When Is Model-Based Sampling Appropriate for EIA Surveys?

Nancy J. Kirkendall, Energy Information Administration

The Goal

There are survey situations in which a model-based sample (perhaps a multiattribute cut-off sample) can result in greater accuracy for estimating totals than can be derived from a traditional probability based sample of the same size. This paper describes the efforts of the Energy Information Administration (EIA) to identify and quantify population characteristics which would lead to the conclusion that a model-based sample warrants consideration.

Within EIA the survey systems where cut-off sampling is appropriate have a two-fold sampling strategy and specific population characteristics. (1) The population is monitored by a periodic (e.g. annual) census survey. It is small and highly skew. Data are collected on multiple variables, and estimates are needed for multiple regions. The population and processes measured by the survey are very stable over time. (2) Totals by region are also estimated by more frequent periodic (e.g. monthly) sample surveys. These collect a subset of variables, from a subset of the population. It is possible to link the data provided by a respondent on the two surveys.

The driving reason for EIA's interest in model-based samples is to minimize the sample size, and therefore the cost. Additionally the smaller respondents complain more about the burden of reporting, they require more non-response follow up, and are more likely to report inaccurately than the larger respondents.

The Data

This paper provides examples of some of EIA's efforts in identifying and quantifying population characteristics which would clarify when model-based sampling is appropriate. The examples in this paper make use of data from EIA's electric power surveys.

The EIA-861 is an annual census survey of all electric utilities. It collects a variety of attributes, those of interest to this project include sales and revenues by State and sector: residential, commercial, industrial, public and other. The examples here use 5 years of annual data. The EIA-826 is a monthly sample survey of the same population. It collects sales and revenues by State and sector (residential, commercial, industrial and "other", where "other" on the EIA-826 is the sum of other and public on the EIA-861).

The sample design which was used for the EIA-826 from 1990 through 1992 was a stratified random sample by State. The stratification was (approximately) based on total sales summed across sectors, and for each State includes a certainty stratum and several sample strata (from 2 to 4) with a sample size of 2 per

stratum. The monthly data from the sampled companies are also used in the examples presented later.

Traditional Justification for Cut-off Sample

Traditional justification for the use of a cut-off sample includes the requirement that there is a high correlation between the two data sources (here between the values of the same variable reported by the same utility at different points in time.), and that the coverage of the sample exceed 80% or 90% of the totals being measured.

For the electricity data, the correlation between the same variables reported on the annual surveys taken two years apart ranges from 95% for the "other" category, to 99.9% for the residential category.

Earlier Evaluations

Among the tests we have performed previously are: comparing the accuracy of the results obtained from the stratified sample and estimation procedure, with accuracy of the model-based estimate using the certainty companies only. These estimates for totals were prepared using annual data from one year and annual data from the "sampled" companies two years later. The accuracy of the resulting estimates can be determined from the actual annual total.

The same procedure can be applied using the annual data from the census and the monthly data from the sample to predict monthly totals. In this case, however, we can only compare the two estimates since the true monthly total is not known.

In many States these comparisons have shown that even using only the certainty companies selected for the stratified sample, estimated totals from the model were as accurate or more accurate than those from the stratified design. In other States, it appeared that the stratified certainty stratum needed to be augmented before the results would be satisfactory.

The Search for Other Characteristics

Figure 1 consists of three plots of the logarithm of residential revenues reported by utilities in California. The first plot shows utilities in the certainty stratum, the second shows utilities in the first sample stratum and the third shows utilities in the second sample stratum. These plots illustrate the skew nature of the population: revenues in the certainty stratum range from 1 thousand dollars, ($\log_{10} 1 = 0$), to 2.4 billion dollars ($\log_{10} 2.4*10^6 = 6.4$). The smallest value in the certainty stratum is for a utility which has large revenues in the other sector, but small revenues in the residential sector.

These plots also show that except for very small utilities, the data from the same utility show up as a straight line. This indicates that there is more information concerning the current value for an individual utility from

the past reports of that utility than in the information currently reported by the other utilities. Of course there may be overall trends in the industry from year to year that are not easily seen in this display.

For sake of argument, the coefficient of variation was computed for the utilities in the first sample stratum for three different cases:

1. A simple random sample of size 2. This is part of the overall stratified design, and leads to a coefficient of variation of 100%.

2. For each utility estimate this year's data with last year's data (no sample). This leads to a coefficient of variation of 8%.

3. Use a smarter time series estimate (a first order autoregression) to forecast this year's data from the previous year's data. This leads to a coefficient of variation of 6%.

In other words, ignoring the time series structure in the data is a mistake. The question is whether the current information provided by the sampled companies provides information concerning the period to period change in the nonsampled companies.

Period to Period Change

Model based estimation uses the model

$$y_i = \beta x_i + \epsilon_i \qquad V(\epsilon^i) = \sigma^2 x_i^{2\gamma}$$

where γ=1/2 corresponds to the ratio estimator, y_i represents the data from company i in the current time period (i.e. the data from the current monthly survey) and x_i represents data from company i in the past time period (i.e. the most recent annual data from the census survey).

This model can be rewritten:

$$\frac{y_i}{x_i} = \beta + \eta_i \qquad V(\eta_i) = \sigma^2 x_i^{2(\gamma-1)}$$

and leads to the idea that the ratio of the current data to the past data by company might provide insight into important population characteristics.

The annual data from the census of the utilities from 1986 through 1990 was used to calculate the following ratios:

$$\frac{y_{i,t}^{(k)}}{y_{i,t-1}^{(k)}}$$

here i indicates specific utility, t=2,3,4 or 5 represents year, with t=1 representing 1986, and k represents sample status. If k=1 then the utilities are in the certainty stratum, if k=2 the utilities are in one of the sample strata, and if k=3 the utilities were not sampled. These annual ratios were run through an analysis of variance procedure to see if the estimated mean (β) was dependent on year and/or sample status.

The result was that year is significant in all but 2 States, and sample status was significant in only 5 States. In these 5 States, for all but Minnesota the mean of the nonsampled utilities was closer to the mean of the certainties than it was to the mean of the sampled utilities. This indicates two things: 1) a sample can provide information on year to year change; 2) the stratified sample design does not do this well.

Monthly Data and Seasonality

The above comparison looked only at year to year change. In a monthly survey the estimate for β will reflect seasonality as well as growth or decline in the industry. If the model-based approach is to be used the ratios which should be examined are given by y_i/x_i, where y_i is the current month's data for company i, and x_i is the average annual value for company i in 1990. The estimated values of β will differ from month to month as a result of seasonality. This seasonality may differ from State to State and from sector to sector. The question is whether the factors that influence seasonality for a sector in a State are the same for all companies regardless of size.

This question was addressed by examining plots showing the ratios for various States and various sectors. In these plots the certainty utilities are plotted with a solid line, and the sampled utilities are plotted with dashed lines. Separate plotting symbols are used for each sample stratum in a State. Figure 2 shows three different plots: one for residential revenue in California; one for residential revenue in South Dakota; and one for commercial revenue in South Dakota.

For residential revenue in California, one sampled utility reported only 4 months of revenue during the 24 month period (shown by a ratio of 3). The remaining utilities all show similar seasonal patterns except that 2 of the 3 remaining sampled utilities appear to have somewhat greater seasonal swings.

For residential revenue in South Dakota, the seasonal patterns look fairly consistent except that two of the certainty companies have pronounced peaks in the summer months whereas most of the other utilities only show a slight move upward in the summer months.

For commercial revenue in South Dakota, the third plot, the month to average ratio is relatively flat except for one very small utility, possibly involved in irrigation, which shows a pronounced peak in the summer months.

For these examples, it is likely that estimating a value of β using the certainty companies only would give a reasonable approximation to the seasonality in the total, even though it is clear that the seasonality is not necessarily the same for each utility.

Conclusions

The time plots illustrate the stability of the population and the reported data over time. The analysis of variance showed that current data from a sample may be informative concerning changes in the industry over time. However, it also indicated that the present stratification does not group utilities which are alike.

The seasonal plots support a hypothesis of similar seasonal patterns for most utilities, but they also show that some utilities do not conform with the common pattern. It is not clear that any sampling mechanism would provide more information concerning the seasonal patterns.

The information presented here is an attempt to identify and display population characteristics in a way which would allow the user to evaluate the possibility that a nonprobability sample would provide cost savings without sacrificing accuracy. These results are preliminary. Comments or suggestions are welcome.

Figure 1

Residential Revenue for California

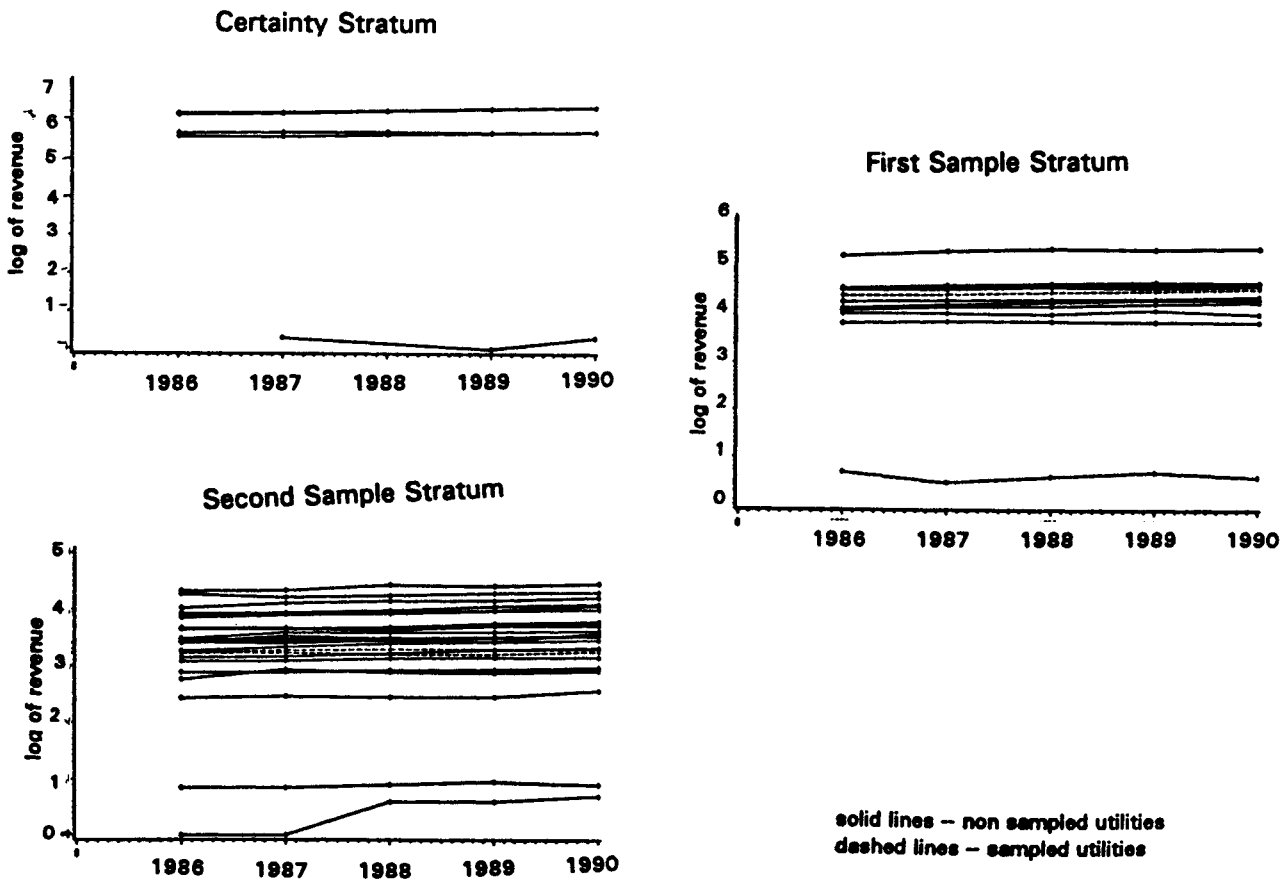
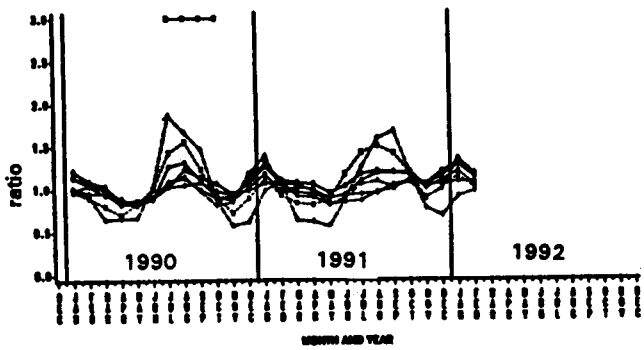


Figure 2

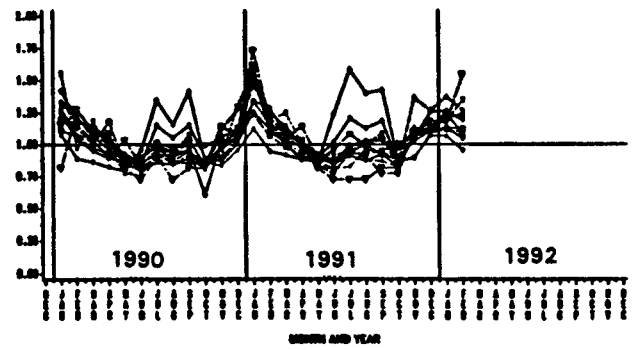
Utility Level Seasonal Patterns by Sector and State

ratio is 12 * month divided by annual total

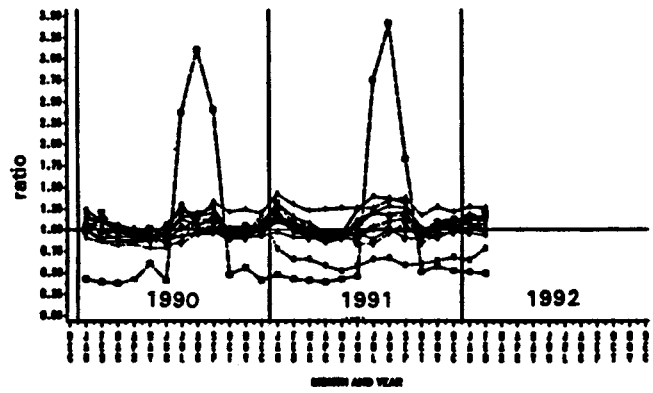
Normalized Residential Revenue for California



Normalized Residential Revenue for Sourth Dakota



Normalized Commercial Revenue for South Dakota



Solid lines certainty utilities

Dashed lines sampled utilities

(different symbols for different strata)