

**Appendix 2:
The Sample Selection Model**

Blank

150

Appendix 2: The Sample Selection Model

2.1 Introduction

The sampling procedure used to obtain the in-person sample for the mail/in-person follow-up portion of the contingent valuation survey was a random selection of households in census tracts surrounding three prechosen interview sites in the five-county study area. The location and number of the interview sites were restricted by the availability of appropriate sites and by budgetary and staffing limitations of the project. As a result, the sampling procedure employed to ensure a representative random sample of the study area was more complex than most standard practices. We developed a computer model to select a sample that was representative of the general population in several important respects (e.g., income, education, race). This appendix explains the details of this model: the problem, the model construction, and the results.

2.2 The Sampling Problem

Under ideal circumstances, a simple random sampling (SRS) procedure will provide the basis for consistent statistical estimation of underlying population parameters. Where SRS procedures are not feasible, cluster sampling, or random sampling within proscribed areas of the study area, are often employed. The clusters themselves are usually selected at random, and various techniques of proportionate random sampling within each cluster are then utilized (Frankel, 1983). This is to allow for ex post facto weighting of responses based on one or more indicators, such as income or race in proportion to their incidence in the study population.

For our study, there were significant constraints on our ability to select our clusters at random and obtain an equal probability sample. We could not perform a simple or even a stratified random sampling procedure to obtain our sample for the mail/in-person follow-up survey over the whole study population. Because of time, cost, and liability concerns, it was decided that enumerators for the study should not be sent out into the neighborhoods of the Greater Houston-Galveston Area to collect questionnaires and interview respondents. Instead, enumerators would be located at three locations around the study area. Respondents could come to these sites at their convenience and be interviewed there. This procedure raised a whole new set of statistical validity concerns for sampling.

If we were to follow the usual procedure of proportionate sampling or simple random sampling from each cluster, a biased sample could be obtained because of the differences in socioeconomic characteristics between households in the sampling areas and households in the whole study area. We felt that selected households should be located close to the interview locations so that the travel time to the interview sites would not become a significant factor in

people's decisions to take part in the interviews. In fact, in order to get a high response rate, we hoped the selected households would be located as close to the interview sites as possible. This requirement would have made the sample more biased if a traditional sampling approach had been used, because the smaller the cluster, the larger the discrepancies between the cluster and the area as a whole could be.

Our problem was thus to select a sample of households, given a limited number of nonrandomly selected small clusters, that permitted valid extrapolation of study results to the whole study area. To our knowledge, neither existing practical experience nor theory in this field were available to provide us with a ready-made solution. Yet, for the validity of our study, this issue had to be given high priority.

2.3 The Sampling Procedure

The basic structure of the sampling procedure we used was as follows. First, we obtained 1990 U.S. census data to examine the socioeconomic information of households in the study area at the census tract level. Around each of the three preselected interview sites, a group of census tracts was chosen. Each of the three census tract groupings had approximately the same population and estimated travel times to their respective interview locations. These three groups of census tracts were used as the basic sampling area from which a final sample of households would be drawn.

Next, a computer model was created to select a subgroup of census tracts from within the basic sampling area. The number of households that were to be sampled from each of the chosen census tracts was also determined by the model. The primary purpose of the model was to ensure the validity (statistically valid representativeness) of our sample. How we achieved this will be discussed in detail below.

Finally, the required number of households was selected randomly from each census tract chosen by the model. A nationally recognized firm specializing in sample selection provided a list of names and addresses of households to invite to be interviewed.

2.4 The Sample Selection Model

2.4.1 Model Objective

Because the census tracts in the basic sampling area were selected around the three predesignated sites, a random selection of households within these nonrandomly chosen tracts might not be representative of the Greater Houston-Galveston Area. Table A2.1 shows the actual statistics of seven socioeconomic variables for the Greater Houston-Galveston Area and the sampling areas around the three interview sites. Clearly, if simple random sampling were performed over the basic sampling area, the selected sample would be expected to be significantly different from

the five-county study area population in terms of income and education. In addition, the seven indicators would not capture all of the differences between census tract populations and other unobserved factors might have significant effects on the households' willingness to pay to support the environmental management plan.³⁹ To address this concern, we decided that the sample should be dispersed over some minimum number of census tracts around each site.

Table A2.1 Comparison of Total Study Area, Cluster Area, and Selected Model Results Means

Area:	Study Area (702 tracts)	Sample Area (65 tracts)	Model Result (36 tracts)
Indicators:			
Mean Income (I)	\$41,515	\$34,984	\$41,932
> H.S. Education (E)	51%	39%	51%
Owner Occupied (tenure) (O)	54%	59%	54%
Caucasian (W)	67%	70%	67%
African-American (A)	18%	17%	19%
Hispanic (H)	21%	24%	21%
Urban (U)	93%	91%	93%

The objective of our sample selection model is that the expected values of important socioeconomic indicators for the selected sample be as close to those of the study population as possible. In order to express the representativeness quantitatively, a set of indicators of representativeness needed to be specified. The differences, or deviations, between expected values of these socioeconomic indicators for the selected sample and the mean values of these indicators for the Greater Houston-Galveston Area study population should be as small as possible (minimized).

³⁹ For example, these several indicators do not address the fact that two of our three interview sites were close to Galveston Bay. The households in our final mail/in-person follow-up sample will thus be more likely to use Galveston Bay for recreational purposes than households selected randomly from the Greater Houston-Galveston Area.

Because we treated minimizing the deviations from each of the seven indicators as separate goals, our model characterized a multiobjective programming problem. The following strategy was adopted to solve the model. First, we calculated the expected mean values of the selected socioeconomic indicators for a proposed sample configuration. We then subtracted these expected means from the mean values of the Greater Houston-Galveston Area population and "normalized" the differences by dividing them by the corresponding standard deviations for each of the indicators. According to statistical theory, each of the normalized deviations should (asymptotically) have a standard normal distribution, $N(0,1)$. Thus, the normalization calculation made all of the indicators comparable. Finally, a minimax mathematical programming model was designed to find an optimal solution. The details of this mathematical programming model will be provided in the following section.

In summary, the objective of the sample selection model used in this study was to minimize the normalized differences between the expected mean values of the socioeconomic indicators for a sample configuration and the mean values for these indicators over the whole Greater Houston-Galveston Area. The model always sought to decrease whichever normalized difference was the greatest by changing the configuration of census tracts and numbers of households selected from each chosen tract in its working solution until no further improvements could be made. The constraints in the model were the requirements of dispersion of the sample around each site, sample size, the minimum and maximum numbers of households that could be selected at each site, and model linearization requirements.

2.4.2 Model Presentation

Decision Variables

The objective of the sampling model was to choose a subset of census tracts from which to sample, and to indicate how many households should be selected from each. The final solution of the model was a list of the number of households that should be selected by a professional sampling firm from each census tract in the basic sampling area. The decision variables of the model are y_i , which denotes whether or not census tract i should be sampled at all, and x_i , which represents the number of households that should be selected from that census tract. The variable y_i was an integer variable that takes on a value of 1 when census tract i is selected, and a value of 0 if the census tract is not selected. The variable x_i is treated as a continuous variable in the model, which was rounded to an integer value for the final count of households to be sampled. The range of i was from 1 to M , where M is the total number of census tracts under consideration for sampling.

Indicator Normalizations

Let us use subscript j with a range from 1 to J to denote each of J indicators to be considered for representativeness (e.g., household income, education, housing tenure, urban versus rural areas, and race). Assuming IND_i^j is the average value of the j th indicator in census tract i ,

IND_{GH}^j is the average value of the j th indicator over the whole study area, and $SDIND_{GH}^j$ is the standard deviation of the j th indicator over the whole Greater Houston-Galveston Area, then the expected average value of the j th indicator in the sample would be $(\sum_{i=1}^M IND_i^j * x_i * y_i) / (\sum_{i=1}^M x_i * y_i)$, where $x_i * y_i$ is the number of households selected in census tract i and $\sum_{i=1}^M x_i * y_i$ is the total number of households selected to be invited to participate in the study. $\sum_{i=1}^M IND_i^j * x_i * y_i$ is the expected total value of indicator j in the sample. The normalized deviation between the expected mean value of the j th indicator in the selected sample and the mean in the whole study area would be:

$$NDIND^j = \frac{|((\sum_{i=1}^M IND_i^j * x_i * y_i) / \sum_{i=1}^M x_i * y_i) - IND_{GH}^j|}{SDIND_{GH}^j / \sqrt{\sum_{i=1}^M x_i * y_i}} \quad (2.1)$$

According to statistical theory, the normalized variable $NDIND^j$ should have an asymptotically standard normal distribution.

Objective Function

With the formulation provided for the normalized deviation above, the objective function of the model can be written as:

$$\begin{matrix} MIN \\ x_i, y_i \end{matrix} \quad \begin{matrix} MAX \\ j \end{matrix} \quad NDIND^j \quad (2.2)$$

Because there are product terms of the decision variables x_i and y_i in the objective function, this is a nonlinear minimax mathematical programming problem.

Constraints

A. Total Sample Size

The total sample size generated should be within an acceptable range, which can be expressed as:

$$N_1 \leq \sum_{i=1}^M x_i * y_i \leq N_2 \quad (2.3)$$

where N_1 and N_2 are respectively lower and upper bounds of the total number of households which would be invited to participate in the study.

B. Minimum and Maximum Numbers of Households to be Selected at Each Site

Due to limited space at interview locations and a need to simplify the allocation of enumerators to interview sites, the number of households selected around each interview site had to be within specified ranges. To satisfy this requirement, following constraints were constructed.

$$\begin{aligned} S_1 &\leq \sum_{i=1}^{M_1} x_i y_i \leq B_1 \\ S_2 &\leq \sum_{i=M_1+1}^{M_2} x_i y_i \leq B_2 \\ S_3 &\leq \sum_{i=M_2+1}^M x_i y_i \leq B_3 \end{aligned} \tag{2.4}$$

where subscripts 1, 2, and 3 denote the three interview sites. S and B denote the upper limit and lower limit of the number of households chosen around each site. M_1 represents the number of census tracts available to be selected by the model around site 1. M_2 is the total number of census tracts available for Sites 1 and 2, so $M_2 - M_1$ represents the number of census tracts around Site 2 that are available to be selected by the model. Likewise, $M - M_2$ represents the number of census tracts available to be selected for Site 3.

C. Dispersion Requirement

To reduce systematic bias in the results of our study arising from unobservable factors that differentiate census tracts, the sample should not be selected from too few census tracts around each site. We required that the number of census tracts selected around each site be at least some minimum number of tracts. This introduced following three constraints:

$$\sum_{i=1}^{M_1} y_i \geq m_1$$

$$\sum_{i=M_1+1}^{M_2} y_i \geq m_2 \quad (2.5)$$

$$\sum_{i=M_2+1}^M y_i \geq m_3$$

where m_1 , m_2 and m_3 are lower bounds on the number of census tracts selected around the three sites. M_1 , M_2 , and M are as defined above.

D. Statistical Validity Requirement of the Deviation Normalization

In equation 2.1, we calculated the mean (expected value) of various indicators for the sample by calculating the product of the number of households selected from census tract i and the mean value of the variable in that census tract, and summing over all selected census tracts. We normalized this mean value by the standard deviation of each variable over the study area. This procedure was based on the assumption that the number of the households selected from each census tract would be large enough so that the households selected would be able to represent the census tract as a whole. If the number of households selected in a census tract is too small, even though the expected mean is still theoretically equal to the population mean, the distribution of the indicator's values may not sufficiently reflect the actual distribution in the census tract. This required that we provide for a minimum number, or lower bound, of households to be selected from each census tract.

We also required that the number of households selected from a given census tract have an upper bound so that the sample would not be selected from too few census tracts. This was done to satisfy our dispersion requirement.

Given these needs for upper and lower bounds on the numbers of households selected from each tract, the decision variable x_i was constrained by the following inequalities:

$$L_i \leq x_i \leq U_i \quad \text{for all } i \quad (2.6)$$

where L_i is the lower bound and U_i is the upper bound.

In fact, two sets of lower bounds and upper bounds were employed in the model. The first set assigned the same absolute values for the upper and lower bounds to all census tracts selected by the model to be sampled ($y_i = 1$). The second set of bounds on x_i established minimum and maximum proportions of the total number of households in census tract i ; these had to be satisfied for all tracts being sampled.

The sizes of census tracts around the three interview sites vary greatly. The first set of bounds treats all census tracts as if they were all the same size. The effect of this constraint is that the solution would tend to favor smaller census tracts. To adjust for this, the proportional bounds limit oversampling of smaller tracts and balance the opportunity for selection among all available tracts. Constraint 2.6 was thus actually formulated in terms of two sets of upper and lower bounds:

$$L \leq x_i \leq U \quad (2.6a)$$

$$l \cdot HH_i \leq x_i \leq u \cdot HH_i \quad (2.6b)$$

where HH_i is the total number of households in census tract i , L is absolute lower bound, U is the absolute upper bound, l is the lower bound on the proportion of households chosen from sampled census tracts, and u is the upper bound proportion.

2.4.3 Solving the Model

The model presented in last section is a nonlinear minimax programming model which cannot be solved by readily available computer software. To solve the model, several modifications needed to be made.

Simplification of the Objective Function

The minimax problem given in the last section can be simplified into a standard minimization problem by introducing a single variable d as the objective in the following way:

$$\begin{array}{ll} \text{MIN} & d \\ & x_i, y_i \end{array} \quad (2.7)$$

subject to:

$$NDIND^j \leq d, \quad \text{for all } j=1, J \quad (2.8)$$

The constraints 2.8 require that d be not less than any normalized value of indicator, and in fact among J constraints, at least one constraint is binding because the objective function 2.7 requires d to take the value of the largest of the minimized normalized indicator values. The formulation 2.7 and 2.8 is thus an alternative way to express the objective function 2.2 of the minimax problem.

Model Linearization

To linearize the model, we imposed a requirement that $x_i=0$ if, and only if, $y_i=0$. The term $\sum_{i=1}^M IND_i^j * x_i * y_i$ in the model thus becomes $\sum_{i=1}^M IND_i^j * x_i$, which is a linear term. The term $\sum_{i=1}^M x_i * y_i$, which is the total number of households that should be selected, reduces to $\sum x_i$. By requiring that the total number of households selected over the whole study area to be fixed at N , function 2.1 becomes the linear function 2.10, below.

To ensure that $x_i=0$ if and only if $y_i=0$ was incorporated into the model, we reconstructed equation 2.6 as follows:

$$L_i y_i \leq x_i \leq U_i y_i, \quad \text{for all } i. \quad (2.9)$$

Under this constraint, when $y_i=0$, $x_i=0$, and when $x_i=0$, $y_i=0$. When $y_i=1$, we have $L_i \leq x_i \leq U_i$, which is constraint 2.6; when $x_i > 0$, y_i should be equal to 1. Inequality 2.9 combines constraint 2.6 and the need for model linearity.

Corresponding to equations 2.6a and 2.6b, we have 2.9a and 2.9b as follows:

$$L y_i \leq x_i \leq U y_i \quad \text{for all } i \quad (2.9a)$$

$$l \cdot HH_i y_i \leq x_i \leq u \cdot HH_i y_i \quad \text{for all } i. \quad (2.9b)$$

Thus, the nonlinear minimax problem has been transformed into a linear programming problem that can be solved by routine linear programming computer software. The mathematical formulation of the model can be summarized as follows:

$$\begin{array}{ll} \text{MIN} & d \\ & x_i, y_i \end{array} \quad (2.7)$$

subject to:

$$NDIND^j \leq d, \quad \text{for all } j=1, J \quad (2.8)$$

$$\sum_{i=1}^M x_i = N \quad (2.3')$$

$$S_1 \leq \sum_{i=1}^{M_1} x_i y_i \leq B_1 \quad (2.4a)$$

$$S_2 \leq \sum_{i=M_1+1}^{M_2} x_i y_i \leq B_2 \quad (2.4b)$$

$$S_3 \leq \sum_{i=M_2+1}^M x_i y_i \leq B_3 \quad (2.4c)$$

$$\sum_{i=1}^{M_1} y_i \geq m_1 \quad (2.5a)$$

$$\sum_{i=M_1+1}^{M_2} y_i \geq m_2 \quad (2.5b)$$

$$\sum_{i=M_2+1}^M y_i \geq m_3 \quad (2.5c)$$

$$L y_i \leq x_i \leq U y_i \quad \text{for all } i \quad (2.9a)$$

$$l \cdot HH_i y_i \leq x_i \leq u \cdot HH_i y_i \quad \text{for all } i. \quad (2.9b)$$

where

$$NDIND^j = \frac{|\left(\sum_{i=1}^M IND_i^j \cdot x_i\right) / N - IND_{GH}^j|}{SDIND_{GH}^j / \sqrt{N}} \quad (2.10)$$

2.4.4 Model Application

Data Preparation

From the 702 census tracts in the five-county study area around Galveston Bay, approximately 20 census tracts were chosen around each of the three interview locations. A total of 65 census tracts ($M=65$) were selected in this first round as the basic sampling area. The groups of census tracts selected around each interview site had approximately the same total population, and estimated travel times to the corresponding interview sites were roughly equivalent.

Based upon the sample stratification criteria and model specifications of two recent landmark studies regarding contingent valuation methodology (Arrow, et al., 1992, and McClelland, et al., 1992), we selected seven socioeconomic and demographic indicators to compare the sample area populations around the three interview sites with the five-county (study area) population characteristics: (1) income, (2) education, (3) percent urban (versus rural), (4) tenure (whether the respondent owns or rents the residence), and (5-7) race/ethnicity (Caucasian, African-American, Hispanic). We used the Summary Tape Files (STF3) of the 1990 United States census for Texas to obtain data at the census tract level for all five counties in the study area on each of the above seven indicators. We calculated means and standard deviations for the seven socioeconomic indicators within each and over all census tracts, for both the 65 census tract areas around the three interview locations and the entire five-county area, as required by the model. Table A2.1 above shows the means for the overall study area and the basic sampling area. Means of several of the indicators from the basic sampling area appear quite different from the study area means.

Some simplifications in the data were made in order to calculate the means and standard deviations for each variable. For example, the census data for educational attainment are divided into seven categories ranging from (1), less than ninth grade, to (7), graduate or professional degree. For our purposes, we split the population into two groups: those with educations completed up through high school, and those with more than a high school education. Our indicator for education was the percent of the population in the census tract that had achieved more than high school education.

Another indicator of some concern was the proportion of persons of Hispanic origin. Under the race delineation tables in the census data files, both Caucasians and African-Americans can be of Hispanic origin, so there could be some overlap between the Hispanic and the African-American and Caucasian populations. This variable was included as a separate indicator, nonetheless, since we wanted to include this important segment of the population in the study.

The standard deviations of the area-wide indicators were also calculated from the census data set. While the standard deviation for income follows the textbook formulation for the standard deviation of a continuous variable, the standard deviations for variables 2-7 are the binomial standard deviations for population parameters. These binomial standard deviations can be easily

calculated from the mean values by the formula $SD = (p(1-p))^{1/2}$, where p is the proportion of households over the area that take on one characteristic, $(1-p)$ is the balance, and SD is the standard deviation.

Model Results

This model construction is a linear programming algorithm that combines both integer and continuous decision variables. The model combines an integer programming routine, which selects the census tracts that will have greater than zero elements, with a continuous linear programming algorithm that defines the number of households, chosen from each census tract i . A Lotus-supported XA linear programming software was used in solving the model. The model has 131 decision variables and 284 constraints, and the XA software provided model results in two minutes with a 386 IBM-compatible computer. Lotus's routine functions provided us with a very convenient way to run the model repeatedly, which was necessary for us to perform sensitivity analyses of many lower and upper bounds for the constraints on the model's objective value (d).

We tried numerous specifications of absolute and proportional upper and lower limits because of the trade-offs between the model's objective value d , which represents the degree of similarity between the characteristics of a selected sample and those of the whole study population, and the lower and upper limits on the numbers of households that can be selected from a chosen tract and still ensure the dispersion of the sample. We also varied the minimum numbers of census tracts required to be selected around each enumeration site. Among the several acceptable simulation results, we selected one that best satisfied our combined objective of minimum d with maximum dispersion. Table A2.2 shows the three best model solutions and the trade-offs on the bounds between them. The researchers had to judge how best to resolve the trade-offs between the upper and lower values for the constraints that provide for dispersion and the d value of the objective function. We chose Model 1 as our best solution since increasing the number of census tracts around each site did not improve the d value of the solution significantly, and reducing the maximum number of households that could be selected around each site increased the d value without much change in the dispersion of households within the selected tracts.

The parameter values of the constraints that were used in the final model (Model 1) are shown in Table A2.2 below. The solution to the objective function (the d value) for this final model was 0.25. The expected mean values of the seven socioeconomic variables calculated from the output of the model run under this specification are shown in Table A2.1 under Model Results, to facilitate comparison with study area means and the means that would be obtained under a simple proportionate sampling procedure from all 65 tracts in the basic sampling area. The values for the seven indicators are almost the same between the whole study area and the model results for the 36 census tracts.

The model results provide a basis for sampling from a population, not the final sample self. Even when the sampling procedure is valid, both in the overall configuration of the census tract/household proportions mix and in the subsequent random selection of the actual households to be surveyed, the final data set may still be biased due to other reasons (e.g., selection effects due to the salience of the research topic to various respondents). Before this model is used, all factors that might influence participation should be considered thoroughly so that as many of these factors as possible can be integrated into the model's indicators or constraints. In our case, we did not include recreational use of the bay as an indicator of representativeness, largely because these data were not available to us from the census.

Blank

1/6/6