LBL-20333
UC-350

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

APPLIED SCIENCE DIVISION

RECEIVED
LAWRENCE
BERKELEY LABORATORY

DEC 5 1989

LIBRARY AND
DOCUMENTS SECTION

A Study of Aggregation Bias in Estimating the Market for Home Heating and Cooling Equipment

D.J. Wood, H. Ruderman, and J.E. McMahon

May 1989

TWO-WEEK LOAN COPY

This is a Library Circulating Copy
which may be borrowed for two weeks.

APPLIED SCIENCE
DIVISION

LBL-20333
c.2

Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. Neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or

Lawrence Berkeley Laboratory Library

University of California, Berkeley

LBL-20333

A STUDY OF AGGREGATION BIAS IN ESTIMATING THE MARKET FOR HOME HEATING AND COOLING EQUIPMENT*

David J. Wood, Henry Ruderman and James E. McMahon

Energy Analysis Program
Applied Science Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

May 1989

ABSTRACT

Econometricians frequently propose parametric models which are contingent on an underlying assumption of rational economic agents maximizing their utility. Accurate estimation of the parameters of these models depends on using data disaggregated to the level of the actual agents, usually individual consumers or firms. Using data at some other level of aggregation introduces bias into the inferences made from the data. Unfortunately, properly disaggregated data is often unavailable, or at least, much more costly to obtain than aggregate data.

Research on consumer choice of home heating equipment has long depended on state-level cross-sectional data. Only recently have investigators been able to build up and successfully use data on consumer attributes and choices at the household level. A study estimated for the Electric Power Research Institute REEPS model is currently one of the best of these.

This paper examines the degree of bias that would be introduced in that study if only average data across SMSAs or states were used at several points in the investigation. We examine the market shares and elasticities estimated from that model using only the mean values of the exogenous variables, and find severe errors to be possible. However, if the models were calibrated on only aggregate data originally, we find that proper treatment allows market shares and elasticities to be found with little error relative to the disaggregate models.

* This work was supported by the Assistant Secretary for Conservation and Renewable Energy, Office of Building and Community Systems, Building Equipment Division of the U.S. Department of Energy, under Contract No. DE-AC03-76SF00098.

A STUDY OF AGGREGATION BIAS IN ESTIMATING THE MARKET FOR HOME HEATING AND COOLING EQUIPMENT

David J. Wood, Henry Ruderman and James E. McMahon

1. Introduction

The uncertain status of future energy supplies relative to demand has been a point of concern in this country for more than a decade. A number of empirical studies have been conducted to estimate the factors that influence demand in one fashion or another [2,3,4,5,6,9,12]. Some of these have been in support of computer models that simulate long-run national energy demand [6,12].

These econometric studies take a common view that energy-related decisions are made by rational economic agents seeking to maximize their own utility. For the studies listed above, that decision is the choice of a specific fuel or technology for residential space heating. The statistical techniques used to estimate the determinants of that decision fall into a general category called discrete choice modeling. In this context, the issue of aggregation bias can be critical.[1]

Aggregation bias is a general term for errors induced by a mismatch of the economic theory (of individual consumers) and the level of aggregation of data (e.g., mean values at the state or national level). It has two distinct forms:

1) bias in the prediction of market shares and elasticities for aggregate groups, using parameters from a household choice model and representative values of independent variables (frequently the means) for the group; and

2) bias in the parameters of a household utility maximization choice model estimated on data aggregated at some regional level.

This paper offers examples of the potential severity of both types of bias. The former is found to be potentially dangerous, leading to predictions of market share which are severely in error; the latter is found to be much less of a problem if handled correctly.

The present paper is organized as follows: Section 2 contains a brief introduction to the theory of aggregation bias. That section is strongly influenced by McFadden and Reid [16], but has been cast in the context of consumer choice of space heating systems.

The first form of aggregation bias (prediction of aggregate market shares or elasticities using only mean values) is taken up in Section 3. This bias has been extensively analyzed in the literature on transportation modal choice [10,11,16,17,18,19,20]. That literature has largely focused on bias-reduction methods generally called *classification* techniques; they depend on the relationship between the bias and the covariance matrix of the independent variables. These techniques were generally proposed as being computational short-cuts to the goal of bias-free market share estimates. That goal can also be achieved through *sample enumeration,* the calculation of predicted choice probabilities for every member of a random sample from the population.

[1] Readers interested in a brief review of the literature on home heating appliance choice are referred to Wood, Ruderman, and McMahon [22]. More extensive reviews may be found in Dohrmann [4] or Hartman [7]. A complete review of issues and techniques in discrete choice modeling is in Amemiya [1].

We show a sample enumeration method of estimating market shares and elasticities which is essentially bias-free (although computationally intensive). The method captures arc elasticity effects of relatively large perturbations in the independent variables at the same time it estimates the point elasticity. In an econometric model of residential heating equipment choice estimated by EPRI [6], using arc elasticities allows significant improvements (relative to the point elasticity) in predicting market shares under large (e.g., greater than 20%) changes of the independent variables.

The second form of the bias is discussed in Section 4. Despite general acknowledgement of some inherent error in estimating a household choice model from aggregate data, the appliance choice literature has long depended on just such data [2,3,4,9]. There is a body of literature indicating that such models may lead to biased parameter estimates [8,11,13,16,18,19]. Again working with the EPRI model [6], we present an empirical example of the potential severity of aggregation bias, and conclude that if the aggregate coefficients are used in a properly "disaggregated" fashion, then the resulting calculation of market shares or elasticities can be without serious error.

In Section 5, we examine the models estimated on aggregate data and test the significance of their parameters and implied elasticities for predicting new market shares. Though the estimated parameters on aggregate data are significantly different from the disaggregated estimates, those differences do not generally lead to economically significant errors.

A brief summary and discussion of results is found in Section 6. Acknowledgements and references to papers mentioned in the text follow.

There are two appendices: the first discusses the aggregation of EPRI's database into groups by SMSA and geographical region; and the second discusses several probit versions of the models estimated by EPRI, with particular attention to the asymptotic nature of the theory suggested by McFadden and Reid [16].

2. Introduction to Aggregation Bias Theory

In 1975, McFadden and Reid published an analysis of the role of aggregation bias in estimating parameters for consumer choice models [16]. That paper dealt with the estimation of probit models for dichotomous choice problems (i.e., choice between only two alternatives).[2] The authors developed their analysis in the context of choice of transportation mode. In this section, we will briefly review issues in their work that are relevant to the current paper. Our presentation will be in the context of consumer choice of space heating systems, but will follow closely the development by McFadden and Reid.

Almost all discrete choice estimation models start from an assumption that the utility a consumer derives from his choice is some linear combination of his own and the choices' attributes, plus a random term with some assumed distribution. Thus, utility is assumed to have a form like this:

$$\text{Utility} \;=\; \sum \beta x + \epsilon \;=\; \mathbf{BX} + \epsilon$$

where $\mathbf{X}$ is a vector of exogenous variables,
$\mathbf{B}$ is a vector of the coefficients β which the researcher wishes to estimate, and
ϵ is a random term.

The probability that a consumer will choose any particular alternative from a given set of choices is just the probability that that alternative is perceived as providing more utility than any other. In a choice between two alternatives, this leads to the following equations:

$$\begin{aligned} PROB\left\{\text{consumer chooses alt 1}\right\} &= PROB\left\{\text{utility of alt 1} > \text{utility of alt 2}\right\} \\ &= PROB\left\{\mathbf{BX}_1 + \epsilon_1 > \mathbf{BX}_2 + \epsilon_2\right\} \\ &= PROB\left\{\epsilon_2 - \epsilon_1 < \mathbf{B}(\mathbf{X}_1 - \mathbf{X}_2)\right\} \\ &= PROB\left\{\epsilon_2 - \epsilon_1 < \mathbf{BZ}\right\} \end{aligned} \tag{2.1}$$

Note that the numeric subscript is associated with each alternative: $\mathbf{X}_i$ refers to the vector of exogenous variables associated with the ith alternative, and $\mathbf{Z}$ is the vector of differences $\mathbf{X}_1 - \mathbf{X}_2$. As long as utility is linear in these variables, it is only the difference between these two vectors that is significant.

Thus, the probability that the consumer will choose alternative 1 depends on the probabilistic nature of the difference between the two random terms. If we assume that the difference has a standard normal distribution, then we have the standard probit model, and equation 2.1 above will look like this:

$$PROB\left\{\text{consumer chooses alternative 1}\right\} \;=\; \Phi(\mathbf{BZ}) \tag{2.2}$$

where Φ is the cumulative standard normal distribution function.

[2] However, their results are at least indicative of the kind of bias that occurs in multichotomous choice models, or models other than the probit (e.g., the logit). Reid [17] took advantage of the similarity of the normal and logistic distributions to extend the analysis to binary logit models.

Then the best prediction of the fraction of N particular individuals choosing the first alternative is:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N}\Phi(\mathbf{B}\mathbf{Z}_i).$$

The quantity $\bar{P}$ is the expectation of the response probability of the empirical distribution (over the N individuals sampled) of the vector of the independent variables $\mathbf{Z}_i$. It is the expected *market share* of alternative 1 in the group of N individuals.

To estimate $\bar{P}$ for an entire state, state k say, it is appropriate to instead consider the expectation of the response probability of the underlying distribution of the independent variables.[3] If that distribution is jointly normal with mean $\bar{\mathbf{Z}}_k$ and covariance matrix $\mathbf{A}_k$, then the term $y = \mathbf{B}\mathbf{Z}_k$ is distributed normally with mean $\mathbf{B}\bar{\mathbf{Z}}_k$ and variance $\sigma_k^2 = \mathbf{B}\mathbf{A}_k\mathbf{B}$. Therefore, the expectation of the response probability is

$$\bar{P}_k \underset{N_k\to\infty}{\longrightarrow} \int_{-\infty}^{+\infty}\Phi(y)\frac{1}{\sigma_k}\phi\left(\frac{y-\mathbf{B}\bar{\mathbf{Z}}_k}{\sigma_k}\right)dy$$

where $\phi(\ )$ is the normal probability density function.

McFadden and Reid simplify this equation further by using the convolution properties of the normal distribution. That simplification results in the following relationship:

$$\bar{P}_k \underset{N_k\to\infty}{\longrightarrow} \Phi\left(\frac{\mathbf{B}\bar{\mathbf{Z}}_k}{\sqrt{1+\sigma_k^2}}\right) \tag{2.3}$$

This result can also be seen by direct consideration of the condition necessary for an individual to choose the first alternative (equation 2.1).

The difference between equations 2.2 (for individuals) and 2.3 (for aggregate regions) is the fundamental bias derived by McFadden and Reid. Its effect can be seen quite easily for aggregation bias of the first kind, the prediction of aggregate market shares. Equation 2.3 suggests that when estimating $\bar{P}_k$, the aggregate explanatory variables $\bar{\mathbf{Z}}_k$ can be viewed as having the same relative weights $\mathbf{B}$ as in the disaggregated model, but that the variables (or the weights) should be divided by the term $(1+\sigma_k^2)^{0.5}$ before applying the nonlinear transformation $\Phi(\ )$. If a region is completely homogeneous with respect to the independent variables $\mathbf{Z}_k$ (all individuals holding the same values $\bar{\mathbf{Z}}_k$), then $\mathbf{A}_k = 0$, and the term $\sigma_k^2 = \mathbf{B}\mathbf{A}_k\mathbf{B}$ will have no effect.

Similarly, when equation 2.3 is substituted into the least-squares equations typically used to estimate a coefficient vector (call it $\mathbf{D}$) on aggregate data (e.g., state-wide averages):

$$\Phi^{-1}(\bar{P}_k) = \mathbf{D}\bar{\mathbf{Z}}_k + \epsilon_k \tag{2.4}$$

$$\Phi^{-1}\left(\Phi\left(\frac{\mathbf{B}\bar{\mathbf{Z}}_k}{\sqrt{1+\sigma_k^2}}\right)\right) = \mathbf{D}\bar{\mathbf{Z}}_k + \epsilon_k \tag{2.5}$$

we find that, in the limit as the number of observations increases, the estimated coefficients $\mathbf{D}$ are functions of the true vector $\mathbf{B}$ and multiplicative bias factors in the terms $(1+\sigma_k^2)^{-0.5}$ for all k regions. If the individual aggregate data (i.e., state-wide averages) have a constant covariance matrix $\mathbf{A}_k = \mathbf{A}$ for all regions k, then the bias terms reduce to a simple form:

$$\mathbf{D} = \frac{1}{\sqrt{1+\mathbf{B}\mathbf{A}\mathbf{B}}}\mathbf{B} \tag{2.6}$$

Only if all regions are completely homogeneous (i.e., all $\mathbf{A}_k$ are zero), will $\mathbf{D}$ be a consistent estimator of $\mathbf{B}$.

[3] The two are equivalent as the number of individuals sampled, N_k, goes to infinity.

3. Naive Estimates of Market Shares and Elasticities

This section is concerned with aggregation bias of the first kind: bias in the estimation of aggregate market shares and elasticities from a disaggregated model. This problem has been extensive analyzed in the literature on transportation modal choice [10,11,16,17,18,19,20].

The transportation literature has largely focused on bias-reduction methods generally called classification techniques. In very simplified form, these techniques work by dividing the population into groups for which the set of exogenous variables is more or less homogeneous. Market shares are calculated for each group using the mean values, and the results combined with appropriate weights to get an overall probability. The number of groups necessary depends on the number of independent variables and how each contributes to the overall covariance matrix. Past efforts have used numbers ranging from as few as four [18] up to several dozen or more.

Naive estimation refers to the use of aggregate mean values of the exogenous variables in analytic expressions for market share or elasticity (without regard to the variance of those variables). It can be viewed as a limiting case of classification techniques, with the population being divided into exactly one group. It is the simplest, and probably the most commonly used, method of getting aggregate market shares. Unfortunately, it can produce highly inaccurate estimates.

Table 1 shows the results of naive estimation of market shares for the population in EPRI's disaggregated database. The relative error between the aggregate and disaggregate predictions ranges from a quite tolerable 0.5% (for electric forced air with central cooling) to a serious 50% (for electric baseboard systems without central cooling).

Table 1

Effect of Aggregation on Market Share Estimation

	estimated market shares:			
system	disaggregated enumeration	"naive"	difference	relative difference
central cooling	0.5264	0.5908	-0.0774	-12.1 %
gas forced air, with ac	0.3097	0.3720	-0.0623	-20.1 %
oil forced air, with ac	0.0266	0.0327	-0.0061	-22.9 %
elec forced air, with ac	0.1595	0.1587	0.0008	0.5 %
heat pump	0.1150	0.1280	-0.0130	-11.3 %
elec baseboard, with ac	0.0306	0.0275	0.0032	10.3 %
gas forced air, w/o ac	0.1915	0.1749	0.0165	8.6 %
gas hydronic, w/o ac	0.0063	0.0053	0.0010	15.8 %
gas non-central, w/o ac	0.0047	0.0044	0.0003	5.5 %
oil forced air, w/o ac	0.0321	0.0267	0.0054	16.9 %
oil hydronic, w/o ac	0.0318	0.0184	0.0134	42.2 %
oil non-central, w/o ac	0.0021	0.0018	0.0003	18.2 %
elec forced air, w/o ac	0.0341	0.0216	0.0125	36.6 %
elec baseboard, w/o ac	0.0559	0.0280	0.0279	49.9 %
all gas systems	0.5121	0.5566	-0.0445	-8.7 %
all oil systems	0.0927	0.0796	0.0131	14.2 %
all conventional elec	0.2802	0.2358	0.0444	15.8 %

A similar bias effect occurs in elasticities which are simply evaluated at the mean values of the exogenous variables. Such an elasticity is, at best, the elasticity of a household all of whose

exogenous variables are at their respective means (which may be completely unrepresentative of the population). In general, it will be a poor estimator of the mean of the individual household elasticities.

However, even the mean of the individual elasticities for all households in the database is not an ideal measure of how an overall market share will respond to changes in an exogenous variable. A better measure can be obtained by taking a weighted mean of household elasticities, where the weights are the probabilities that the household selects the alternative whose elasticity is being sought. That calculation results in an overall market elasticity.[4] Let

e_x^i	=	household elasticity of alternative i with respect to variable x
P_i	=	household probability of selecting alternative i
MS_i	=	estimate of overall market share of alternative i $= \frac{1}{N}\sum P_i$

Then the overall market share elasticity of alternative i with respect to x is:

$$\eta_x^i = \frac{\partial MS_i}{\partial x} \cdot \frac{x}{MS_i} = \frac{\sum P_i e_x^i}{\sum P_i} \tag{4.1}$$

We estimated these market elasticities by simulating economic changes in the dataset. Our method is to perturb by a small fraction δ (e.g., $\delta = 0.01$) the exogenous variable and calculate new market shares (under the perturbed conditions) by summing the household probabilities. The difference between the new and unperturbed market shares is due to the change in the exogenous variable. Dividing that difference by the perturbation size δ and the original market share gives an estimate of the overall market arc-elasticity for that perturbation:

$$\eta(\delta) = \frac{MS_{new} - MS_{old}}{\delta MS_{old}} \tag{4.2}$$

Both the new market shares and the resulting market share arc elasticities above are functions of the perturbation size δ. Due to the nature of the logit model, they can be approximated quite well by second-order polynomials.[5] We can exploit this relationship to express a whole range of arc elasticities using only the three parameters of a quadratic curve.

We do this by calculating the arc elasticity above for several discrete perturbations over the range -33% to +50%. Using ordinary least squares techniques, these values are regressed on a quadratic curve in δ. If the resulting fit is close enough, the parameters of the fitted quadratic convey all the information necessary to find any arc elasticity in the fitted range. Furthermore, the constant term in that quadratic (the intercept) is the limiting value of $\eta(\delta)$ as the perturbation size goes to zero. As such, it is the point elasticity of the overall market share.[6] This process is pictured in Figure 1.

The difference between these bias-free elasticities of overall market share and a naive projection of the elasticity at the mean values of the exogenous variables can be quite significant. Table 2 shows comparable values for a sample of seven elasticities on the EPRI dataset.

---

[4] In effect, the more likely a household is to choose a particular alternative, the more important that household's elasticity is in determining the overall market elasticity.

[5] Specifically, due to the fact that individual probabilities are expressed as ratios of exponentials.

[6] Readers interested in a more detailed discussion of the "simulation" approach to calculating elasticities should see Wood, Ruderman, and McMahon [22]. In that paper, however, this methodology was applied to an extensively modified version of the EPRI study. Specific disaggregate elasticities reported there will differ from those reported here.

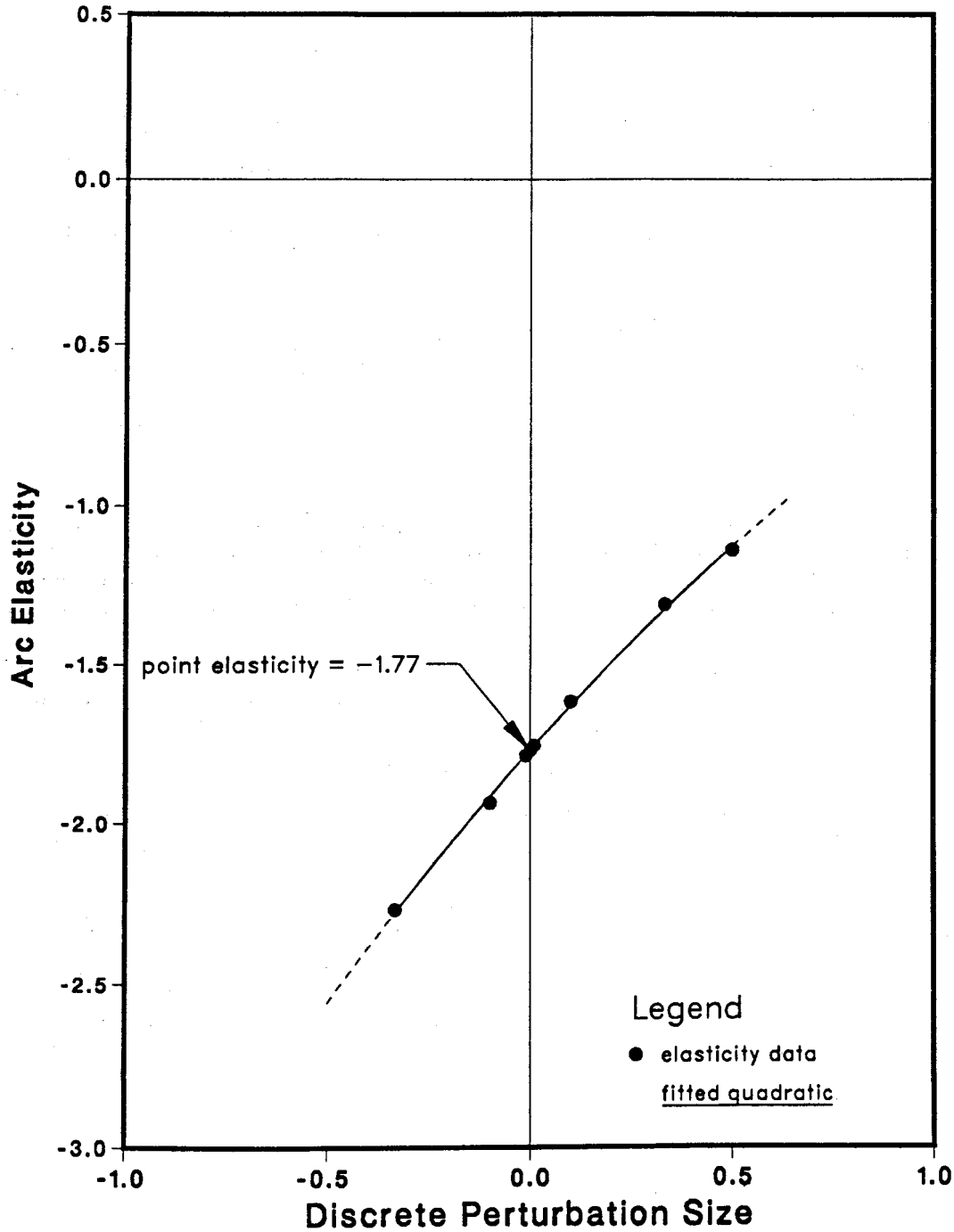

Figure 1. Elasticity of Heat Pumps With Respect To Own Capital Cost.

Table 2

Effect of Aggregation on Elasticity Estimation

market share	exogenous variable	estimated market share elasticities: disaggregated	naive	difference	relative difference
central cooling	price of gas	-0.143	-0.162	0.019	-13.3 %
all gas space heat	price of gas	-0.529	-0.536	0.007	-1.3 %
all oil space heat	price of gas	0.424	0.643	-0.219	-51.7 %
all elec space heat	price of gas	0.572	0.674	-0.102	-17.8 %
central cooling	own capital cost	-0.293	-0.149	-0.144	49.1 %
heat pump	own capital cost	-1.770	-1.350	-0.420	23.7 %
central cooling	household income	0.282	0.375	-0.093	-33.0 %

Our simulation method provides bias-free estimates by using sample enumeration to get unbiased market shares and appropriately combining them to get an overall market elasticity. Furthermore, representation of an entire range of arc elasticities in only three parameters allows an investigator to capture changes in market share which are non-linear in the size of the perturbation of an exogenous variable.

A new market share (under a change in some exogenous variable) can be calculated as a function of the old market share, the elasticity, and the relative perturbation size necessary to reach the new value of the exogenous variable:

$$MS_{new} = MS_{old}(1 + \delta\eta(\delta)) \tag{4.3}$$

If only the point elasticity is available, calculating new market shares from equation 4.3 amounts to an assumption that $\eta(\delta) = \eta$, a constant. The result will be slightly biased, with error more significant as δ is larger or the true arc-elasticities less constant.[7]

This error in estimating new market shares is graphically shown in Figure 2. Since the relative change in market share is just the product of the perturbation δ and the elasticity $\eta(\delta)$, the correct relative change in market share for a 33% increase in the exogenous variable is just the area of the smaller rectangle (= 0.14) shown in Figure 2. The naive elasticity produces a noticeably larger estimate (area = 0.21). In this example, the curve of arc elasticities is nearly flat, so the results of using a point elasticity instead of the correct arc elasticity are only slightly in error. This would not be the case for elasticities and market shares estimated from the curve shown in Figure 1, above.

[7] Arc-elasticities are not constant in δ because of the non-linearity of the logistic distribution. This is expressed in the coefficients of δ and δ^2 in our quadratic fit: very roughly, the larger the absolute values of these coefficients, the more error in predicting new market shares using only the point elasticity.

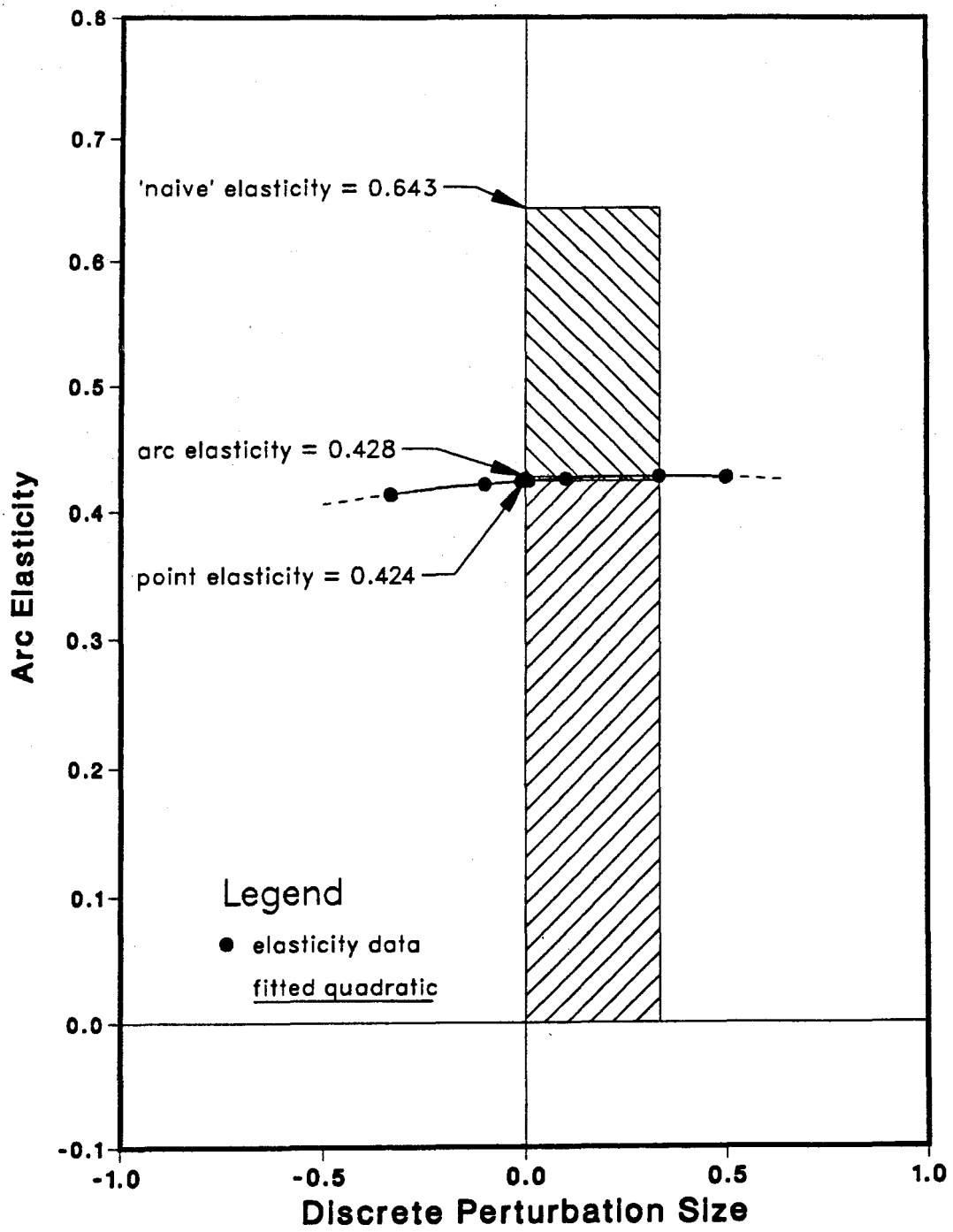

Figure 2. Relative Changes in Market Share: Oil Systems With Respect To Price of Gas.

The two error-inducing effects (naive estimation and the use of point elasticities alone) may be either compounding or partly canceling.[8] The result depends, at least in part, on the direction of the proposed movement of the exogenous variable. Results for particular examples of combined effects are shown in Table 3. There, we report new market shares for seven different technology groups implied by the elasticities in Table 2. All are due to a 33% increase in the exogenous variable shown. Existing (i.e., unperturbed) market shares by disaggregated methods are listed in the first column.

Table 3

Predicted Market Shares for Selected Technologies

			with 33% increase in exogenous variable			
market share	exogenous variable	unperturbed	disaggregated mkt share	"naive" mkt share	difference	relative difference
central cooling	price of gas	0.5264	0.5025	0.5590	-0.0565	-11.2 %
all gas space heat	price of gas	0.5121	0.4249	0.4372	-0.0323	-7.6 %
all oil space heat	price of gas	0.0927	0.1059	0.0966	0.0093	8.8 %
all conv elec heat	price of gas	0.2802	0.3312	0.2887	0.0425	12.8 %
central cooling	own capital cost	0.5264	0.4754	0.5615	-0.0861	-18.1 %
heat pump	own capital cost	0.1150	0.0642	0.0704	-0.0062	-9.7 %
central cooling	household income	0.5264	0.5724	0.6646	-0.0922	-16.1 %

[8] For arc elasticities with significant slope, it may well turn out that the naive elasticity is exactly equal to the arc elasticity for some specific perturbation size. This very nearly occurs for our example in Figure 1.

4. Examination of Aggregation Bias in a Logit Model

This section considers a different model than the probit analyzed by McFadden and Reid. Instead, we examine aggregation bias as it would occur in estimating the parameters of a nested logit model.

Our analysis is made possible by the availability of a disaggregated dataset developed for the Electric Power Research Institute [6]. That dataset was developed to estimate submodels in EPRI's Residential End-Use Energy Planning System (REEPS). The EPRI study used a database of some 1300 households drawn from the Annual Housing Surveys conducted by the Bureau of the Census. Only new, single-family, owner-occupied housing was included. The AHS data were supplemented by estimates of household energy requirements calculated using the MIT Thermal Load Model [15].

The EPRI database allows us to see how much difference aggregated data makes in estimating the coefficients of an econometric model. To this end, we defined two aggregate datasets: each of the independent and dependent variables in EPRI's model was averaged by Standard Metropolitan Statistical Area (SMSA) and by geographical region (one or more contiguous states).[9]

EPRI used a nested logit structure in which the choice of space heating technology is dependent on the choice to have central cooling or not. Specifically, at the root of the nested logit tree, the consumer chooses a central cooling alternative. Given that decision, he selects a space heating alternative. There are five possible choices given central air conditioning (including heat pumps), and eight choices given no central air conditioning.

The logit formulation specifies that consumer utility is a linear combination of the exogenous variables (consisting of appliance attributes such as operating cost, consumer attributes such as income, and geographical characteristics such as weather). The probability of selecting any alternative at a given level of the tree is the ratio of the exponential of the utility of that alternative to the sum of exponentials of utilities of all the alternatives on that level.[10]

The EPRI study used normalized values of the capital and operating costs. Normalization was by the operating cost of an electric baseboard system in the case of space heating and by the annual air conditioning energy usage in the case of central cooling. (The improved statistical properties of the model with this normalization suggest that it is the relative costs of different heating systems which are relevant for the consumer's choice.) In this paper, we assume that normalization would be carried out using aggregated values. Thus, the exogenous cost variables for a given space heating choice are its own average costs in that aggregation group, divided by the average cost of an electric baseboard system. This normalization reflects what would be possible using only the aggregate data. We consider it highly unlikely that an investigator would be able to obtain an aggregate version of the normalized variable.

We calculated mean values for each of the independent variables in each of the three regressions of EPRI's model, averaging over both SMSAs and regions as shown in Appendix 1. In each case we then carried out the same logit regression as in EPRI's study. The regressions were carried out by maximum likelihood techniques similar to those used for the household data. Each entry in the sum of log-likelihoods to be maximized was weighted by the number of household observations over which the mean data had been averaged.

The best choice of a weighting scheme to use in aggregate choice estimation is not a simple question, and depends on the goal of the regression. For a goal of estimating aggregate coefficients consistent with the disaggregate ones, Section 2 suggests the best choice of weight on observation k would be $(1+\mathbf{B}\mathbf{A}_k\mathbf{B})^{-0.5}$ if the values of $\mathbf{B}$ were known (at least for probit models). The choice made here (i.e., to weight each entry in the sum of aggregate log likelihoods by the number of

[9] The aggregation groups used are detailed in Appendix 1.

[10] See EPRI [6] or McFadden [14] for a complete description of the nested logit model.

households in the aggregate) was made for both *ad hoc* and theoretical reasons:

i) From a purely *ad hoc* viewpoint, this weighting scheme outperformed the two other leading candidates: equal weights on all data and weighting by the square root of the number of households. The improved performance was judged both in the overall log-likelihood of each estimation, and in the number of coefficients whose sign or value was inconsistent with utility maximization.

ii) From a more theoretical viewpoint, the aggregate regression would produce exactly the same results as the disaggregate only if a) all aggregation groups were completely homogeneous, and b) each aggregate data point was weighted by the number of households it represented. Thus, the weighting scheme we use means that any differences between aggregate and disaggregate coefficients are due to a pure "heterogeneity" effect.

Tables 4 through 6 show the results of each aggregated regression in comparison with the equivalent results from EPRI's household database.[11] The statistical significance of the differences found, and any special comments on the results, are noted after each table. One general comment should be observed in advance: We use a standard likelihood ratio test to test for the statistical significance of the differences between two sets of estimated parameters. This test is simple both conceptually and in practice, but it effectively dismisses the uncertainty in the aggregate coefficients, treating them as constants and comparing them to the disaggregate estimates (whose uncertainty is used for the test of statistical significance). If the covariance matrix between these two sets of estimates is sufficiently similar to the covariance matrix of the disaggregate estimates, then the resulting statistic will underestimate the true significance of their difference.

Table 4

Effect of Aggregation on Choice of Central Cooling

logit estimates (t-statistics in parentheses)

variable name	Disaggregated (household level data)		Aggregated by SMSA		Aggregated by Region	
normalized capital cost	-5.152	(-5.076)	-6.793	(-1.878)	-9.221	(-0.819)
normalized operating cost	-126.1	(-3.464)	-144.9	(-1.086)	-309.4	(-0.586)
weather	0.1396	(10.250)	0.1210	(2.715)	0.09641	(0.698)
central a/c choice × income	0.07177	(7.605)	0.1531	(1.596)	0.2599	(0.391)
inclusive value	0.3013	(2.704)	0.5522	(1.118)	0.3825	(0.229)
central a/c choice	-1.250	(-2.016)	-2.170	(-0.708)	-2.623	(-0.140)
Likelihood Ratio Test						
log likelihood household model:	-592.6					
log likelihood household model restricted to aggregated results:			-634.3		-840.4	
2 × difference ($\approx \chi^2_6$)			83.4	$p < .01$	495.6	$p < .01$

The likelihood ratio test is used to test the significance of the difference between two sets of parameter estimates. The test compares the log likelihood of the unrestricted household regression

[11] There were some minor errors in the data used by EPRI to estimate the coefficients of their model (Wood, Ruderman, & McMahon [21]). These errors caused some of the coefficients to be reported incorrectly in their paper. Corrected versions of the estimates are used in this study and are reported in Tables 4 through 6.

(-592.6) with the log likelihood of the household database restricted to the parameter estimates from the aggregated regressions (-634.3 and -840.4). Twice the difference between them is asymptotically distributed as a χ^2 random variable with degrees of freedom equal to the number of parameters being restricted (6 in this case). The values of the likelihood ratio test statistic (83.4 and 495.6) are highly significant for six degrees of freedom.

Table 5

Effect of Aggregation on Choice of Space Heating
(given no central cooling)

logit estimates (t-statistics in parentheses)

variable name	Disaggregated (household level data)		Aggregated by SMSA		Aggregated by Region	
normalized capital cost	-0.4843	(-4.683)	-0.5326	(-1.925)	-0.5007	(-0.756)
normalized operating cost	-2.259	(-5.773)	-2.163	(-2.439)	-2.972	(-1.406)
gas restrictions type 1	-2.715	(5.781)	-5.265	(-2.765)	-6.871	(-1.374)
gas restrictions type 2	-1.400	(-3.477)	-1.890	(-1.718)	-3.600	(-0.732)
gas restrictions type 3	-1.033	(-3.539)	-1.312	(-1.449)	1.327*	(0.236)
gas hydronic choice	-1.826	(3.779)	-1.656	(-1.378)	-1.754	(-0.629)
gas non-central choice	-3.562	(-8.550)	-3.521	(-3.652)	-3.555	(-1.692)
oil forced air choice	-1.001	(-4.456)	-1.279	(-2.318)	-1.190	(-0.955)
oil hydronic choice	0.09985	(0.252)	-0.06773	(-0.066)	-0.1177	(-0.050)
oil non-central choice	-3.976	(-6.672)	-4.260	(-3.073)	-4.176	(-1.377)
elec forced air choice	-1.187	(-3.746)	-1.512	(-2.020)	-0.8597	(-0.477)
elec baseboard choice	-1.183	(-3.946)	-1.456	(-2.053)	-0.8309	(-0.488)
Likelihood Ratio Test						
log likelihood household model:	-560.0					
log likelihood household model restricted to aggregated results:			-570.5		-623.4	
2 × difference ($\approx \chi^2_{12}$)			21.0	$p < .10$	126.8	$p < .01$

* This parameter estimate has the wrong (i.e., counterintuitive) sign; however, the variance of the estimate is so large that it is not significantly different from zero.

The values of the likelihood ratio statistic found here (21.0 and 126.8) are significant at the 10% level (for the aggregation by SMSA) and at better than the 1% level (for aggregation by state).

Table 6

Effect of Aggregation on Choice of Space Heating
(given central cooling)

logit estimates (t-statistics in parentheses)

variable name	Disaggregated (household level data)		Aggregated by SMSA		Aggregated by Region	
normalized capital cost	-0.1657	(-6.640)	-0.2507	(-2.585)	-0.2724	(-0.447)
normalized operating cost	-2.473	(-4.757)	-0.7140	(-0.225)	4.809*	(0.342)
operating cost × income	-0.04503	(-3.009)	-0.1753	(-1.427)	-0.3644	(-0.671)
gas restrictions type 1	-2.428	(-7.598)	-3.658	(-2.547)	-5.522	(-1.026)
gas restrictions type 2	-1.001	(-2.209)	-1.182	(-0.717)	2.506*	(0.302)
gas restrictions type 3	-1.793	(-5.999)	-2.837	(-2.087)	-2.319	(-0.396)
heat pump trend	0.04506	(3.855)	0.08418	(1.461)	0.07206	(0.430)
oil forced air choice	-1.827	(-8.750)	-1.713	(-2.590)	-1.950	(-0.986)
elec forced air choice	0.9907	(3.811)	1.496	(1.719)	0.7834	(0.230)
heat pump choice	-0.3947	(-1.700)	-0.2780	(-0.335)	-0.2200	(-0.072)
elec baseboard choice	-1.097	(-4.170)	-0.7775	(-0.924)	-1.237	(-0.401)
Likelihood Ratio Test						
log likelihood household model:	-938.0					
log likelihood household model restricted to aggregated results:			-990.8		-1188.0	
2 × difference ($\approx \chi^2_{11}$)			105.6	$p < .01$	500.0	$p < .01$

* Both of these two parameter estimates have the wrong (i.e., counterintuitive) sign; however, the variance of the estimates is so large that they are not significantly different from zero.

The values of the likelihood ratio test statistic found here (105.6 and 500.0) are highly significant for eleven degrees of freedom.

5. Effects of Aggregation Bias on Elasticities and Projected Market Shares

In this section, we isolate the effects of aggregate data used in estimating coefficients on elasticities resulting from the model. In particular, we would like to exclude further bias effects due to naive estimation of market shares and elasticities.

To this end, we found elasticities using the coefficients estimated on aggregate data, described in the previous section. We used both the naive approach and the sample enumeration approach described in Section 3. Sample enumeration was over the 122 SMSAs (or 20 regions) in the database, reflecting the best a researcher could do using only aggregate data. Examples of this process are shown in Figures 3 and 4, below. The resulting elasticities are tabulated in Tables 7 and 8, along with "aggregate naive" and disaggregated elasticities .

Some consequences of this aggregation bias are shown in Tables 9 and 10, where the predicted market shares for the seven different heating/cooling choices are shown under 33% increases in the exogenous variable listed. The true (i.e., calculated by enumeration using the disaggregated coefficients), unperturbed market shares are shown in the center column.[12]

As can be seen from both the figures and tabulated results, elasticities estimated from sample enumeration on SMSA-aggregate data are generally close enough to the disaggregate elasticities to be acceptable. Naive estimation produces unacceptable errors with both aggregate and disaggregate data.

[12] Market shares for both enumeration methods were calculated from arc elasticities. All market shares in Tables 9 and 10 were calculated as changes from an unperturbed "base" figure, using the elasticities in Tables 7 and 8. However, that base was different for the different methods. Each was found by procedures equivalent to those used for the elasticities: a) enumeration on household data using coefficients from the disaggregate regressions, b) naive market share estimation, and c) enumeration on the aggregate data using coefficients from the aggregate regressions.

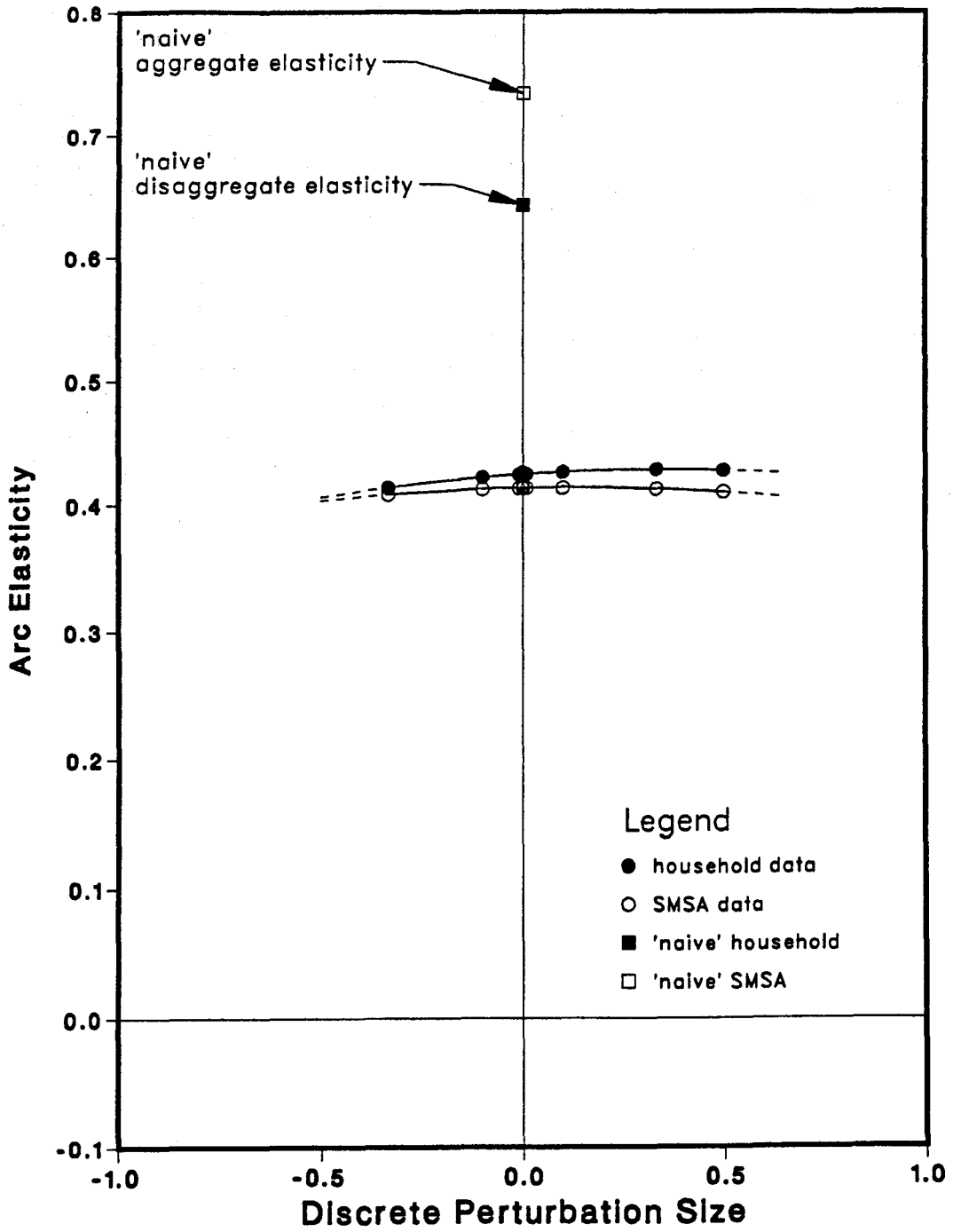

Figure 3. Disaggregate and SMSA-Aggregate Elasticities: Oil Systems With Respect To Price of Gas.

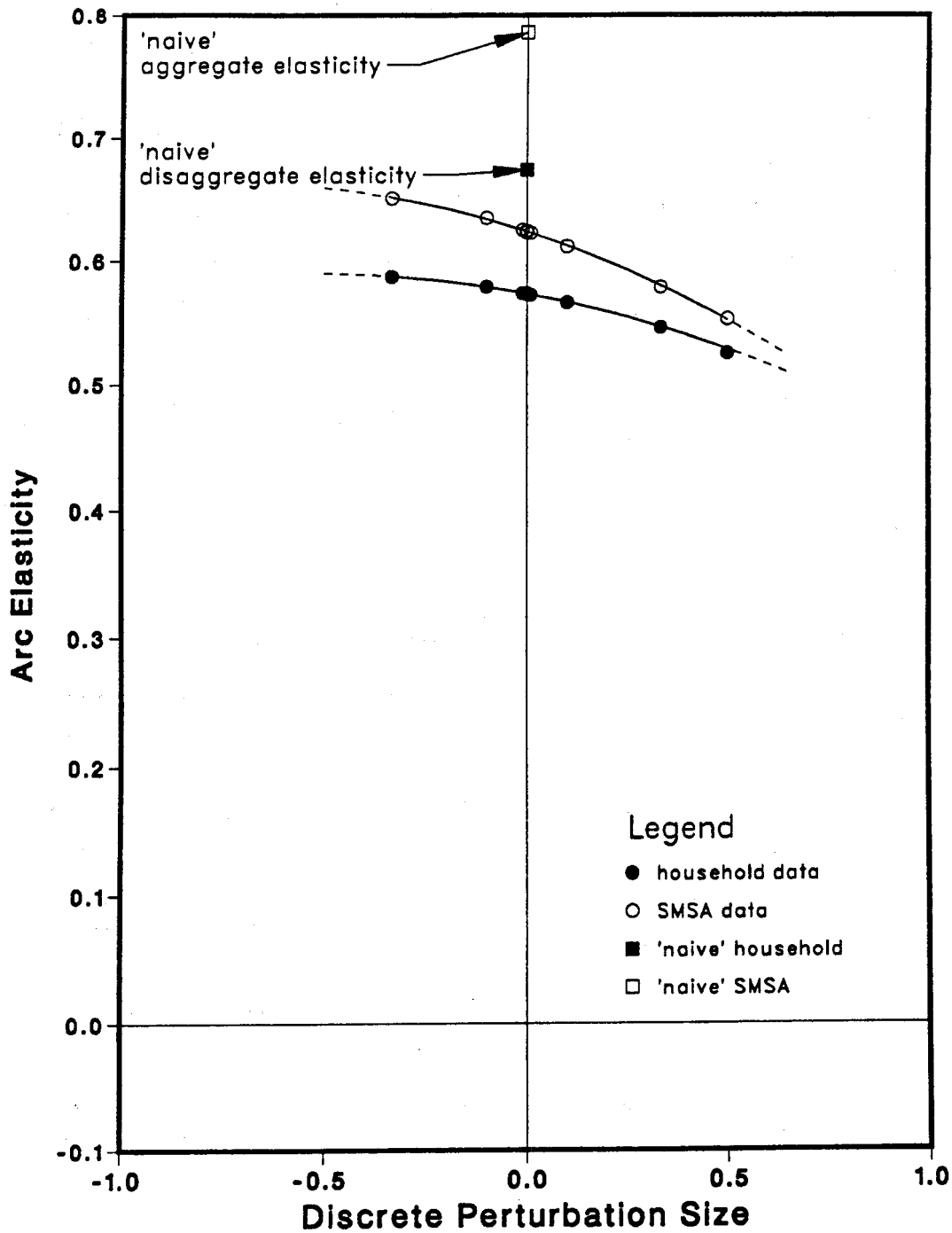

Figure 4. Disaggregate and SMSA-Aggregate Elasticities: Electric Systems With Respect To Price of Gas.

Table 7

Effect of Aggregation on Elasticity Estimation

		estimated elasticities from logit models:		
			aggregated by SMSA	
market share	exogenous variable	disaggregated	naive	enumerated on SMSAs
central cooling	price of gas	-0.143	-0.249	-0.195
all gas space heat	price of gas	-0.529	-0.674	-0.590
all oil space heat	price of gas	0.424	0.734	0.413
all conv elec heat	price of gas	0.572	0.786	0.624
central cooling	own capital cost	-0.293	-0.245	-0.413
heat pump	own capital cost	-1.770	-2.034	-2.752
central cooling	household income	0.282	0.516	0.390

Table 8

Effect of Aggregation on Elasticity Estimation

		estimated elasticities from logit models:		
			aggregated by regions	
market share	exogenous variable	disaggregated	naive	enumerated on regions
central cooling	price of gas	-0.143	-0.087	-0.091
all gas space heat	price of gas	-0.529	-0.562	-0.512
all oil space heat	price of gas	0.424	0.623	0.394
all conv elec heat	price of gas	0.572	0.565	0.517
central cooling	own capital cost	-0.293	-0.295	-0.503
heat pump	own capital cost	-1.770	-2.207	-2.384
central cooling	household income	0.282	0.883	0.770

Table 9

Predicted Market Shares for Selected Technologies					
			with 33% increase in exogenous variable		
				aggregated by SMSAs	
market share	exogenous variable	unperturbed	disaggregated mkt share	naive mkt share	enumerated on SMSAs
central cooling	price of gas	0.5264	0.5025	0.5400	0.4989
all gas space heat	price of gas	0.5121	0.4249	0.4177	0.4132
all oil space heat	price of gas	0.0927	0.1059	0.0950	0.1073
all conv elec heat	price of gas	0.2802	0.3312	0.3129	0.3464
central cooling	own capital cost	0.5264	0.4754	0.5408	0.4576
heat pump	own capital cost	0.1150	0.0642	0.0442	0.0461
central cooling	household income	0.5264	0.5724	0.6902	0.5979

Table 10

Predicted Market Shares for Selected Technologies					
			with 33% increase in exogenous variable		
				aggregated by regions	
market share	exogenous variable	unperturbed	disaggregated mkt share	naive mkt share	enumerated on regions
central cooling	price of gas	0.5264	0.5025	0.5675	0.5044
all gas space heat	price of gas	0.5121	0.4249	0.4071	0.4078
all oil space heat	price of gas	0.0927	0.1059	0.0856	0.1058
all conv elec heat	price of gas	0.2802	0.3312	0.3487	0.3460
central cooling	own capital cost	0.5264	0.4754	0.5270	0.4366
heat pump	own capital cost	0.1150	0.0642	0.0356	0.0530
central cooling	household income	0.5264	0.5724	0.7563	0.6494

6. Summary

This paper has tried to explain the ways in which aggregation bias can affect econometric investigations, and to provide a clear example of that bias. By reworking a correctly disaggregated EPRI study, using only aggregate data at particular points, we found:

i) potentially severe aggregation bias in naive estimation of market shares (Table 1),

ii) large relative errors in naive estimation of market share elasticities (Table 2), with potential errors in prediction resulting from the use of naive elasticities (Table 3),

iii) statistically significant errors in estimating the coefficients of logit models on aggregate data (Tables 4, 5, and 6),

iv) coefficients of aggregate models leading to large errors in estimated elasticities (Tables 7 and 8), with consequent errors in predicting market shares (Tables 9 and 10).

We also found that the errors resulting from models estimated on aggregate data can be reduced by aggregating over smaller geographical units (SMSAs rather than regions) and by calculating market shares and elasticities using enumeration over the aggregate groups.

7. Acknowledgements

The authors are indebted to Dr. Andrew Goett of Cambridge Systematics, the author of the original EPRI study on which this work is based. We are also grateful to Dan McFadden and Fred Reid, whose 1975 exposition of aggregation bias formed the foundation for our own more modest understanding; and to Mark Levine of the Lawrence Berkeley Laboratory, whose interest and support have helped bring this study to fruition.

References

1. Amemiya, T. [1981]: "Qualitative Response Models: A Review," *Journal of Economic Literature*, **19**, December, pp. 1483-1536.

2. Anderson, K. [1973]: "Residential Energy Use: An Econometric Analysis,

3. Baughman, M. and Joskow, P. [1975]: "The Effects of Fuel Prices on Residential Appliance Choice in the United States," *Land Economics*, February.

4. Dohrmann, D. [1980]: "Estimation of Fuel Choice Elasticities for the Residential Sector," Report prepared for Lawrence Berkeley Laboratories.

5. Dubin, J. and McFadden, D. [1984]: "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," *Econometrica*, **52**, No. 2, March, pp. 345-362.

6. Electric Power Research Institute (EPRI) [1984]: "Household Appliance Choice: Revision of REEPS Behavioral Models," EPRI EA-3409, Research Project 1918-1, Final Report.

7. Hartman, R. [1978]: "A Critical Review of Single Fuel and Interfuel Substitution Residential Energy Demand Models," MIT Energy Laboratory Report MIT-EL-78-003, March.

8. Hartman, R. [1982]: "A Note on the Use of Aggregate Data in Individual Choice Models: Discrete Consumer Choice Among Alternative Fuels for Residential Appliances," *Journal of Econometrics*, **18**, pp. 313.

9. Hartman, R. and Hollyer, M. [1977]: "An Examination of the Use of Probability Modeling for the Analysis of Interfuel Substitution in Residential Energy Demand", MIT Energy Laboratory Working Paper MIT-EL-77-018WP, July.

10. Koppelman, F. [1975]: "Travel Predictions with Models of Individual Choice Behavior," Center for Transportation Studies, Massachusetts Institute of Technology, Cambridge, Report 75-7.

11. Landau, U. [1981]: "Reducing the Aggregation Bias: A Heuristic Approach," *Transportation Research -- A*, Vol. **15A**, No. 5, pp. 367-375.

12. Lin, W., Hirst, E. and Cohn, S. [1976]: "Fuel Choices in the Household Sector," Oak Ridge National Laboratory, ORNL/CON-3, October.

13. McCarthy, P. [1982]: "Further Evidence on the Temporal Stability of Disaggregate Travel Demand Models," *Transportation Research -- B*, Vol. **16B**, No. 4, pp. 263-278.

14. McFadden, D. [1978]: "Modeling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models*, edited by Karlqvist, A. *et al.*, Amsterdam: North-Holland.

15. McFadden, D. and Dubin, J. [1982]: "A Thermal Model for Single Family Owner-Occupied Detached Dwellings in the National Interim Energy Consumption Survey," MIT Energy Laboratory Discussion Paper No. 25, MIT-EL-82-040WP, June.

16. McFadden, D., and Reid, F. [1975]: "Aggregate Travel Demand Forecasting From Disaggregated Behavioral Models," National Academy of Science, National Research Council, Transportation Research Board, *Record,* No. 534, Washington, D.C.

17. Reid, F. [1974]: "A Set of Models for Optimizing the Benefits of a Transportation Plan," M.S. Dissertation, University of California, Berkeley.

18. Reid, F. [1978]: "Aggregation Methods and Tests," *Final Report Series, Volume VII,* Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley.

19. Reid, F. [1979]: "Minimizing Error in Aggregate Predictions From Disaggregated Models," National Academy of Science, National Research Council, Transportation Research Board, *Record,* No. 534, Washington, D.C.

20. Talvitie, A. [1973]: "Aggregate Travel Demand Analysis with Disaggregate or Aggregate Travel Demand Models," *Proceedings of the Transportation Research Forum,* **14**(1), pp. 583-603.

21. Wood, D., Ruderman, H. and McMahon, J. [1989]: "A Review of Assumptions and Analysis in EPRI Report EA-3409: 'Household Appliance Choice: Revision of REEPS Behavioral Models,'" LBL Technical Report LBL-20332.

22. Wood, D., Ruderman, H. and McMahon, J. [1989]: "Market Share Elasticities for Fuel and Technology Choice in Home Heating and Cooling," LBL Technical Report LBL-20090.

Appendix 1 Aggregation by SMSAs and by Regions

The database used by EPRI consisted of more than 1300 households in 122 SMSAs and 41 states. We aggregated each of the independent variables in EPRI's model by averaging the values for:

1) SMSAs: Each SMSA (Standard Metropolitan Statistical Area) was allowed to be represented by as many or as few households as was available in the database. In some cases, this meant that the average value of the explanatory variables for an SMSA might be based on only one or two households. This is a potential source of error in the aggregated results.

2) Regions: We felt it was unrealistic to allow entire states to be represented by only one or two household observations, so we aggregated contiguous states together to make larger regions with a minimum number of households in each region for each of the three different regressions in the model. This left us with twenty regions, each geographically contiguous and more or less homogeneous in climate.

The state/region grouping analysis showed that the nine states not represented in the database were almost all from the northern tier of the country (Montana, North Dakota, Vermont, etc.) Although not large in population, the absence of these cold-weather states may be skewing the results found in EPRI's study.

The tables on the following pages show the state/regional groupings used, the SMSAs within each state, and the number of household observations available for aggregation in each of the three regressions of the model. There were twenty regions (representing different numbers of households) in each of the three regressions. But there were as few as 88 SMSAs represented in the "heating choice, no central cooling" regression, and as many as 121 SMSAs in the "central cooling choice" regression.

Appendix 1
SMSAs, Regional Groupings, and Number of Households

Regional Grouping	heat choice no ac	heat choice ac	ac choice
ALABAMA			
Birmingham, AL	2	14	16
Mobile, AL	1	3	4
FLORIDA			
Fort Lauderdale-Hollywood, FL	1	5	6
Jacksonville, FL	1	7	8
Miami, FL	0	9	9
Orlando, FL	1	13	14
Tampa-St. Petersburg, FL	0	19	19
West Palm Beach-Boca Raton, FL	2	11	13
GEORGIA			
Atlanta, GA	0	37	37
Augusta, GA-SC	0	4	4
LOUISIANA			
New Orleans, LA	0	14	14
Shreveport, LA	0	3	3
MISSISSIPPI			
Jackson, MS	0	2	2
Regional Total	8	141	149
ARIZONA			
Phoenix, AZ	0	33	33
Tucson, AZ	13	3	16
Regional Total	13	36	49
ARKANSAS			
Little Rock-North Little Rock, AR	0	6	6
KANSAS			
Wichita, KS	0	5	5
IOWA			
Davenport-Rock Island-Moline, IA-IL	1	2	3
Des Moines, IA	0	7	7
MISSOURI			
Kansas City, MO-KS	6	18	24
St. Louis, MO-IL	1	18	19
NEBRASKA			
Omaha, NE-IA	0	6	6
OKLAHOMA			
Oklahoma City, OK	0	12	12
Tulsa, OK	3	12	15
Regional Total	11	86	97

Appendix 1, continued
SMSAs, Regional Groupings, and Number of Households

Regional Grouping	heat choice no ac	heat choice ac	ac choice
CALIFORNIA			
Anaheim-Santa Ana, CA	15	14	29
Bakersfield, CA	0	7	7
Fresno, CA	1	6	7
Los Angeles-Long Beach, CA	5	26	31
Oxnard-Simi Valley-Ventura, CA	5	7	12
Sacramento, CA	0	27	28
Salinas-Seaside-Monterey, CA	5	0	5
San Bernardino-Riverside-Ontario, CA	1	30	31
San Diego, CA	18	7	26
San Francisco-Oakland, CA	15	4	19
San Jose, CA	16	2	19
Santa Barbara-Santa Maria-Lompoc, CA	4	0	4
Stockton, CA	0	12	12
Regional Total	85	142	230
COLORADO			
Denver-Boulder, CO	33	8	41
CONNECTICUT			
Bridgeport-Milford, CT	6	1	7
Hartford, CT	2	0	2
New Haven-Meriden, CT	1	0	1
MASSACHUSETTS			
Boston, MA	7	3	11
Springfield, MA-CT	2	0	2
Worcester, MA	2	0	2
RHODE ISLAND			
Providence, RI	11	0	11
Regional Total	31	4	36
DELAWARE			
Wilmington, DE-NJ-MD	2	1	3
DISTRICT OF COLUMBIA			
Washington, DC-MD-VA	1	29	30
MARYLAND			
Baltimore, MD	4	12	16
Regional Total	7	42	49

Appendix 1, continued
SMSAs, Regional Groupings, and Number of Households

Regional Grouping	heat choice no ac	heat choice ac	ac choice
ILLINOIS			
Chicago, IL	6	32	38
Peoria, IL	3	6	9
Rockford, IL	2	2	4
Regional Total	11	40	51
INDIANA			
Fort Wayne, IN	2	3	5
Gary-Hammond-East Chicago, IN	0	9	9
Indianapolis, IN	5	10	15
South Bend, IN	1	0	1
Regional Total	8	22	30
KENTUCKY			
Louisville, KY-IN	0	9	9
TENNESSEE			
Chattanooga, TN-GA	0	1	1
Knoxville, TN	1	6	7
Memphis, TN-AR-MS	0	16	16
Nashville-Davidson, TN	0	3	3
NORTH CAROLINA			
Charlotte-Gastonia-Rock Hill, NC-SC	1	6	7
Greensboro-Winston-Salem-etc., NC	4	12	16
SOUTH CAROLINA			
Charleston, SC	1	4	6
Columbia, SC	1	6	7
Greenville-Spartanburg, SC	1	2	3
VIRGINIA			
Newport News-Hampton, VA	0	3	3
Norfork-Virginia Beach, VA-NC	0	6	6
Richmond, VA	1	12	13
WEST VIRGINIA			
Huntington-Ashland, WV-KY-OH	0	1	1
Regional Total	10	87	98
MICHIGAN			
Detroit, MI	23	12	35
Flint, MI	2	0	2
Grand Rapids, MI	5	0	5
Lansing-East Lansing, MI	2	0	2
Regional Total	32	12	44

Appendix 1, continued
SMSAs, Regional Groupings, and Number of Households

Regional Grouping	heat choice no ac	heat choice ac	ac choice
MINNESOTA			
Duluth-Superior, MN-WI	1	0	1
Minneapolis-St. Paul, MN-WI	9	15	24
Regional Total	10	15	25
NEVADA			
Las Vegas, NV	0	12	12
UTAH			
Salt Lake City-Ogden, UT	12	4	16
Regional Total	12	16	28
NEW JERSEY			
Newark, NJ	6	0	7
Paterson-Clifton-Passaic, NJ	2	1	4
Trenton, NJ	0	1	1
NEW YORK			
Albany-Schenectady-Troy, NY	6	0	6
Binghampton, NY-PA	1	0	1
Buffalo, NY	6	0	6
New York, NY	2	2	4
Rochester, NY	5	0	5
Syracuse, NY	7	0	7
Regional Total	35	4	41
NEW MEXICO			
Albuquerque, NM	6	10	16
OHIO			
Akron, OH	0	2	2
Canton, OH	1	1	2
Cincinnati, OH-KY-IN	3	16	18
Cleveland, OH	7	6	13
Columbus, OH	3	7	10
Dayton-Springfield, OH	3	8	11
Lorain-Elyria, OH	3	0	3
Toledo, OH	5	2	7
Youngstown-Warren, OH	3	1	4
Regional Total	28	43	70

Appendix 1, continued
SMSAs, Regional Groupings, and Number of Households

Regional Grouping	heat choice no ac	heat choice ac	ac choice
OREGON			
Portland, OR-WA	19	2	21
WASHINGTON			
Seattle-Everett, WA	27	2	30
Spokane, WA	3	1	4
Tacoma, WA	11	1	12
Regional Total	60	6	67
PENNSYLVANIA			
Allentown-Bethlehem, PA-NJ	5	0	5
Erie, PA	2	0	2
Harrisburg, PA	4	1	5
Johnstown, PA	1	0	1
Lancaster, PA	1	0	1
Philadelphia, PA	12	14	27
Pittsburgh, PA	6	6	12
Reading, PA	4	0	4
Wilkes-Barre-Hazelton, PA	1	0	1
York, PA	5	2	7
Regional Total	41	23	65
TEXAS			
Austin, TX	0	7	7
Beaumont-Port Arthur-Orange, TX	0	6	6
Corpus Christi, TX	0	3	3
Dallas, TX	0	26	26
El Paso, TX	7	1	8
Fort Worth-Arlington, TX	0	23	23
Houston, TX	0	20	22
San Antonio, TX	1	9	10
Regional Total	8	95	105
WISCONSIN			
Appleton-Oshkosh-Neenah, WI	3	0	3
Madison, WI	2	3	5
Milwaukee, WI	7	5	12
Regional Total	12	8	20

Appendix 2 Aggregation Bias in a Probit Model

The aggregation bias theory developed by McFadden and Reid [16] shows that, given reasonable assumptions, equation 2.3 holds as an asymptotic relationship as the number of observations in each aggregation group goes to infinity.

$$\overline{P}_k \underset{N_k \to \infty}{\longrightarrow} \Phi\left(\frac{\mathbf{B}\overline{\mathbf{Z}}_k}{\sqrt{1+\sigma_k^2}}\right) \tag{2.3}$$

where $\overline{P}_k$ is the market share of alternative 1 in aggregation group k, $\sigma_k^2 = \mathbf{B}\mathbf{A}_k\mathbf{B}$, and $\mathbf{A}_k$ is the covariance matrix of the exogenous variables in aggregation group k.

Since we were in possession of a fully disaggregated dataset, from which we had estimates of the true (i.e., the disaggregated) parameters $\mathbf{B}$, we decided to test to see if a correction suggested by equation 2.3 could improve the estimation of a probit model on aggregated data.

To that end, we initially calculated a disaggregated probit model of the the central cooling choice at the root of EPRI's nested logit. The results of that estimation are shown in Table 11. The probit coefficients are fairly similar to the (appropriately scaled) logit coefficients reported in Table 1.[13] For the purposes of the probit modeling, the inclusive value of the nested logit was treated as an arbitrary independent variable.

Table 11

Disaggregated Probit Model of Central Cooling Choice		
variable name	coefficient estimate	t-statistic
normalized capital cost	-2.6122	(-6.881)
normalized operating cost	-77.4534	(-4.179)
weather	0.0838	(11.595)
income	0.0413	(8.079)
inclusive value	0.1567	(3.757)
central cooling choice	-0.7580	(-2.742)

We then calculated the empirical means $\overline{\mathbf{Z}}_k$ and covariance matrix $\mathbf{A}_k$ of the six independent variables for each of the twenty regional groupings. From these we formed $\sigma_k^2 = \mathbf{B}\mathbf{A}_k\mathbf{B}$, using the $\mathbf{B}$ estimated from the disaggregate probit model, and estimated via weighted least squares two models of the following forms:

$$\Phi^{-1}(\overline{P}_k) = \mathbf{D}\overline{\mathbf{Z}}_k + \epsilon_k \tag{2.4}$$

and

$$\Phi^{-1}(\overline{P}_k) = \frac{\mathbf{C}\overline{\mathbf{Z}}_k}{\sqrt{1+\sigma_k^2}} + \epsilon_k. \tag{APP 2.1}$$

[13] The actual quantity estimated in most qualitative choice models is not the vector of coefficients $\mathbf{B}$, but rather the vector $\mathbf{B}/\sigma$, where σ^2 is the variance of the stochastic term in the random utility model. Since the probit and logit models are driven by the standard normal and logistic distributions, respectively, coefficients estimated on the same data using these two models will differ at least by a scale factor representing the relative size of their standard deviations. Amemiya [1] has shown that dividing logit coefficients by 1.6 allows very close approximation to the probit coefficients on the same data.

Values of Φ^{-1} were found by rational approximations using standard techniques. Weights were inversely proportional to the square root of the number of households in each aggregation group.

The results of this aggregated estimation (the parameters **D** and **C**) are compared with the results of the disaggregated probit estimation (the parameters **B**) in Table 12, below. Differences between the estimates can be attributed to aggregation bias.

Table 12

Effect of Aggregation on Choice of Central Cooling
Probit Model

variable name	Aggregated by State/Region: Aggregated **D**	Aggregated by State/Region: Corrected **C**	Disaggregated **B**
normalized capital cost	-2.477	-2.403	-2.6122
normalized operating cost	386.3	418.3	-77.4534
weather	0.06353	0.06977	0.0838
income	0.1104	0.1131	0.0414
inclusive value	0.6066	0.6679	0.1567
central cooling choice	-4.987	-5.299	-0.7580
Likelihood Ratio Test			
log likelihood household model:			-593.95
log likelihood household model restricted to aggregated results:	-1432.8	-1515.7	
$2 \times$ difference ($\approx \chi^2_6$)	1677.7	1843.5	

Equation 2.3 suggests the estimates **C** should be closer to the true **B** than **D**, and their difference is tested in the table with a logistic ratio test.[14] By that test, at least, the corrected results are even worse than the uncorrected ones (i.e., they have a higher χ^2 statistic).

The failure of the estimates to improve significantly when the attenuation term $(1 + \sigma_k^2)^{0.5}$ is divided into the independent variables is not entirely surprising. Equation 2.3 holds only asymptotically, and requires a joint normal distribution of the exogenous variables $\mathbf{Z}_k$. This last assumption was not even remotely met by the EPRI dataset.

We examined the joint normalcy assumption by testing the marginal normalcy of each independent variable in each aggregation group.[15] The results of that test were more or less what we could have expected: only the independent variables of income and inclusive value were approximately normal in most aggregation groups. The central cooling choice dummy variable is constant across all households, so it contributes nothing to the covariance matrix (and therefore

[14] This is *not* the technique that McFadden and Reid proposed to estimate the parameters **B** from aggregate data. Their approach was to consider the problem of estimating **D** (the vector of estimated parameters from aggregated data) as an implicit function of itself. The approach we take here is only possible because we are in possession of a disaggregated dataset. It would not be possible for an investigation with only aggregate data.

[15] If the variables are distributed as a joint normal, their marginal distributions must be normal. Therefore failure to find normal marginal distributions indicates that the joint normalcy assumption was not met by the data. Normalcy was tested by the Kolmogorov-Smirnov test using SPSS-X, with corrections for the use of the empirical mean and variance of the data itself.

nothing to the aggregation bias). The climate variable is constant for all households in a given SMSA; therefore it tends to be "lumpy" for a state or region with households from several SMSAs. Similarly, the cost variables tend to be close to constant across an SMSA (where common fuel prices and common climate do more to make costs uniform than variation in dwelling size does to make them varied). But over a larger area, the distribution of costs tends to be multi-modal, unless many SMSAs have been included.

This lack of a joint normal distribution of the independent variables probably doomed the correction from the start. But for the reasons discussed above, this same lack is likely to occur in other datasets and other studies of household space heating appliance choice. Unfortunately, this tends to cast doubt on the usefulness of most classification techniques (which require normal, or at least symmetric, distributions for the independent variables) to reduce aggregation bias.

Interestingly enough, we find a potential conflict between two desirable goals in choosing the size of an aggregation group: small groups are likely to have smaller variance across the individuals in the group (e.g., climate is common to an entire city), but larger groups may have a wider and more nearly continuous set of determining conditions and therefore a better chance of insuring approximately normal distributions for the independent variables (e.g., capital costs). Researchers on transportation choice generally found that *no* geographical classification scheme was practical, although grouping on smaller units generally worked better than larger units. But if a classification-type of correction is going to be made, larger groups may come closer to satisfying the assumptions: more symmetric distributions of the independent variables and better estimates of the covariance matrices. Smaller groups will have less overall bias, but larger groups may be more amenable to correction of that bias.

We also looked at one other aspect of equation 2.3: its asymptotic properties. Clearly, as an asymptotic result it holds completely only for an infinite number of observations in each aggregation group. Equally clearly, for some large enough number, it ought to hold reasonably well. How large is large enough?

The answer will certainly vary with the circumstances and the data used. Sufficient observations for one problem may be insufficient for another, and different datasets will require greater or lesser samples to get reasonable estimation of the covariance matrices $\mathbf{A}_k$.

We can get some idea of what is involved by looking at this particular problem and its associated data. If our data had a joint normal distribution, with different means but a common covariance matrix $\mathbf{A}_k = \mathbf{A}$ in each aggregation group, then the conditions of equation 2.3 are met. In that circumstance, the parameters $\mathbf{D}$ estimated from

$$\Phi^{-1}(\overline{P}_k) \;=\; \mathbf{D}\overline{\mathbf{Z}}_k \;+\; \epsilon_k \tag{2.4}$$

would be asymptotically just simple multiples of the true coefficients $\mathbf{B}$:

$$\mathbf{D} \underset{N_k \to \infty}{\longrightarrow} \frac{\mathbf{B}}{\sqrt{1 + \mathbf{BAB}}}. \tag{APP 2.2}$$

How many observations per aggregation group do we need in the EPRI dataset (if it were normally distributed) for the law of large numbers to come to our aid and give us approximately the result in equation APP 2.2? The answer can be hinted at by a simple simulation experiment.

We generated 50,000 independent standard normal variates and transformed them by matrix techniques into five vectors of 10,000 joint-normal variates. Each vector corresponds to one of the first five independent variables in EPRI's cooling choice model. (We eliminated the dummy for the central cooling choice, since as a constant it does not contribute to aggregation bias). The five vectors were organized into 20 groups of 500 observations per vector. Each group was given a joint normal distribution with mean vector equal to the empirical mean of one of the aggregate states/regions in the EPRI database. All groups were given a common covariance matrix $\mathbf{A}$, equal to the overall covariance of the EPRI database. We also formed a dummy choice vector, using a known linear combination of the five independent variables and one additional normal random

variate to "blur" the choices.

We then estimated the disaggregate probit model of equation 2.2, and the two least-squares aggregate models 2.4 and APP 2.1 on datasets of gradually increasing size. Specifically, we estimated those three models for five observations per aggregation group (100 total), 10 observations per group (200 total), 25 per group (500 total), 50 (1,000), 100 (2,000), 250 (5,000), and finally 500 observations per group (10,000 total). The resulting coefficients from each model were used to estimate the market share of the choice variable on the disaggregated dataset. The results of that estimation are displayed in Table 13 and Figure 5, below.

Table 13

Asymptotic Effects in Correcting Aggregation Bias

Estimated Market Shares of Central Air Conditioning

number of observations	Aggregated D	Corrected C	Disaggregated B
5/group (100 total)	0.7273	0.7374	0.7639
10/group (200 total)	0.6766	0.6998	0.7329
25/group (500 total)	0.7181	0.7936	0.7551
50/group (1,000 total)	0.7247	0.8132	0.7616
100/group (2,000 total)	0.7004	0.7613	0.7527
250/group (5,000 total)	0.7011	0.7670	0.7520
500/group (10,000 total)	0.6938	0.7582	0.7444

The true market share is not displayed, but varies from sample to sample as the amount of included data increases. The disaggregated estimate was consistently within 0.001 of the true market share.

Note that the correction moves the aggregated estimate in the right direction in every sample, but in some samples (25/group and 50/group), it overshoots its mark and actually ends up with a slightly worse estimate than the uncorrected one. Not until we reach a level of 100 observations in each aggregation group does the correction consistently outperform the raw aggregate estimate.

We should be careful about the extent of the conclusions to be drawn here. This does not constitute a full simulation study of the problem (the 10,000 observations would have to be replicated many times, using different random seeds).[16] Even if this were done, the generality of the conclusions is still limited by the particular characteristics of the EPRI dataset. Other problems and other datasets might require fewer observations in each group to make satisfactory use of classification bias-reduction schemes. But for at least this one empirical dataset, a minimum of 50 to 100 observations per aggregation group seems to be required for a classification type of bias-reduction scheme to be effective.

[16] A second run produced qualitatively similar results, however.

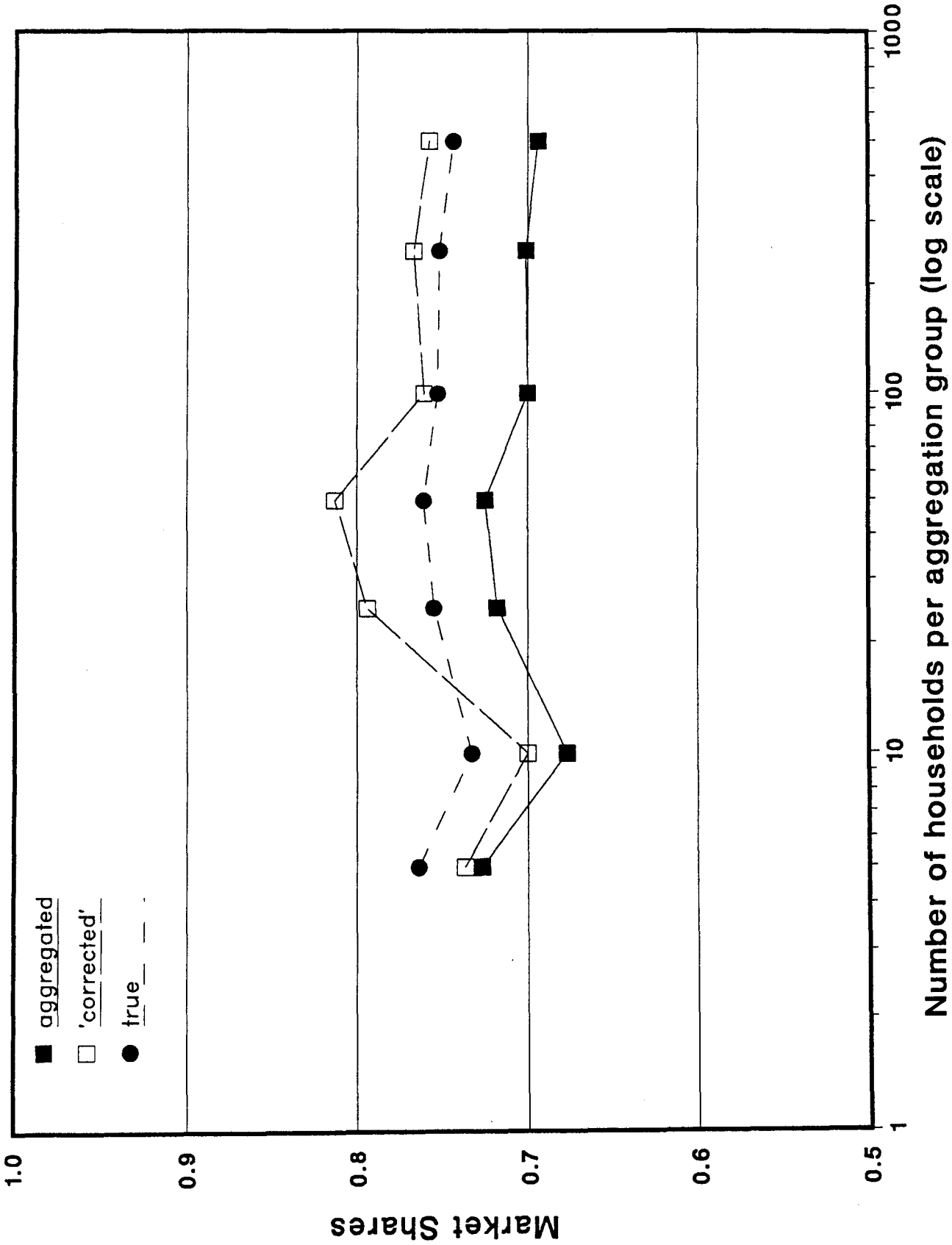

Figure 5. Asymptotic Effects in Correcting Aggregation Bias: Estimated Market Shares of Central A/C.