SOME EFFECTS OF FRAME DEFICIENCIES ON ESTIMATES

William E. Winkler
Energy Information Administration
Department of Energy

A frame is a list of companies, along with associated attributes, which can serve as a basis for designing surveys. Data collected with the surveys are then used to obtain estimates of specified (energy) activities of all, or a subset, of these companies. Examples of associated attributes are the type of company (refiner, retailer, importer, etc.), volume of activity (sales to residential consumers, number 2 distillate supplied to market, etc), and reporting status (current respondent, inactive, out-of-scope, etc.). These attributes can be used to choose a sample of companies from the frame. Surveying a sample of companies, in contrast to surveying all the companies, yields a great reduction in overall respondent burden. This allows for an increased amount of information for the same information collection budget.

The sampling techniques now in use or under investigation by the Energy Information Administration (EIA) will require accurate frames. The underlying assumption for improving EIA sampling strategies is that frames will be in place and updated based on periodic censuses of the universe. This approach is critical to EIA objectives of reducing respondent burden. The EIA is working to improve its survey frames.

WHY ARE GOOD FRAMES IMPORTANT?

Four general categories of error which can affect the accuracy of estimates are:

1. Frame deficiencies
2. Nonresponse Error
3. Processing and Respondent Error
4. Sampling Error

Before presenting a number of specific examples of how frame deficiencies can lead to inaccurate estimates, we briefly indicate how frame deficiencies can be connected, at least partially, with the last three general categories of error.

Nonresponse Error- If a large number of out-of-scope companies are included in the survey frame and no telephone numbers are available for them, then nonresponse followup can be very difficult. Improper nonresponse followup or lack of time to do the nonresponse followup can mean that out-of-scope companies are counted as nonrespondents. At the very least, loss of precision in the estimates is the result.

Processing Error- Survey personnel may not be aware of erroneous frame data or many duplicates in the frame. If survey system design does not account for duplicates in the frame, then some nonrespondents may actually be duplicates that have responded under different names, addresses, and respondent identification numbers.

Sampling Error- The measures of size associated with individual companies which are used in sampling programs may be very different than the values that must be estimated. The projected sampling error estimates obtained using frame data will be much different from the sampling error estimates obtained from the submitted data.

The following specific examples indicate why some frame problems can occur so easily and why they are so difficult to resolve.

Example 1. (name and address changes in the frame, possible changes in corporate structure and method of reporting)

An example of the tedious nature of maintaining frames is illustrated by considering the sulfur producers in the Texas Oil Directory for the years 1980 and 1981. The listed sulfur producing establishments are gas processing plants associated with specific Texas gas fields. Only the "A" entries are shown in Figure 1.

FIGURE 1

'A' Sulfur Producers from the 1980 and 1981 Texas Oil Directories

1980

AMERICAN PETROFINA COMPANY OF TEXAS, Box 2159, Dallas, TX 75221
Port Arthur Plant, Jefferson County, Texas
AMINOIL U.S.A., Box 94193, Houston, TX 77018
Birthright Plant, Hopkins County, Texas
AMOCO PRODUCTION COMPANY, Box 591, Tulsa, OK 94102
Cowden, North Plant, Ector County, Texas
Edgewood Plant, Van Zandt County, Texas
Fullerton, South Plant, Andrews County, Texas
Midland Farms Plant, Andrews County, Texas
Slaughter Plant, Hockley County, Texas
Yantis, West Plant, Wood County, Texas
AMOCO TEXAS REFINING COMPANY, Box 401, Texas City, TX 77001
Texas City Refinery Plant, Galveston County, Texas
ATLANTIC RICHFIELD COMPANY, Box 2451, Houston, TX 77001
Fashing Plant, Atascosa County, Texas
Houston Plant, Harris County, Texas

1981

AMERICAN PETROFINA COMPANY OF TEXAS, Box 2159, Dallas, TX 75221
Port Arthur Plant, Jefferson County, Texas
AMINOIL U.S.A., Box 94193, Houston, TX 77018
Birthright Plant, Hopkins County, Texas
AMOCO OIL COMPANY, Box 401, Texas City, TX 77001
Texas City Refinery Plant, Galveston County, Texas
AMOCO PRODUCTION COMPANY, Box 591, Tulsa, OK 94102
Cowden, North Plant, Ector County, Texas
Edgewood Plant, Van Zandt County, Texas
Fullerton, South Plant, Andrews County, Texas
Midland Farms Plant, Andrews County, Texas
Slaughter Plant, Hockley County, Texas
Yantis, West Plant, Wood County, Texas
ARCO OIL & GAS COMPANY, A DIVISION OF ATLANTIC RICHFIELD COMPANY, Box 2819, Dallas, TX 75221
Fashing Plant, Atascosa County, Texas
Northeast Edgewood Plant, Atascosa County, Texas
ATLANTIC RICHFIELD COMPANY, Box 2451, Houston, TX 77001
Houston Plant, Harris County, Texas

The following observations are of interest.

- Apparently, Amoco Oil has changed their name and contact. This leads to questions of whether corporate relations have changed or whether these are really two different companies. Also, Amoco has an Oklahoma mailing address for five out of six of it plants.

- Atlantic Richfield has set up a new division. This leads to numerous questions. What are the new relationships? Where did the Northeast Edgewood plant come from? What are the new reporting entities? How does this impact historical data?

This example shows that even with a small portion of a relatively small list, problems can occur. Also, slight name changes mean that manual processing (in contrast to processing by computer programs) must be done.

Example 2. (Frame contains no State-level or consumption sector disaggregation of sales volumes, out-of-scope information inaccurate)

In 1980 the EIA sent out a new survey, Form EIA-172, "Sales of Fuel Oil and Kerosene." The survey was a successor to a previous voluntary survey, the BOM-6-1345, while the new one was mandatory. The 1979 EIA-172 survey form was sent to a probability sample of 6,537 companies out of 29,600 companies on the EIA-402 frame of fuel oil dealers.

The EIA-402 frame was constructed in 1979 by mailing a survey form to a list of companies constructed from the April 1979 Dun's Marketing List (SICs 5171-Petroleum Bulk Stations and Terminals, 5172- Petroleum Products nec, and 5983- Fuel Oil Dealers) and at least one EIA list. The EIA-402 survey form asked if each company was in business and if so, how much kerosene, distillate fuel oil and residual fuel oil it sold nationally. The EIA-402 national totals did not disaggregate volumetric sale to consumers from sale-for-resale volumes. Although it did not request volumes by state, it did ask in which states the companies did business.

The EIA-172 survey had to obtain data which allow obtaining estimates of final sales volumes for sectors such as residential, commercial, agricultural, industrial, and eight other categories at the State-level. Since State-level volumes were not

available, all EIA-402 respondents listed as doing business in more than one state were included in the sample with certainty. A probability sample of the remaining EIA-402 respondents was drawn based on the EIA-402 distillate volume. Due to time constraints, not all processing of EIA-402 "nonrespondents" (i.e. both out-of-scope categories and nonrespondents) could be completed. In order to try to protect against any severe errors caused by lack of proper identification of out-of-scope companies, the EIA-172 sample included five percent of the "nonrespondents."

Two problems are already evident:

1. Only EIA-402 national total sales volumes (including sale-for-resale) were available. Consequently, national and State-level volumetric coverage could not be determined.

2. National total sales volumes might not be a good predictor of total sales by consumption sector even at the national level. Consequently, estimated coefficients of variation associated with the sample might vary significantly from coefficients of variation of sector totals obtained from the EIA-172 data.

Figure 2 shows a plot of national distillate sales to consumers against national distillate sales to residential consumers obtained from the 1979 EIA-172 data base. A large number of different ways of partitioning the population using information available from the EIA-402 was tried. The plot chosen for display was the one for which total national distillate sales to consumers appeared to do the best job of predicting sales to residential consumers. Since many of the companies with the largest EIA-402 volumes did not sell or sold very little to residential consumers, only companies below various cutoffs were considered. The plot chosen for display has values which yielded the highest R-square value among the different groups considered. The plot clearly shows that total distillate sales cannot predict residential sales at the national level.

Figure 3 shows State-level R-Square values if total distillate sales to consumers is used to predict sales to residential consumers. Although R-square values are valuable summary statistics, they obscure what is happening in individual states. Figures 4 and 5 are plots for New Hampshire and Pennsylvania respectively. Although the R-square values for each state is 0.69, the plots are considerably different. One might confidently use total distillate sales in New Hampshire to predict residential sales. One might not in Pennsylvania. A number of the companies plotted having no residential sales are large multistate companies (sometimes refiners) who make sales to commercial, industrial, or on-highway diesel customers only.

The implication for frame building is that if residential volumes are not included as a frame attribute, then residential volumes may not be accurately estimated from sample surveys even if survey size is substantially increased. An extreme situation which illustrates the difficulty in obtaining information suitable for computing sector totals using samples is the following. Although the situation illustrated is more extreme than occurs in most states, it is presented because it represents the worst case of difficulties that can occur.

The largest (based on total sales volume) 30 percent of distillate retailers in a state make little or no sales to residential consumers while the remaining 70 percent make a large proportion of their sales to residential consumers. In order to obtain accuracy (i.e. low coefficient of variation) of estimates of total distillate sales to consumers and sales of distillate to residential consumers, disjoint samples in the two groups would have to be drawn. Basically, the largest 30 percent would provide most of the frame for obtaining estimates of total distillate sales while the smallest 70 percent would provide most of the frame for obtaining distillate sales to residential consumers.

Another frame problem which showed up only after receipt of the EIA-172 survey forms was the following. Two refiners from Texas who had not responded to the EIA-402 survey and who happened to be included in the five percent sample of EIA-402 "nonrespondents" submitted the EIA-172 survey form. One refiner erroneously submitted units in gallons instead of barrels. Inclusion of the responses of these two refiners with the corresponding national weight adjustment factor of 26 (nonresponse factor times sampling weight of 20) caused estimates of total distillate

sales to consumers in Texas to increase by more than 40 percent.

This is an example where a purely national frame used in drawing the five percent sample was not suitable for obtaining a State-level estimate in at least one state. If a volume had been previously present for the refiner which submitted volumes in gallons, the erroneous submission could have been quickly identified by an edit program. The refiner had to be identified by first obtaining State-level estimates and then determining which companies' submissions affected the estimates the most. Such identification can take several manweeks while an edit check, when appropriate volumetric data is available, takes very little time since the edits are part of normal verification procedures.

FRAME ATTRIBUTE SURVEY

In January 1982, the Administrator, EIA, approved a survey to obtain attributes for a Petroelum Products Sales Frame. This frame will be used to design the samples for Form EIA-782, "Monthly Petroleum Product Sales Report," and Form EIA-172, "Sales of Fuel Oil and Kerosene."

The mailing list for the frame attribute survey was constructed using:

- 11 EIA systems,
- 41 State lists, and
- 6 Industry lists.

The list is the largest ever constructed by EIA, consisting of 58,000 names and addresses. In addition, all names and addresses from EIA systems can be connected back to their original systems through a control number table. The merged system interconnects virtually all information about sales of petroleum products from all levels of the marketing chain that EIA collects.

Two versions, EIA-764-A and EIA-764-B, of a survey form which would collect attribute information were designed. The first survey form would request information about:

- clarification of address information,
- whether a company was still in business,
- volume of sales by State in the following categories:
  - distillate to residential consumers,
  - distillate to nonresidential consumers,
  - distillate sold for resale,
  - residual fuel sold to consumers,
  - residual fuel sold for resale,
  - gasoline sold to consumers, and
  - gasoline sold for resale.
- whether a firm is a parent firm, the names and addresses of the subsidiaries for which it reports, and
- whether a firm is a subidiary, the name and address of its parent if it does not report for itself.

Form EIA-764-B will request all the information collected by Form EIA-764-A except volumes of distillate sales. Form EIA-764-B will be sent to 6072 firms currently supplying distillate volumes on Form EIA-172, "Sales of Fuel Oil and Kerosene."

Clarification of address information is important even for out-of-scope companies since updating with non-EIA lists could add an out-of-scope company back into the frame. As long as the different statuses of out-of-scope companies are clearly identified in the frame, those companies will not be mistakenly surveyed. Volumes are necessary for evaluating coverage of the frame and for allowing efficient (i.e. small size) samples that allow obtaining accurate estimates. Sampling can yield a great reduction in overall respondent burden.

CONCLUSION

The survey frame is the structure which allows suitable statistical analysis to be done. Frame deficiencies can cause greater errors in the accuracy of estimates than errors caused by an other source.

REFERENCES

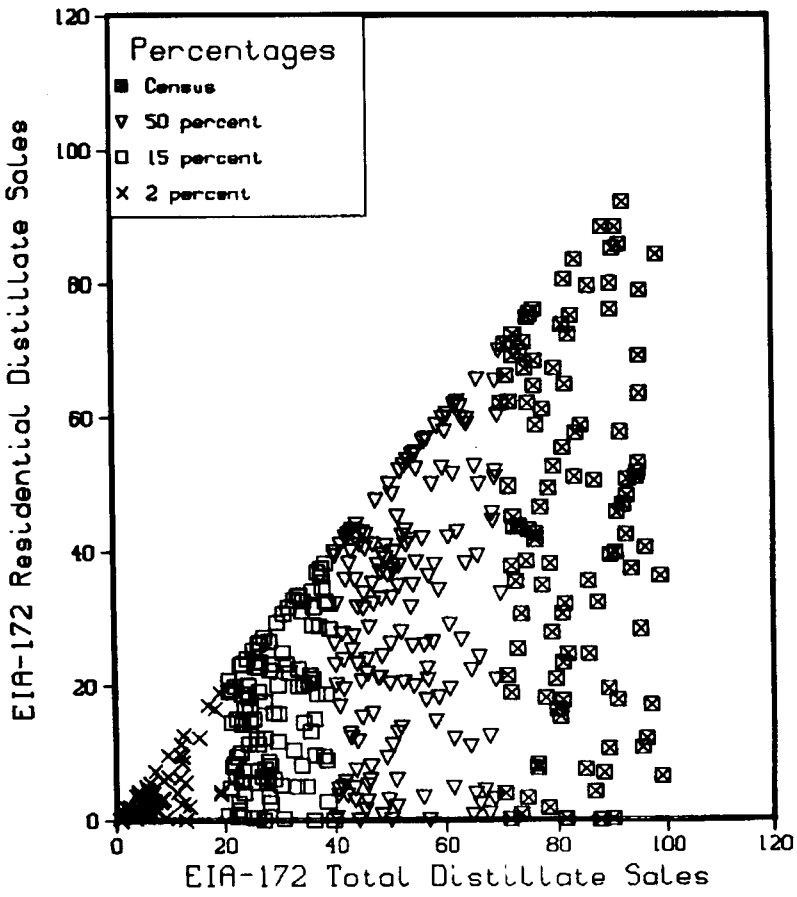
Cochran, W. (1977). "Survey Sampling." J. Wiley. New York

Hansen, M., Hurwitz, W., and Madow, W. (1953). "Sample Survey Methods and Theory, Vol II." J. Wiley. New York

Winkler, W. (1982). "The Energy Information Administration's Frames and Frame Development." Presented to the ASA Committee on Energy Statistics.

FIGURE 2

1980 EIA-172 National Distillate Volumes
Total Sales to Consumers less than 100,000 Barrels
Sample of Companies less than 70,000 Barrels



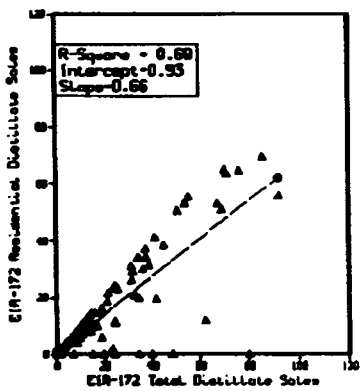
1980 EIA-172 Distillate Volumes
Total Sales to Consumers vs. Residential Sales
Units are Thousands of Barrels

FIGURE 3

Plot of R-Squares by State
Highest Residential Volume States

FIGURE 4

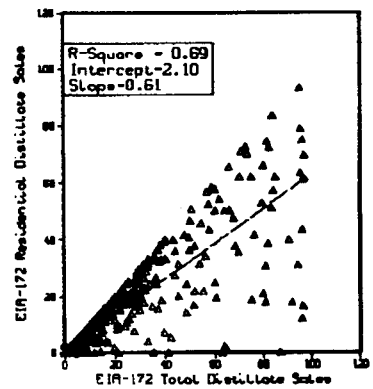
State of New Hampshire



Straight Line is OLS Regression Line

FIGURE 5

State of Pennsylvania



Straight Line is OLS Regression Line