations purely by chance. It is easy to get a high correlation where there is no real relationship. Often by pure chance we may discover (for example) that the LPMS tonnage of oil from a lock correlates with the tonnage of metal from WCSC. Sometimes upbound lock figures correlate with downbound WCSC figures. A small-tonnage lock may correlate highly with much larger WCSC totals. Many of these were pure accidents, not reliable relationships. If we use them to predict next year's tonnage, we may be lucky, but we are most likely to be inaccurate.

Thus, when logic tells us that no connection is reasonable, even high correlations should be used only with caution when there are other, more logical candidates.

Here are some techniques based on commonsense information: Lock tonnages were only used to estimate same-direction WCSC tonnages. More tonnage in a lock is logically likely to be related to more tonnage in the WCSC figure (so we omit variables with negative correlations). The total of all commodities should be equal to the sum of each commodity group estimated. The lock on the waterway with the most tonnage for a direction is likely to contain the best "sample" of tonnage for the waterway. We prefer bigger samples to measures with larger correlations.

This method can also be used to determine which locks give tonnage consistent with the WCSC figures. If all the river tonnage goes through a lock, the lock tonnage should equal the WCSC tonnage. From a graph of several locks, the one closest to the WCSC tonnage or the nearest pattern in recent years can be selected. In this case, we may omit the constant in the regression equation.

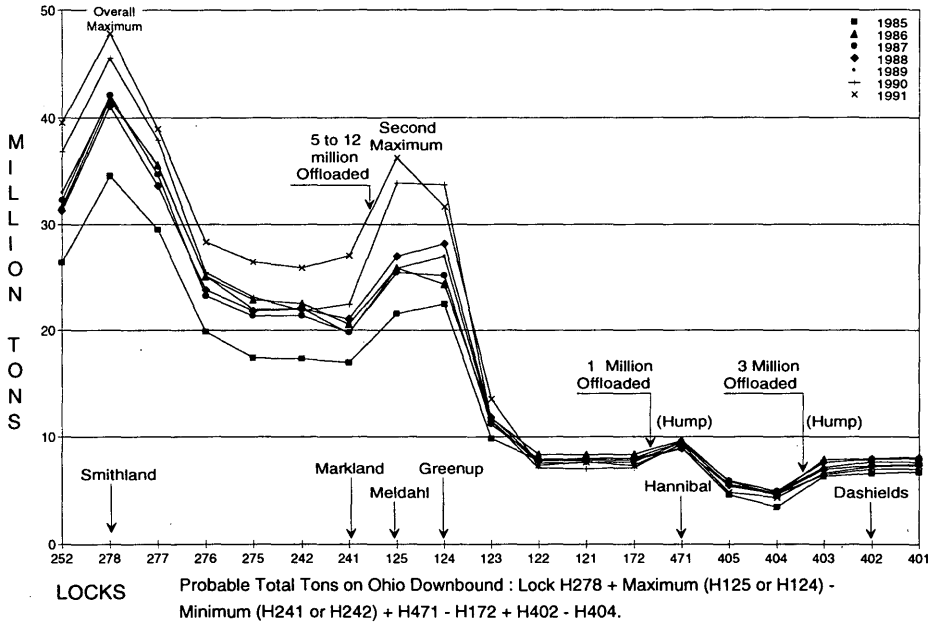

FIGURE 2 Ohio River, downbound, coal (1985–1991).

Multicollinearity

We should expect high correlations of locks with WCSC data, since they are from the same river. Often we find them.

When we find many correlations, the problem of multicollinearity presents itself. This involves multiple regression analysis with explanatory variables that correlate highly with each other, leading to unstable regression coefficients and erroneous inferences about the model (*7*,*8*). For example, another variable may be added that makes a variable coefficient flip over to negative (contrary to the logic of the analysis). Another highly significant variable may then be reduced to insignificance. At times, regression analysis seems to be a matter of chance rather than science. We were unsuccessful in using factor analysis, first differences, or two-stage least-squares regression to solve the problem. Eventually common sense and graphical techniques were applied for more satisfying equations.

CHARTING THE FLOW OF COMMODITIES

High correlations and regression equations can suggest candidates for estimates, but we must guard against spurious correlations. We can use the WCSC data file for the latest year to document the flow of a commodity within a waterway or system of waterways. This is an empirical method to find how much tonnage actually went through the various locks under consideration in a recent year.

Initially we found high correlations with WCSC tonnages for river locks or for the lock of a nearby waterway (e.g., an Ohio lock for the Tennessee River). Later flow analysis gave empirical evidence that relationships were justified (in addition to probabilistic deduction). Sometimes we found that it was not. Sometimes most of the river's tonnage does not go through a reporting lock. The Alabama River is an example.

The flow of commodities can show which locks are the best estimators that have the most tonnage flowing through the river studied. Figure 2 is an example of a graph of the flow chart of the river with the locks in geographical order. This method shows which locks have the most tonnage. A river with several maximums (humps) indicates that there are short hauls or local tonnage, and the total may be estimated by summing the locks giving the maximum for each hump (and perhaps subtracting the locks with the minimum between them). An example was nonmetals in the downbound Allegheny. This hypothesis was tested by calculating the error level and correlation for each likely combination. If a higher correlation is the result of such a combination, greater confidence in our final estimate is justified. However, adding variables also means adding to the error factor.

CONSISTENT AND STABLE PATTERN

Estimation is based on the assumption that the system dynamics will remain stable so that the future will be like the past (*1*,p.82; *4*,p.3). Reliable estimates will then be possible.

However, the process for the year studied may not be typical of previous years. Changes in extent of usage or operation (such as the Tennessee-Tombigbee), a drought, or shifting patterns of economics may change the underlying process of flow of commodities. Graphing the processes can reveal when the recent system is different from the past. Flow analysis revealed changes in patterns. Only the years reflecting the present system ideally should be used for the regression equation. Sometimes only 1 or 2 years define the current system. Unfortunately, this bases the estimates on few data with a

low level of reliability. However, the ideal method should involve understanding the system and trying to use all useful available information.

Thus, forecasting works when the same structure operating in the sample period is still in force in the postsample period. A correction for trends over time ("year") may aid the process. The lock or WCSC tonnage may be increasing in the accuracy of reporting. The tonnage of a commodity may be steadily increasing or decreasing over time for shipments that do not go through the best lock measure. "Year" alone was used in the equation if the best lock gave an unreliable estimate according to "average error" or no lock correlated with WCSC tonnage.

GRAPHICAL ANALYSIS

The stability of the pattern may not be reflected by a simple correlation. It can, however, be tested by a graphical analysis of the relationship between the data points over time. A change in system dynamics can be identified and only the more recent system data used in the regression analysis (Figure 1).

Present software makes it much easier to draw graphs quickly. Recent books and statistics courses include techniques of data analysis through graphing (*7,9*). "Look closely at the raw data . . . before carrying out a multiple regression analysis . . . since computer programs are blind to many anomalies in the data" (*7*,p.6). The estimate of 1990 data involved line graphs of tonnages with time and also geography and some scattergrams. Lotus 4 drew the graphs. Line graphs of lock tonnages over time were constructed along with the variable of WCSC tons to determine which years did not show a consistent relationship between locks and WCSC tons. Outliers can be spotted and dropped from inclusion in a regression equation (inferring that they are bad data).

THE PROBLEM OF BAD DATA: OUTLIERS

In 1929 the British economist Josiah Stamp said, "Government(s) are very keen on amassing statistics—they collect them, add them, raise them to the nth power, take the cube root, and prepare wonderful diagrams. But what you must never forget is that everyone of those figures comes . . . from the village watchman, who just puts down what he damn pleases" (*10*).

The most logical predictor to estimate the WCSC river tonnage for each direction and commodity should be the largest commodity tonnage of any lock. If the correlation is low or negative, a better measure may be found, however. Measurement error was a problem. Some districts or locks may not record all the tonnage for the year (*11*). The processing of the data may be incomplete. Some data may be double counted. A change might develop in accuracy (increasing accuracy or enforcement of reporting).

Several districts did not report for several years (1983, 1986). Lock data are incomplete for some years or months. Other observers may classify commodities differently than the operators reporting to WCSC. Some record more nonmetals than does WCSC but considerably less miscellaneous tonnage, for example. Not all operators report their cargoes to WCSC in some years.

WCSC is familiar with the problems of incomplete reporting and poor accuracy, since many operators must be painstakingly coaxed to send accurate and prompt reports of their cargo movements. Experience with LPMS data revealed the pitfalls of high-tech analysis of poor data, which may result in faulty and inaccurate estimates.

Bad data probably accounted for the less accurate estimates. Scatter diagrams and line graphs show the correlation between a likely annual lock total and the WCSC figure. They might indicate that a value is out of line (an outlier). We may then proceed on the expectation that such data points are unreliable. The LPMS staff can double-check the data, or we may omit such points from the analysis.

BEST FIT REGRESSION

How can we be sure that the right variables and the right number of variables in a multiple regression equation are used to maximize the estimation accuracy?

The addition of any variable (even an unrelated one) will decrease the unexplained variation and increase the multiple correlation (*4*,pp.435–436). A dozen unjustified variables can be used in an equation, which would give a perfect correlation but a perfectly awful estimate.

Initially, I relied too much on computer or formula-derived regression equations. Much information was ignored. Regression equation formulas for 1989 were often automatically based on the highest correlation among lock variables and the highest multiple correlation.

Bowerman and O'Connell (*4*) give some better measures for the best fit of an equation: a corrected R square, a mean square error, and a C-statistic. The "best" equations have the largest corrected R square (multiple correlation squared), the smallest mean square error, and the smallest C-statistic (*4*,pp.436–441). The corrected R square was adopted (also called corrected multiple coefficient of determination):

$$\text{Adjusted } R \text{ square} = \left[R \text{ square} - \frac{(np - 1)}{(n - 1)}\right] * \frac{(n - 1)}{(n - np)} \qquad (2)$$

where n is the sample number and np is the number of variables used in the regression equation, including the constant.

$$C\text{-statistic} = \frac{\text{SSE}}{sp(\text{squared})} - (n - 2k) \qquad (3)$$

where SSE is the unexplained variation (sum of the squares of the error), k is the number of variables in the equation (not including the constant), and sp(squared) is the mean square error calculated from the model with k variables and a constant.

The solution is to select the most logical variables that maximize the adjusted R. Of course each variable in the equation should also have a significant t-score when its coefficient is divided by its error factor, at a .10 probability level.

TABLE 1 Comparison of 1988 and 1989 Measures

	1988	1989
Mean Correlation	.89	.86
Median Absolute Percent Error	16.10	15.37
Average Error 1982-87	13.97	16.83
Average Tons Error	952,819	651,333
WCSC Tons	11,097,162	9,596,500
Number of Estimates	160	278
Correlation of Estimate with Actual tons	.9976	.9982

For the naive model for 1989, Median Absolute % Error = 15.75
Tons Error = 759,102

TABLE 2 Percentage Error Levels for Waterways, 1988 to 1990

	Total Tons By Direction Mean Percent Error			Change
River	1988	1989	1990	
Ohio	1.4	1.4	2.4	Worse in '90
Mississippi	6.8	7.3[a]	6.3	Improvement
Tennessee	7.4	8.1	2.5	Improvement '90
Monongahela	9.1	8.3	2.6	Improvement
Arkansas	9.2	9.3[a]	14.3[a]	Improvement
Black Warrior	12.9	15.8	5.6	Improvement '90
Kanawha	9.0	8.3	3.4	Improvement
Cumberland	10.4	11.6[a]	2.9	Improvement '90
Illinois	8.9	6.6	7.6	Improvement '89
Allegheny	22.6	31.7[a]	15.0	Improvement '90
Columbia	16.3	13.4	11.4	Improvement
Tennessee-Tombigbee	24.5	34.8	11.9	Improvement '90
Apalachicola	37.4	60.0	5.6	Improvement '90
Snake	34.7	24.0	25.0[a]	Improvement '89
Alabama-Coosa	54.5	41.3	5.2	Improvement

[a]On the average of all commodity groups and directions, there was substantial improvement from the previous year.

EVALUATING FORECASTS

When the estimated year's actual WCSC data are available, a check of accuracy is done. Improvement over the previous year's predictions is assessed.

The errors of our estimate can be compared with those of a "naive model" (*1*,pp.84–85; *12*,p.572). The estimate can be compared with a no-change model. In this naive model, next year's value is set to this year's value. If the regression equation "does no better than this naive model, the implication is that it does not abstract any of the essential forces making for change, that it is of zero value as a theory explaining year to year change" (*13*,p.109).

The formula ratio for determining the efficiency of the model is

$$\frac{\text{RMSE (model)}}{\text{RMSE (naive)}} \tag{4}$$

RMSE is the root mean square error of the forecast (deviation from perfection). For only one estimation, it would be

$$\frac{\text{Absolute (estimate} - \text{actual)}}{\text{Absolute (last year} - \text{actual)}} \tag{5}$$

A similar method would be to determine whether the direction of change is the same as predicted.

The estimating process is assessed annually. The least accurate estimates and least accurate waterways are screened for the causes of error. Two measures of error were percentage of error and error-tons. Measures of the different variables (correlation, average error, tonnage of commodity in lock) that might affect error are correlated with the measures of error (see Table 1). Estimates are ordered by groups and the average error of each group calculated by waterway (Table 2), commodity group, and number of variables. I went back to the equations to see what pattern could be found. A good assessment question is, Was our estimate better than a random guess? Usually it was. When it was not, why not?

ESTIMATING PROCESS, 1989 VERSUS 1988

An assessment of the 1989 waterway estimates was done to improve the process, assess its value, and avoid pitfalls for the 1990 process.

Validity of Model

Estimates correlated with WCSC data at a slightly improved 0.99816 for 1989 compared with 0.99758 for 1988. Figure 3 shows this high precision for our estimates compared with the actual WCSC tonnage. Of the 278 estimates, the larger tonnage points are close to the ideal of accurate estimates (equality for the two axes). Only the smaller tonnage points (on the left) show some random scatter and outliers. Smaller tonnage sizes account for this, but accuracy on the smaller waterways and commodities is less important.

Average ton error per estimate was 651,333 for 1989 (15 percent) compared with 952,819 for 1988. The 1988 error was reduced a third for 1989 (see Table 1).

The formula for the efficiency of the model is mean error for the model divided by mean error for a "naive" model that compares the previous year (1988) with the year estimated (*1,13*). When using average absolute tonnage in error (WCSC minus estimated), an efficiency of 0.864 was calculated (the smaller the better). By using average percentage error, the figure is 0.959. A figure above 1.0 would give evidence of an invalid model. Our evidence shows that our model does work better than chance. The 1989 process was better than that of 1988 on the basis of these measures.

Our conclusion is that if we estimated 1989 figures by using the 1988 figures (assuming no changes), we would be less accurate than by using our model. Thus, the formula above might have but does not prove that our model is invalid. (By inference, it is valid.) Another method developed was to determine whether our model predicts the right direction of change between 1988 and 1989. For 62 percent of the time the correct direction of change was predicted: when a commodity of a river was predicted to increase, it did. Predictions for the Apalachicola, Columbia, and Kanawha were very poor: the correct direction of change was predicted only 40 percent of the time (chance alone should give us 50 percent). When these were omitted, the rivers were predicted correctly 68 percent of the time.

Lock data for the Apalachicola, Alabama-Coosa, and Columbia were poor and erratic. Much of the commerce does not go through the only reported lock, and tonnages are small for the former two. Using our validity checks, estimates for these waterways were not valid in 1989.

Comparisons with the 1988 Process

These correlations show that for both years, "average error" correlates highly with absolute percentage error for the year studied. The regression formula is absolute percentage error = 1.7 * average error. We can expect the confidence level for our prediction to be within plus or minus 1.7 times the average error. The regression equation for error-tons is 0.0444 * tons estimated. Thus we can expect an average error of 4.4 percent of the tons for that estimate for 1989.

"Year" as a variable increased accuracy somewhat, especially in reducing error-tons. It was used sparingly. Use of more variables in the equation helps lower the percentage of error, especially when two variables are used (a constant was always used for 1989), but one variable was usually better than two. The adjusted R square was not used for 1989, but its use for 1990 indicated when several variables were justified and when they were not, making multivariable equations more efficacious.

The comparison of estimates of totals for waterways is given in Table 2. Estimates with more tonnage have a smaller percentage of error: above 1 million, percentage error is 10 percent; below, it is 44 percent. Higher correlations between variables strongly reduce error, and a lower "average error" for an equation shows the strongest effect on reducing percentage error (with a 0.55 correlation). For all waterways except the Apalachicola and the Tennessee to Black Warrior

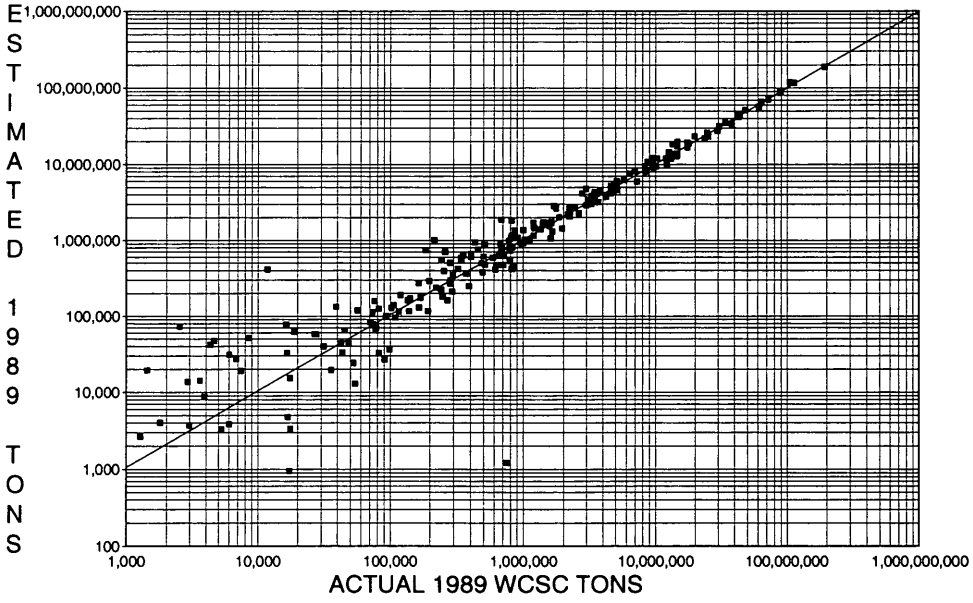

FIGURE 3 Analysis of 1989 estimation process (logarithmic scale).

system, 1989 was an improvement over 1988. Improvement in all waterways in 1990 is noted except for the Ohio and Illinois.

Large-tonnage commodities are more accurate than smaller ones. Fewer locks result in less accurate estimations: the Mississippi is less reliable than the Ohio. For 1990 the one-lock waterways were less reliable: Alabama (mean error of 98 percent), Apalachicola (69 percent), Black Warrior (27 percent). Two- and three- (reporting) lock waterways were somewhat better: Tennessee-Tombigbee (44 percent), Kanawha (30 percent), and Cumberland (30 percent). Past faulty collection of LPMS data also seems a factor in lowering reliability: Alabama (98 percent), Columbia (41 percent), Apalachicola (69 percent). Rivers without these problems were better: Ohio (7 percent), Mississippi (11 percent), Gulf Intracoastal (13 percent), Tennessee (15 percent), Arkansas (21 percent), Monongahela (16 percent), Allegheny (36 percent), and Illinois (35 percent for 1988).

The model for 1989 shows improvement, and data indicate that it is a valid model. Reliability can now be predicted, and several factors increasing reliability are known.

REITERATIVE PROCESS: LEARNING FROM EXPERIENCE

Available recent statistical sources were consulted. Government library facilities were not adequate (except for the loan system), but university libraries were helpful. Opportunities to consult with experienced statisticians working on similar problems might help this process, as might a sense of fallibility and a willingness to drop less successful methods.

Our past regression equations were consulted, but each equation was updated with the newly available WCSC figures every year.

Better methods include using regression equations, larger-tonnage locks, "logical measures" of waterway tonnage even if correlations are lower, and fewer variables except on the larger tonnages.

Graphical analysis was used: scattergrams may pinpoint a year in which the results are out of line. Line graphs of tonnages by commodity between all the locks of a river [e.g., Ohio (Figure 2)] show which lock or sum of locks gives the maximum tonnage for the waterway. Graphs by year of various measures, including actual WCSC data and the more promising regression estimates, may show which is the best estimator, or when the lock or WCSC has not collected all the tonnage data, and perhaps which years might be omitted from the equation (Figure 1).

The adjusted R square, along with common sense, can help select variables. Graphs can show whether there is any pattern to past tonnage relationships to tell whether estimation is justified and, if so, what years to use.

These methods (graphical analysis, adjusted regression equations, common sense, and corrected R square) seem sounder than previous ones. If not, other ways will be sought to perfect the methodology. Advance 1990 WCSC data showed improvement for all waterways except the Ohio and the Illinois. The latter may contain problems, which will be studied. Otherwise the improvement would have been more dramatic. The Mississippi improved slightly in spite of the absence of lock data for the lower Mississippi. Total tonnage error for all estimates decreased by 11 million tons or 7 percent for 1990 over 1989.

ACKNOWLEDGMENTS

I wish to thank Dave Penick, Tom Mire, and Sid Andrus for developing this project, inputting ideas, providing the resources for making it possible, and for critiquing the paper; my colleague Yun Chan for assistance and graphical work for this project and the paper; Martha Broussard for useful comments; and the automated data processing personnel for their technical assistance.

REFERENCES

1. C. W. Ostrom. *Time Series Analysis: Regression Techniques.* Sage Publications, Beverly Hills, Calif., 1990.
2. L. R. Klein. *An Essay on the Theory of Economic Prediction.* Markham, Chicago, Ill., 1971.
3. J. Kmenta. *Elements of Econometrics* (second edition). MacMillan, New York, 1986.
4. B. L. Bowerman and R. T. O'Connell. *Time Series Forecasting: Unified Concepts and Computer Implementation.* Duxbury, Boston, Mass., 1987.
5. D. Kleinbaum, L. Kupper, and K. Miller. *Applied Regression Analysis and Other Multivariate Methods.* PWS-Kent, Boston, Mass., 1988.
6. L. R. Klein. *A Textbook of Econometrics.* Prentice-Hall, Englewood Cliffs, N.J., 1974.
7. B. C. Carson and S. Fullerton. Graphical Techniques in Data Analysis. Presented at Harris Users Conference, Miami, Fla., March 1989.
8. G. C. Judge, W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. Lee. *Theory and Practice of Econometrics.* Wiley, New York, 1985.
9. W. S. Cleveland. *The Elements of Graphing Data.* Wadsworth, Monterey, Calif., 1985.
10. P. Kopac. Guide for Conducting Questionnaire Surveys. Presented at 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
11. L. G. Bray. *A Methodology for Forecasting Short Run Lock and River Tonnage Activity.* Tennessee Valley Authority, Knoxville, 1989.
12. C. Christ. *Econometric Models and Methods.* Wiley, New York, 1966.
13. M. Friedman. Comments. *Conference on Business Cycles.* National Bureau of Economic Research, New York, 1951.

Publication of this paper sponsored by Committee on Inland Water Transportation.