Compensating for Missing Values and Invalid Responses
In Contingent Valuation Surveys

Richard T. Carson, University of California at Berkeley

## Introduction

The benefits of programs aimed at improving environmental quality are measured with increasing frequency by contingent valuation (CV) surveys. These CV surveys incorporate elements of public opinion polls and marketing surveys, as well as a theoretical framework based on modern welfare economics.[1] In the simplest sense, CV surveys present a fairly complex scenario in which a hypothetical market is set up and respondents are asked the maximum amount of money that they would be willing to pay in order to obtain specified increases in the level of an environmental amenity. The need for a survey instrument that asks respondents these questions, in a manner consistent with relevant economic theory, and, at the same time, is understandable to respondents, gives rise to a large number of potential non-sampling errors (Carson and Mitchell, 1984a). The most easily identified of these non-sampling errors (and we believe potentially the most serious) is a large item non-response rate for the WTP questions.

In this paper, we consider several different methods of imputing values for the missing or unusable responses to the WTP questions in a recent survey on WTP for national fresh water quality improvements (Carson and Mitchell, 1984b). The methods include imputation of the mean and median from imputation classes defined in different ways: sequential and random hot decks, maximum likelihood estimation, and several variants of CART, a tree structured regression and classification procedure (Breiman *et al.*, 1984). Our problem of imputing values for the missing and invalid WTP responses is difficult since we are interested in three statistics, the mean, the standard error of the mean, and the median.

This paper is divided into six sections. The first describes how the statistics of interest are used in estimating the benefits of national water quality improvements. The second section describes the characteristics of the data set, which are relevant to the missing value problem. The third section briefly presents each of the imputation methods used. In the fourth, we discuss those features of CART which may be useful in imputing missing values, and which are likely to be unfamiliar to the reader. The fifth section presents the results of the different imputation exercises. The final section includes a discussion of these results and some concluding remarks.

## 1. Measuring Water Quality Benefits

A household's maximum willingness to pay for an improvement in water quality from an initial specified level, $q_0$, to a specified higher level, $q_1$, is an economic quantity known as compensating surplus. Since water quality is an amenity that may be enjoyed by everyone, it is the sum of the public's willingness to pay that determines the demand side. As it is difficult, if not impossible, to poll every household in the United States, the standard practice of taking a sample survey is used.

For a specified change in water quality the desired economic measure is:

$$J = \sum_{i=1}^{H} WTP_i \, , \tag{1}$$

R the number of respondents and M the number of households that did not respond, then N = R + M. The expectation of WTP is

$$E(WTP) = (R/N^2)\sum_{i=1}^{R} WTP_R + (M/N^2)\sum_{i=1}^{M} RWTP_M \tag{3}$$

The bias of treating $(R/N^2)(WTP_R)$ as E(WTP) is,

$$(M/N^2)(WTP_R - WTP_M); \tag{4}$$

where this bias is small if M is small or if $WTP_M$ is close to $WTP_R$.

## 2. A Description of the Data Set

The data set used in this paper is from a survey of 813 households in the contiguous United States. The survey queried respondents primarily about their use of water based recreation, and their attitudes toward water quality including their willingness to pay for water quality improvements. The survey was conducted by Opinion Research Corporation (ORC) of Princeton, New Jersey, for Resources for the Future in November and December of 1983. The sample was drawn using a multistage areal probability design.

To compensate for differential response rates (for the questionnaire as a whole), ORC supplied a set of weights based on the 1980 Census in order to produce sample statistics (assuming no item non-response) representative of the non-institutionalized U.S. population. However, as with most sample surveys, there is both a non-response and an item non-response problem. Results from different methods of imputing values for the item non-responses on willingness to pay for water quality are given with, and without, the ORC sample weights. One of the advantages of of imputing values for the item non-response is that the where H is the total number of households, and the subscripts denoting the exact quality change $q_0$ to $q_1$ have been dropped to avoid notational clutter. Using a sample survey, J can be estimated by multiplying the sample mean WTP from the survey by H:

$$\hat{J} = H \sum_{i=1}^{n} \frac{1}{n} WTP_i \, , \tag{2}$$

where n is the number of households surveyed and every household in the population of interest has an equal chance of being one of the n households responding to the survey. The standard error of the sample mean for WTP can be used to form confidence intervals around $\hat{J}$.

The shape of the distribution of WTP is also of interest since benefits (as well as costs) of improvements in water quality may not be distributed uniformly among the public. The median, or more correctly the median relative to the mean, is frequently used as a single summary measure of the shape of the distribution. For costs, familiar terms are used: regressive (median greater than the mean), proportional (median and mean equal), and progressive (median less than the mean). For benefits, the opposite relations hold. The median also has a natural interpretation in the context of a referendum, the real world analogue of a contingent valuation survey, as the highest flat tax that would be approved by the voters.

Due to varying response rates to sample surveys, the sample weights of 1/n from equation (2) are usually replaced with $1/w_i$ where $w_i$ varies inversely with the response rate of different types of households indexed by i. If we let N correspond to the the number of households in the original equal probability sample, original sample weights can be used. We also present results for reweighting the sample (without imputation) back to the original Census specifications.[2]

The variable for which missing values were imputed is WTPTOT, the total sum of willingness to pay for three marginal quality changes (unusable to boatable, boatable to fishable, fishable to swimmable).

The unusable WTPTOT responses can be divided into four categories:

(1) Don't know/refused (87 respondents or 10% of the respondents);

(2) Protest zeros (136 respondents or 17% of the sample);[3]

(3) Failed edit for WTPTOT larger than .05*income (16 respondents or 2% of the sample); or

(4) Failed edit for low WTPTOT which was inconsistent with income and other responses (9 respondents or 1% of the sample). The variables used in the analysis to follow are defined in Table 2.1.

### 3. An Overview of the Imputation Method Used

The choice of imputation classes is usually the most difficult and influential decision to be made in imputing values for missing responses. We used six different ways to define imputation classes. Each of these forms a series of three to four imputation exercises. They are defined in the two tables (3.1 and 3.2) below.

Table 3.1

Imputation Class Definitions

G1 Series: Eight imputation classes defined by the combinations of three variables [AGE < 45; EDUC < college graduate; USERD=1,0].

G2 Series: Eight imputation classes defined by the combinations of four income categories [income < 15 thousand; $15 \geq income < 30$; $30 \geq income < 45$; $45 \geq income$] and USERD.

CR Series: two imputation classes defined by CART using WTPTOT as the dependent variable and sum of squared deviations as the loss function.

CRA Series: four imputation classes defined by CART using WTPTOT as the dependent variable and the sum of least absolute deviations as the loss function.

CL Series: six imputation classes defined by CART using log(WTPTOT) as the dependent variable and sum of squared deviations as the loss function.

CC Series: seventeen imputation classes defined by CART using a 6 categorized (labeled 1, 2, 3, 4, 5, 6) version of WTPTOT as the dependent variable, ordered twoing as the classification criteria, and a cost criteria of the form i - j where i was the correct class and j was the class the observation was classified as being. The classes in order were [1: 0-25; 2: 26-74; 3: 75-149; 4: 150-249; 5: 250-499; 6: 500+]. The divisions correspond roughly to equal percentiles of the distribution and natural breaks in the data.

Table 3.2

Imputation Type

M: Mean WTPTOT of the imputation class assigned to all observations within that class having missing WTPTOT values.

D: Median WTPTOT of the imputation class assigned to all observation within that class having missing WTPTOT values.

S: Observations ordered sequentially within imputation class by Census region and sampling point. Missing WTPTOT values were replaced sequentially by last usable WTPTOT response. No limit was imposed on the number of times an observation could denoted its WTPTOT value. Sequential imputation assumes that there is positive spatial correlation between observations. While possible sequential variants using the CART procedure were not used.

R: Missing WTP values with in the imputation class are replaced with a valid WTPTOT value chosen randomly (with replacement) from those observations in the imputation class having valid WTPTOT values. An observations could donate its value up to five time.

In addition, three variants of the EM algorithm are used. The EM algorithm is an interrative maximum likelihood procedure (Orchard and Woodbury, 1972; Dempster, Laird, and Rubin, 1977). EM1 estimated values for the missing WTPTOT responses. EM1C is the same as EM1 except that any missing values estimated to be less than zero are set to zero. EM2 estimates values for the the log of WTPTOT of the missing responses, and then takes the the antilog of that response to obtain an estimate of WTPTOT. All of the EM imputations assume normality and a squared error loss function.

The G1 and G2 series imputation classes motivated by results from a prior national survey on willingness to pay for water quality (Mitchell and Carson, 1981). The G1 imputations avoid problems with the missing values on INCOME since valid values are, with minor exceptions, always available for AGE, EDUC, and USERD. Further AGE and EDUC have sizable correlations with income which both theory and available empirical evidence suggest to be the best predictor of WTPTOT. The G2 series uses the G1 values when income is missing in order to emphasis any differences between the two series. Defining imputation classes with CART is discussed in the next section.

The S and the R variants of the G1 and G2 series represent the most the most popular forms of the hot *deck*.[4] The mean of the imputation class is perhaps the most commonly used method of imputing missing values. The median is also used here because we are interested in the median value of the sample WTPTOT and because it has a number of desirable robust properties in many situation. In non-symmetric distributions, the case here, the mean and the median are estimators of different locations and hence can not be compared on an efficiency basis. In our case, statistical test tended to reject both symmetry and normality. Lognormality could not be rejected if fairly coarse grouping effects were allowed at values such as 25 and 100 dollars.

Calculating the three statistics of interest for WTPTOT (the mean, the standard error of the mean, and the median) are fairly straightforward. However, the estimated standard error of the mean of WTPTOT, $\sqrt{(1/n)s^2}$, after imputing values for the missing WTPTOT responses, systematically underestimates the true standard error. Kalton and Kasprzky (1982) give an approximate correction formula for the standard error of the mean based on $\sqrt{(1+I)/r)\sigma^2}$ where r is the number of non-imputed WTPTOT values, $\sigma^2$ is the true estimated variance, and I is the proportionate increase in variance arising from the imputation variance. This formula corrects for two sources of bias arising from the use of the standard error of the mean estimated from the sample after imputing the missing values. The first correction is to multiply by the factor n/r, because the estimate is in actually based only on r observations. For deterministic imputations (M, D, and the EM estimates) I equals zero, but $s^2$, the observed variance, underestimates $\sigma^2$ by a factor of $[(r-1)/(n-1)]$. For stochastic imputation procedures (S and R), $s^2$ is an unbiased estimate of $\sigma^2$, but I equals $[(m/n)(1 - (m/n))]$ where m is the number of imputed WTPTOT values. This correction formula is derived under the assumption that the values are missing at random (within imputation class) and i.i.d. within those imputation classes. Large sample assumptions are also made for the stochastic imputation procedures, and, hence, the estimates of the corrected standard error of the mean given in Tables 5.1 and 5.2, should be considered very rough the process generating the missing values is unknown.

Although we do not know what the true values for the missing WTPTOT responses are, we do have some fairly strong priors based upon past findings (Mitchell and Carson, 1981) and from other evidence available in the present survey. The respondents giving "don't knows," or refusing to answer the WTPTOT question (and to a lesser degree those registering protest zeros) tend to be older, less educated and non-water recreators--all characteristics having sizable negative correlations with WTPTOT. Protest zeros also tend to be associated with negative attitudes toward expenditures on most public goods except fighting crime. The WTPTOT responses, set to missing for being to high, are all from respondents with very low incomes while those few observations set to missing for being too low resemble protest zeros. All of these factors suggest that the true mean WTPTOT should be lower than it is without imputing values for the missing observations. A large number of the respondents with invalid WTPTOT responses also exhibited a pervasive pattern of item non-response on other questions, particularly those dealing with attitudes toward, and knowledge, about water quality. This suggests that the standard error of mean WTPTOT should if anything be increased. Our prior on the median is less clear but we would be suspicious of any imputation procedure that resulted in a sizable shift, particularly an increase.

## 4. A Digression on CART

CART (Breiman *et al.* , 1984) is a set of recursive partitioning procedures which are similar in many respects to the Michigan AID/SEARCH program, which has been used successfully in the past to define imputation classes (Kalton, 1983; Chapman, 1983). While the algorithms underlying CART represent an improvement over existing routines in terms of speed and efficiency, we concentrate here on two new features of CART: surrogate splits, and techniques for determining the optimal number of imputation classes to define.[5]

After locating the predictor split that best minimizes the loss function, CART will search out surrogate splits. Surrogate splits are splits on other variables that best approximate the best predictor split. This is a useful feature, if, as is typically the case, the predictor variables also have a significant number of missing values. A measure similar to a correlation coefficient is available which indicates how much better the surrogate split is at mimicking the best predictor split than sending the cases with missing values down the tree in the same direction as the majority of the cases in that node were sent. The output for the first split of the CL series is reproduced in figure 4.1. Note that competitor splits, the "next best" splits at minimizing the loss function, are also available and in general the same splits.[6]

Perhaps the major problem with the Michigan AID/SEARCH program is the difficulty in deciding when to stop growing the tree. This problem is generic to any recursive partition scheme because they give *much too* optimistic estimates of how well they predict in anything but very large data sets. Standard statistical test of variance explained tend to be misleading. It is worth quoting Kalton (109; 1983) at length here on using AID/SEARCH to define imputation classes:

> In practice there is usually little information to guide the choice of imputation classes to satisfy the missing at random assumption, and therefore attention is mainly focussed on forming classes within which the potential donors are as homogeneous as possible with regard to the survey variable being considered. The general principle is then to form imputation classes that minimize the variance of the survey variable within classes, or equivalently that maximize the variance between classes. This is the principle behind the SEARCH/AID technique used for determining the imputation classes in the last section. In general the SEARCH technique seems a valuable tool for guiding the choice of imputation classes. However, since the technique capitalizes on chance patterns in the data, it should be applied with caution.

This ability of recursive partitioning programs to find meaning in random data (Einhorn, 1972) has lead many survey researchers to reject techniques like AID. CART embodies a solution to this problem.

Breiman *et al.* (1984) after unsuccessful attempts to find optimal rules for stopping the growth of a tree found that the solution lay not in stopping the growth of a tree but in how to "prune off" the lower tree nodes after the tree had "grown" as large as it could.[7] In very large data sets this can be done by dividing the data into two groups, a learning set and a test set. The learning set is used to grow a very large tree and is then used to predict the values of the dependent variable of the test set and the squared error (or other loss function) at each node. In general this loss function will decrease with the number of nodes up to some point, remain flat for a while and then increase as the number of tree nodes gets larger.[8] The first section of this traced out loss function corresponds to true predictability, while the second and third parts correspond to false prediction due to over fitting in the learning set. The tree grown by the learning set can be pruned upward eliminating nodes until the squared error in predicting the test set observations begins to increase.

Unfortunately, most data sets are not sufficiently large to use the test set methodology for pruning and estimation of the true explanatory power of the predictor tree. For data sets the size of most regular surveys (200-2000 observations), it is possible to use v-fold cross-validation to determine how many nodes should be pruned off the tree. This can be done by growing v trees each of which omits N/v percent of the data. Each of the v cross validation trees is used to predict the observations that were not used in growing that tree. The error sum of squares (or other loss criteria) from predicting the out of sample observation by each of the v cross-validation trees is averaged at each tree size to determine how large the main tree should be. The tree sequence output from the CL CART run is shown in figure 4.2. Relative error under a squared error loss function is equivalent to $1-R^2$. Note the U-shaped relative error curve from the cross-validation trees while the relative error indicated by the main tree (resubstitution) continues to decrease as the number of nodes increased. Breiman *et al.* (1984) found on the basis of Monte Carlo experiments that v=10 was appropriate for most purposes and that is what we used to grow and prune the trees in figure 4.3.

The CART trees used to define the imputation classes for the CR, CRA, CL, and CC imputations are shown in figure 4.3. Each of these CART estimations uses either a different form of the dependent variable or different loss functions. The CR CART estimation uses WTPTOT and least squares and produces two imputation classes. The problem with the CR estimation is that it dominated by trying to explain the variance of several large WTPTOT observations and for this reason produces a very unbalanced split. The CRA estimation still used WTPTOT as the dependent variable but minimizes the sum of squared deviations instead. This criteria produces a much more balanced set of imputation classes since it does not put as much weight on the large WTPTOT observations. The CL estimation uses the log of WTPTOT as the dependent variable which further increases the importance of explaining the small and medium WTPTOT observations which are those which we most likely need to be able to predict and separate. The CL estimation produces 6 imputation classes. Another type of prediction rule which gives equal or greater weight to different parts of the data set can be used by assigning different ranges of WTPTOT to a series of ordered classes as the CC CART estimation does.[9] The CC estimation produces 17 imputation class and uses a much larger number of the variables than the other CART estimates which are based primarily on income.

## 5. Results

Table 5.1 displays the results of each of the imputation exercise without using the sample weights. Table 5.2 displays the results of the same imputation exercises using the sample weights.[10]

## 6. Discussion and Concluding Remarks

There is uniformity in the results of the imputation exercises to the extent that they all suggest that mean WTPTOT without correcting for the non-response to the survey and item non-response to the WTPTOT question is biased upward. Using the weighted results of Table 5.2, this bias ranges from thirty to sixty dollars. These amounts that translate into a reduction of 2.5 to 5.5 billion dollars in the estimates of the public's willingness to pay for water quality programs.

There is also fairly uniform agreement that imputing the mean value of the imputation class distorts the median while imputing the median distorts the mean.[11] The correlation between any of the procedures that impute either the mean or median is generally fairly high. All of the random imputations produce data sets with much better distributional properties than any of the systematic imputations.[12]

It is troublesome that the two applications of the EM algorithm produce such different results.[13] Of these two estimates EM2 is probably preferred since the distribution of WTPTOT is fairly close to log normal and because of the large number of negative estimates produced by EM1. The data set used, however, is anything but clean and well behaved, which diminishes our enthusiasm for the maximum likelihood approach.

Two of the series using CART defined imputation classes, CR and CRA, do not perform particularly well. This was to be expected because of the difficulties in explaining WTPTOT in its untransformed form. It is interesting to note that these two imputation series look very much like the G2 series. This is due to the fact that income was the primary determinant of the imputation classes in all three series. The CL and CC imputation classes produce similar results although the correlations between these imputations are only around .5 in the deterministic M and D cases.

The CLR and CCR imputations appear to be of the highest quality. They both have similar and reasonable means, standard errors of the mean, and medians. The distribution of WTPTOT after the imputations looks fairly smooth with none of the spikes associated with the deterministic methods. The better "apparent" success of these two imputation methods attests to the need to consider the range in which the missing values are likely to fall when considering how those values should be imputed.[14] The mean values of the CLR and CCR imputations suggest that the reduction in the estimate of the public's willingness to pay for water quality programs should be approximately three billion dollars. This is a reduction in the estimated total willingness to pay of the American public for water quality improvements from 22 billion dollars to 19 billion dollars.

## Footnotes

1/The welfare economic theory behind CV is covered in Freeman (1979), Just, Hueth, and Schmitz (1982), and Mitchell and Carson (1984). A classic example of a CV survey is Randall *et al.* (1974).

2/Reweighting is not in general an appropriate imputation method if there are missing values on a number of different variables of interest.

3/Protest zeros occur when the respondent is not willing to pay anything because they feel the government waste money and should not be given any more, that it should be possible to raise water quality without paying for it, or that industry should pay the cost since they were responsible.

4/There is no generally agreed upon definition of a hot deck (Panel on Incomplete Data, 1983) and the term has come to generically stand for any technique in which missing values are replaced with values of observations from the current sample. The sequential hot deck described by Bailar and Bailar (1978) was used.

5/CART also has a number of useful features not discussed in this paper which may be of interest to the reader. These include node sub-sampling techniques for large data sets, a variety of classification rules and loss functions, measures of variable importance and tree complexity, and linear combinations of predictor variables. In addition, the book by Breiman *et al.* (1984) presents a number of useful proofs on the statistical properties of CART as non parametric regression and classification techniques.

6/Surrogate splits with a high degree of association which are also competitor splits indicate stability in the model and switching back and forth between close surrogates in different data sets does not indicate instability.

7/Unique pruning schemes require the defining of the notion of tree complexity. The implications of different definitions of tree complexity are discussed at length by Breiman *et al.* (1984).

8/The same general U shaped curve holds for the addition of variables in an ordinary regression framework. See Breiman and Freedman (1983) for a discussion.

9/We also hypothesis that while whether a respondent says 15 or 20 dollars is largely random the difference between say 25 and 30 is not since 25 is a natural dividing line.

10/Miscellaneous notes on Tables 5.1 and 5.2. Twenty one imputed values on G1S failed an edit (WTPTOT > .05*income) were replaced by the closest prior value which would pass the edit. Twenty eight imputed values on G1R failed failed the subsequent edit and new random values were drawn. A constraint that an observation could donate its value 5 times for the R imputations was never binding. Nine was the maximum number of times an observation donated its value in the S imputations. Thirty negative EM1 estimates of WTPTOT (smallest, -333) were set to zero for EM1C. The medians in Table 5.2 were calculated by treating the weights as integer frequencies.

11/EM1 and EM1C also distort the median while EM2 appears to distorts the mean, although this is less clear. Three of the four cases in which CART was used to define the imputation classes, CRA, CL, and CC appear to produce less distortion.

12/However, G1S, G1R, G2S, and the G2R imputations appear to underestimate the standard error of the mean.

13/R. Little has suggested that a correction needs to be made for a downward bias in the antilog transformation used with the EM2 procedure.

14/One obvious direction for future research would be to systematically "punch holes" in a data set with similar distributional properties and problems and then impute values for those variables set to missing.

## Acknowledgements

The author would like to thank Leo Breiman, Robert Mitchell, and Gary Casterline for helpful comments and discussions.

## References

Bailar, B.A. and J.C. Bailar, III (1978), "Comparison of Two Procedures for Imputing Missing Survey Values," *1978 Proceedings of the Section on Survey Research Methods* (Washington: ASA), 462-467.

Breiman, L. and D. Freedman (1983), "How Many Variables Should Be Retained in a Regression Equation," *Journal of the American Statistical Association* , 78, 131-136.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984), *Classification and Regression Trees* (Belmont, CA: Wadsworth).

Carson, R.T. and R.C. Mitchell (1984a), "Non-Sampling Errors in Contingent Valuation Surveys," discussion paper D-120, Resources for the Future, Washington.

Carson, R.T. and R.C. Mitchell (1984b), "Fresh Water Quality Benefits: Findings from a New National Contingent Valuation Survey," paper presented at the Association of Environmental and Resource Economists Annual Meeting, Cornell University, August 1984.

Chapman, D.W. (1983), "An Investigation of Nonresponse Imputation Procedures for the Health and Nutrition Examination Survey," in *Incomplete Data in Sample Surveys* , Report of the Panel on Incomplete Data (New York: Academic).

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, B* ,39, 1-38.

Einhorn, H.J. (1972), "Alchemy in the Behavior Sciences," *Public Opinion Quarterly* , 36, 367-378.

Freeman, M.A. (1979) *The Benefits of Environmental Improvements in Theory and Practice* (Baltimore: Resources for the Future/Johns Hopkins University Press).

Just, R.E., D.L. Hueth, and A. Schmitz (1982) *Applied Welfare Economics and Public Policy* (Englewood Cliffs, NJ: Prentice-Hall).

Kalton, G. and D. Kasprzyk (1982), "Imputing for Missing Survey Responses," *1982 Proceedings of the Section on Survey Research Section* (Washington: ASA), 22-31.

Kalton, G. (1983), *Compensating for Missing Survey Data* (Ann Arbor: ISR, University of Michigan).

Mitchell, R.C. and R.T. Carson (1981), "An Experiment in Determining Willingness to Pay for National Water Quality Expenditures," a report to the U.S.E.P.A. (Washington: Resources for the Future).

Mitchell, R.C. and R.T. Carson (1984), "Using Surveys to Value Public Goods: The Contingent Valuation Approach," draft manuscript, Resources for the Future, Washington.

Mitchell, R.C. and R.T. Carson (1984b), "Fresh Water Quality Benefits: Findings from a New National Contingent Valuation Survey," paper presented at the Association of Environmental and Resource Economists Meeting, Cornell, August 1984.

Orchard, T. and M.A. Woodbury (1972), "A Missing Information Principle: Theory and Applications," *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* , 1, 697-715.

Panel on Incomplete Data (1983). *Incomplete Data in Sample Surveys (3 vol.)* (New York: Academic).

Randall, A., B. Ives, and C. Eastman (1974), "Bidding Games for Valuation of Aesthetic Environmental Improvements," *Journal of Environmental Economics and Management* , 1, 132-149.

Sonquist, J.A., E.L. Baker, and J.N. Morgan (1974), *Searching for Structure* (Ann Arbor: ISR, University of Michigan).


Table 2.1

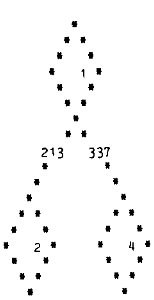
Variables From Willingness to Pay for Water Quality Survey Used

| Name | Description |
|---|---|
| AGE | Age of respondent |
| AREAT | Area type (consolidated metropolitan statistical area to non-metropolitan statistical area) |
| CSIZE | Area size (under 2,500 to 1,000,000+) |
| EDUC | Educational level of respondent |
| INCOME | Household income |
| LOCWATQ | Respondent's perception of water quality of nearest fresh water body |
| POLLUTE | Personal importance of pollution control as national goal |
| POOLOC | Dummy variable for someone in household swimming in a pool or the ocean in past year |
| RACE | Dummy variable for race |
| REGION | 9 Census regions |
| R1USERD | Dummy variable for whether or not the respondent participated in fresh water recreation in past year |
| SEXD | Dummy variable for sex |
| SHORE | Dummy variable for whether anyone in the household recreated near fresh water bodies in past year |
| TDR1USE | Number of water based recreation days by respondent |
| TDUSE | Number of water based recreation days by household members |
| USERD | Dummy variable for whether anyone in the household water recreated in the past year |
| VPOLCST | Attitude on pollution control cost |
| WSPEND | Desire to see increased, decreased, or current level of national expenditures on water quality |
| WTPTOT | Willingness to pay yearly for maintaining and improving (to swimmable) national fresh water quality |
| YISTATE | Number of years lived in current state |

The following variables in addition to WTPTOT have significant item non-response problems: INCOME (7%), WSPEND (16%), and LOCWATQ (8%). Of the missing WTPTOT values, 12% also had INCOME missing, 23% had WSPEND missing, and 15% had LOCWATQ missing.

Figure 4.1

The First CART Split for the CL Series




node 1 was split on variable income
a case goes left if income .le. 1.50e+02
improvement = 4.4e+05 (c. t. = 2.4e+08)

| node | cases | average | standard dev. |
|---|---|---|---|
| 1 | 550 | 0.46e+04 | 0.15e+04 |
| 2 | 213 | 0.38e+04 | 0.14e+04 |
| 4 | 337 | 0.51e+04 | 0.13e+04 |

| | surrogate | | split | assoc. | improve. |
|---|---|---|---|---|---|
| 1 | age | r | 6.35e+01 | 0.25 | 2.4e+05 |
| 2 | yistate | r | 6.25e+01 | 0.19 | 2.2e+05 |
| 3 | educ | s | 2.50e+00 | 0.18 | 1.5e+05 |
| 4 | pooloc | r | 1.50e+00 | 0.14 | 2.0e+05 |
| 5 | csize | s | 3.50e+00 | 0.06 | 3.4e+04 |
| 6 | userd | s | 1.50e+00 | 0.05 | 1.0e+05 |
| 7 | pollute | r | 3.50e+00 | 0.02 | 9.8e+03 |

| | competitor | split | improve. |
|---|---|---|---|
| 1 | age | 6.25e+01 | 2.4e+05 |
| 2 | yistate | 5.85e+01 | 2.2e+05 |
| 3 | pooloc | 1.50e+00 | 2.0e+05 |
| 4 | educ | 2.50e+00 | 1.5e+05 |
| 5 | shore | 1 | 1.4e+05 |

Figure 4.2

Cross-Validation Tree Sequence for CL Cart Estimation

| tree | terminal nodes | cross-validated relative error | resubstitution relative error | complexity parameter |
|---|---|---|---|---|
| 1 | 31 | 0.82 +/- 0.049 | 0.43 | 0.650e+07 |
| 2 | 30 | 0.80 +/- 0.047 | 0.44 | 0.733e+07 |
| 3 | 28 | 0.79 +/- 0.047 | 0.45 | 0.774e+07 |
| 4 | 26 | 0.81 +/- 0.046 | 0.46 | 0.842e+07 |
| 5 | 25 | 0.81 +/- 0.046 | 0.47 | 0.873e+07 |
| 6 | 23 | 0.80 +/- 0.046 | 0.48 | 0.884e+07 |
| 7 | 20 | 0.79 +/- 0.043 | 0.51 | 0.937e+07 |
| 8 | 19 | 0.78 +/- 0.043 | 0.52 | 0.100e+08 |
| 9 | 18 | 0.76 +/- 0.040 | 0.52 | 0.101e+08 |
| 10 | 17 | 0.76 +/- 0.040 | 0.53 | 0.112e+08 |
| 11 | 14 | 0.76 +/- 0.040 | 0.56 | 0.117e+08 |
| 12 | 10 | 0.73 +/- 0.037 | 0.60 | 0.119e+08 |
| 13 | 9 | 0.72 +/- 0.033 | 0.62 | 0.164e+08 |
| 14 | 7 | 0.72 +/- 0.031 | 0.65 | 0.186e+08 |
| 15* | 6 | 0.72 +/- 0.031 | 0.66 | 0.203e+08 |
| 16 | 5 | 0.73 +/- 0.030 | 0.68 | 0.217e+08 |
| 17 | 4 | 0.77 +/- 0.029 | 0.71 | 0.294e+08 |
| 18 | 3 | 0.77 +/- 0.027 | 0.73 | 0.360e+08 |
| 19 | 2 | 0.84 +/- 0.026 | 0.81 | 0.856e+08 |
| 20 | 1 | 1.00 +/- 0. | 1.00 | 0.236e+09 |

Figures 4.3 a, b, c, d

CART Trees for CR, CRA, CL, and CC Series

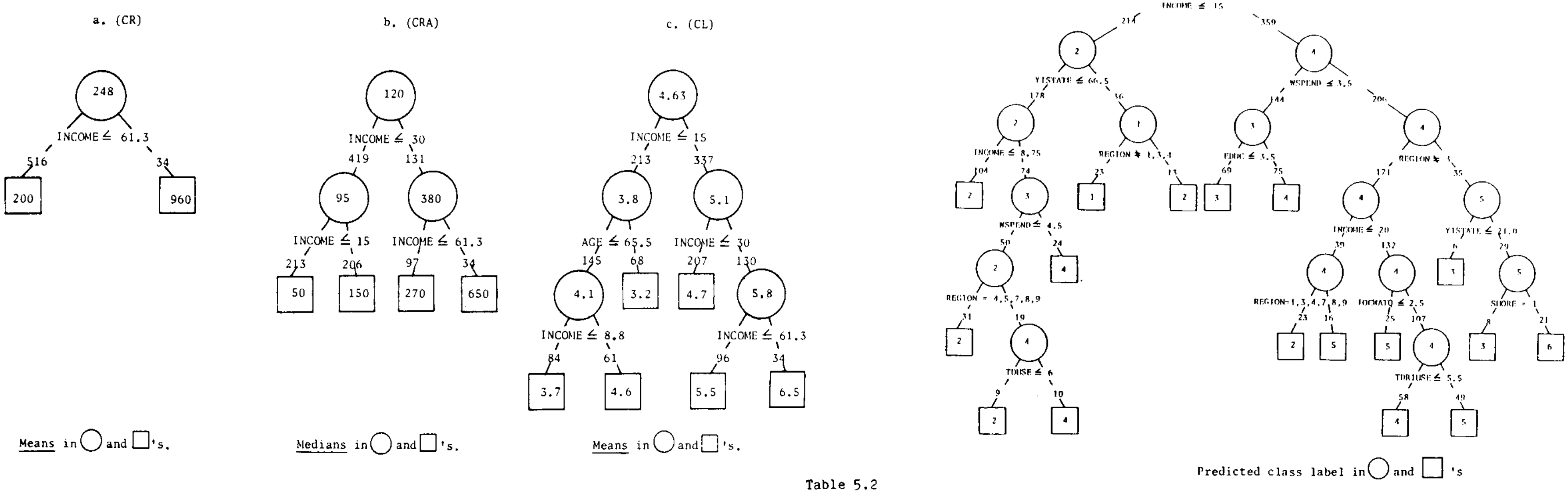



Table 5.1

Unweighted Means, Standard Errors, and Medians

| Imputation Method | MEAN WTPTOT | STAN. ERROR MEAN WTPTOT | CORRECTED S.E.M. WTPTOT | Median WTPTOT |
|---|---|---|---|---|
| None | 275.20 | 25.29 | - | 120 |
| G1M | 261.47 | 17.66 | 25.43 | 160 |
| G1D | 226.07 | 17.80 | 25.63 | 140 |
| G1S | 243.72 | 18.44 | 24.36 | 100 |
| G1R | 248.48 | 18.60 | 25.57 | 105 |
| G2M | 260.35 | 17.93 | 25.82 | 142 |
| G2D | 232.20 | 17.50 | 25.20 | 125 |
| G2S | 251.11 | 18.32 | 24.20 | 120 |
| G2R | 246.55 | 18.33 | 24.21 | 100 |
| EM1 | 251.47 | 18.08 | 26.03 | 141 |
| EM1C | 254.37 | 18.01 | 25.94 | 141 |
| EM2 | 227.65 | 18.49 | 26.63 | 94 |
| CRM | 264.41 | 17.89 | 25.76 | 200 |
| CRD | 230.07 | 17.85 | 25.71 | 103 |
| CRR | 266.90 | 19.87 | 26.24 | 120 |
| CRAM | 259.69 | 18.00 | 25.92 | 110 |
| CRAD | 232.88 | 17.89 | 25.76 | 110 |
| CRAR | 249.49 | 18.63 | 24.61 | 110 |
| CLM | 258.64 | 18.03 | 25.97 | 150 |
| CLD | 235.61 | 17.87 | 25.73 | 150 |
| CLR | 255.86 | 19.19 | 25.35 | 110 |
| CCM | 253.12 | 17.88 | 25.76 | 144 |
| CCD | 226.94 | 17.89 | 25.76 | 100 |
| CCR | 256.78 | 19.09 | 25.22 | 100 |

Table 5.2

Weighted Means, Standard Errors, and Medians

| Imputation Method | MEAN WTPTOT | STAN. ERROR MEAN WTPTOT | CORRECTED S.E.M. WTPTOT | Median WTPTOT |
|---|---|---|---|---|
| None | 255.73 | 25.76 | - | 110 |
| Reweighted | 254.37 | 28.99 | - | 110 |
| G1M | 245.51 | 17.99 | 25.91 | 150 |
| G1D | 209.92 | 18.14 | 26.12 | 120 |
| G1S | 228.14 | 18.89 | 24.95 | 100 |
| G1R | 231.83 | 19.06 | 25.17 | 100 |
| G2M | 242.30 | 18.30 | 26.35 | 130 |
| G2D | 217.98 | 18.08 | 26.03 | 112 |
| G2S | 235.62 | 18.79 | 24.82 | 110 |
| G2R | 230.06 | 18.71 | 24.71 | 100 |
| EM1 | 232.83 | 18.49 | 26.63 | 130 |
| EM1C | 236.58 | 18.40 | 26.50 | 130 |
| EM2 | 209.51 | 18.87 | 27.17 | 89 |
| CRM | 248.17 | 18.21 | 26.22 | 200 |
| CRD | 213.95 | 18.18 | 26.18 | 103 |
| CRR | 254.55 | 21.36 | 28.21 | 110 |
| CRAM | 242.55 | 18.55 | 26.51 | 107 |
| CRAD | 216.29 | 18.24 | 26.27 | 100 |
| CRAR | 232.62 | 19.14 | 25.28 | 100 |
| CLM | 240.77 | 18.41 | 26.51 | 105 |
| CLD | 218.67 | 18.22 | 26.24 | 105 |
| CLR | 237.61 | 19.64 | 25.95 | 100 |
| CCM | 237.21 | 18.26 | 26.30 | 125 |
| CCD | 210.95 | 18.25 | 26.28 | 100 |
| CCR | 242.16 | 19.81 | 26.17 | 100 |

Table 5.3

Correlation Between Imputed Values

| Variables | EM1 | EM2 | G1M | G1D | G1S | G1R | G2M | G2D | G2S | G2R |
|---|---|---|---|---|---|---|---|---|---|---|
| EM2 | .72 | | | | | | | | | |
| G1M | .33 | .19 | | | | | | | | |
| G1D | .35 | .21 | .95 | | | | | | | |
| G1S | .30 | .08 | .17 | .14 | | | | | | |
| G1R | .30 | .15 | .21 | .17 | .06 | | | | | |
| G2M | .78 | .55 | .39 | .37 | .25 | .39 | | | | |
| G2D | .76 | .53 | .35 | .34 | .23 | .38 | .96 | | | |
| G2S | .35 | .26 | .18 | .17 | .32 | .18 | .46 | .42 | | |
| G2R | .46 | .29 | .07 | .06 | .22 | .31 | .46 | .49 | .18 | |
| CRM | .61 | .60 | .08 | .07 | .10 | .29 | .73 | .74 | .27 | .33 |
| CRD | .61 | .60 | .08 | .07 | .10 | .29 | .73 | .74 | .27 | .33 |
| CRR | .06 | .05 | -.02 | -.02 | .07 | .00 | .10 | .09 | .02 | .00 |
| CRAM | .76 | .58 | .25 | .24 | .25 | .40 | .93 | .92 | .44 | .48 |
| CRAD | .76 | .57 | .27 | .27 | .26 | .41 | .93 | .92 | .45 | .50 |
| CRAR | .35 | .29 | .12 | .09 | .04 | .35 | .39 | .39 | .24 | .23 |
| CLM | .78 | .57 | .27 | .27 | .27 | .40 | .95 | .92 | .46 | .48 |
| CLD | .78 | .56 | .28 | .28 | .27 | .40 | .94 | .91 | .46 | .48 |
| CLR | .48 | .49 | .09 | .08 | .10 | .29 | .55 | .52 | .27 | .33 |
| CCM | .63 | .39 | .29 | .33 | .24 | .20 | .52 | .53 | .28 | .39 |
| CCD | .58 | .40 | .31 | .33 | .24 | .17 | .47 | .49 | .20 | .37 |
| CCR | .28 | .12 | .18 | .21 | .07 | .07 | .27 | .27 | .20 | .11 |

| Variables | CRM | CRD | CRR | CRAM | CRAD | CRAR | CLM | CLD | CLR | CCM | CCD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CRD | 1.00 | | | | | | | | | | |
| CRR | .19 | .19 | | | | | | | | | |
| CRAM | .84 | .84 | .13 | | | | | | | | |
| CRAD | .80 | .80 | .12 | .99 | | | | | | | |
| CRAR | .35 | .35 | .06 | .44 | .44 | | | | | | |
| CLM | .81 | .81 | .14 | .97 | .97 | .44 | | | | | |
| CLD | .79 | .79 | .15 | .96 | .96 | .44 | .99 | | | | |
| CLR | .56 | .56 | .06 | .59 | .58 | .14 | .59 | .59 | | | |
| CCM | .22 | .22 | .00 | .48 | .53 | .19 | .54 | .57 | .30 | | |
| CCD | .21 | .21 | .00 | .43 | .47 | .19 | .48 | .50 | .24 | .91 | |