Practical Methods for Electric Power Survey Data

James R. Knaub, Jr.
US Dept. of Energy, Energy Information Administration, EI-53, Washington, DC 20585

Key Words: design-based sampling, model-based sampling, regression weights, small area estimation, imputation, data editing, data quality

Introduction:

In the sections that follow, there will be a progression from a historical perspective, to current methodology, and finally to the newest research at the Energy Information Administration (EIA), in the Office of Coal, Nuclear, Electric and Alternate Fuels (CNEAF). This work has primarily been in the electric power surveys arena. Prior to the late 1980s, CNEAF had done no sampling, and apparently no imputation other than to substitute a response by the same establishment for a prior period. The primary function in the electric power area had been to track data for certain key utilities. Sampling began with an attempt to estimate for sales of electricity and associated revenue for the universe of utilities. A stratified random sample, with a certainty stratum for the largest respondents by State, was developed at the national level, for revenue per kilowatthour. Auxiliary data were not used for model-assisted inference. However, a ratio model was applied to State level sales by economic end-use sector. There were no variance calculations made for these State level models. Also, it was decided not to estimate revenue. However, this was a start. By late 1988 the author was involved and developed a comprehensive model-assisted, design-based approach and instituted a common identification system between the monthly sales and revenue sample and the corresponding annual census, so that auxiliary data could be more easily applied with fewer processing problems. By 1990, after a seminar by Nancy Kirkendall and others, Kirkendall, *et.al.* (1990), the author explored the use of regression modeling, and a number of applications to various CNEAF reported data elements have arisen since. In particular, regression imputation, when feasible, has the advantage over other forms of imputation of having an estimate of accuracy, the variance of the prediction error, readily available. Below are some details regarding developments from 1989 to the present. The author's opinions are his own and not EIA policy unless designated by other documents. An expanded version of this paper has been published in the on-line journal, ***InterStat*** (Knaub(2002)).

Data Quality/Processing:

Due to resource shortages, the nature of the data, and data customer demand, it has become increasingly important to make data handling as simple as possible. For electric power data, there are many customers who want a great deal of information on a monthly basis. The availability of quality data, and the ability CNEAF has to process them, are problematic. Nonsampling error can be overwhelming, especially for smaller respondents. Many people preparing and using the data may have industry experience or other strengths, but be very unfamiliar with statistics. This makes single imputation, scatterplot graphical edits, and the simplification of forms, data collection, file layouts and procedures, important to good data quality and timeliness in such a production environment.

Design- and Model-Based Sampling/Inference and Imputation:

Brewer(1995 and 2002) show how design-based and model-based inference can be used together. Currently, most sample surveys are design-based, or model-assisted design-based. (See Saerndal, Swensson and Wretman (1992) or Chaudhuri and Stenger (1992).) In Knaub (1989), the Keyfitz method of stratified random sampling with two observations per stratum was used for estimating electric sales and revenue. However, for electric power data, collecting monthly sample data from the smallest respondents is generally problematic. Data collected from such respondents may have relatively large nonsampling error. With regard to sales and revenue, a small utility may not read its meters every month, and in all cases, the billing periods are likely to be staggered rather than beginning and ending by calendar month. This impacts upon the accuracy of data collected monthly, especially for the smaller establishments that may not have the needed expertise at filing government forms. Beginning with Kirkendall, *et.al.* (1990), and Knaub (1990), model-based sampling and inference were considered at CNEAF. For sales and revenue data, it was found that using a previously established certainty stratum, of generally larger utilities, in a regression model, and dropping the two observations per stratum for smaller utilities from the sample, results were similar to the full, design-based case.

Using such a cutoff model-based sample is often criticized because model failure for the smaller values could be a problem. However, years of use and testing have shown this to be viable for various electric data elements. Some have greater variance than others (say generation by fuel type as opposed to sales by end-use sector), and various regressor data are available (although previous annual census data on the same data element is generally a good regressor), and even the number of regressors could vary, but model-based inference has generally proved itself useful throughout. Part of this could be due to the inherent problems in collecting data monthly from smaller establishments. Some larger utilities may have a department that provides such data. A small utility or unregulated facility (a small 'nonutility') may not have anyone available to fill out a form who really knows the difference between generation and capacity. As mentioned earlier, even a good guess at a monthly number may be unlikely to be forthcoming for many smaller entities. One might want to impute for these plants, regardless. An imputation implies the use of some kind of model. When regression can be used, at least an indication of the quality of the resulting data will be possible: the standard error of the prediction error. Although randomization can provide protection from model failure, if the data requested are best imputed anyway, then this protection is at least partly an illusion. Further, electric power data, like establishment surveys in general, are highly skewed. Therefore, there can be many small establishments that, if excluded from a sample, do not need to be estimated very well to still have a good estimate overall. Therefore, it is advisable to collect a monthly sample of generally larger establishments with as little nonsampling error as possible, and an annual or less frequent census, also with great care. CNEAF has found it is often very difficult to collect a census well, even on an annual basis.

Thus, the methods developed recently for CNEAF estimation and imputation are solely model-based. However, these methods are consistent with use not only with model-based sampling, but also as imputation for design-based sampling, and imputation for incomplete census surveys. (See the bottom of page 317 in Lee, H., Rancourt, E., and Saerndal, C.-E. (1999/2002).) Further, these methods are complementary with scatterplot graphical editing, which is highly efficient and effective in maintaining data quality (although nothing substitutes for collecting data well from the very beginning, to include better form and file designs). More will be said about scatterplot editing in a succeeding section.

Add-on Data:
Another detail worth mentioning is a term used by the author to describe a response collected in a current survey, that has no counterpart among the regressor data**:** an "add-on." If, for example, a previous census is used for regressor data, but there has been a 'birth,' or new member added to the population from which a current survey is taken, then data for that case would be an 'add-on.' Suppose there were a stratum containing 100 members of the current population, and say that there were 97 of those members that corresponded to regressor data. If 25 observations were taken, 22 should be from among the 97 members of the population that have regressor data, and the other three should be add-ons. These add-ons only represent themselves. If there are odd members of the population inextricably scattered among both observed and unobserved data, then those are not to be treated as add-ons, unless it is determined that the part unobserved is miniscule and will thus cause little downward bias.

Frame Maintenance:
This leads to the concept of frame maintenance. If, for example, establishments in the universe were to merge, and the resulting establishment was sampled, or it was a response to a census where there is nonresponse, then any regressor data should be correspondingly merged before estimation/imputation takes place. This is best done as a part of a frame maintenance system that tracks what data and software are used each reporting period. 'Births' can be covered as in the paragraph above. 'Deaths' must result in the removal of an establishment from appropriate frames. If edits/investigations show regressor data to be unreliable, then the corresponding datum in the current sample (or perhaps preliminary census) must be considered an add-on. (If a census has nonresponse, it may often be useful to impute to provide preliminary aggregate numbers, but if that same census is going to be used as regressor data later, one would still pursue responses, if feasible. That would also help test the reliability of the imputations.) Further, if a response is believed to be unreliable, it might be best to impute to replace it, as if it were a nonresponse, or in the case of a model-based sample, as if the respondent were not in the sample. Proper file maintenance requires appropriate software tools and care. It is best to keep everything as simple as possible, and to keep appropriate records of changes. "Flags" in data record fields may be helpful. (See Kovar and Whitridge (1995).)

Papers and Articles Most Relevant to CNEAF Data:
(Note other relevant references and additional information that appear in Knaub(2002). In the interest of space, the list below is abbreviated.)

Knaub (1989), "Ratio Estimation and Approximate Optimum Stratification in Electric Power Surveys," presents the use of the model-assisted design-based Keyfitz method (two observations per stratum for stratified random sampling), with a certainty stratum of the largest respondents, to estimate sales, revenue and revenue per kilowatthour at State and more aggregate levels, by economic end-use sector. (Note that whenever there is a certainty stratum, there is no contribution to an estimate of survey variance from that stratum, and generally there is no nonsampling error considered, which therefore assumes 'perfect' data. There will be more on this in a succeeding section.) Dean Fennell in the CNEAF suggested testing this method by taking an annual census, temporarily removing responses gathered from members of the universe not in the sample, doing the estimations, and then comparing results to what had been obtained. Since then, this procedure has been used repeatedly at the CNEAF to test new methodology. Although it has obvious advantages, one must also consider the impact of seasonality.

Kirkendall, *et.al*. (1990), "Sampling and Estimation: Making Best Use of Available Data," was a seminar which provided an introduction to the use of regression modeling for analysis and estimation of energy data.

Knaub (1991), "Some Applications of Model Sampling to Electric Power Data," mentions that "One method developed may be of particularly wide application." This shows that in the case of model-based ratio estimation, relative standard errors may still be estimated even when only the subtotal for a regressor is known for the unobserved data.

Knaub (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," looks at a method for measuring heteroscedasticity by decomposing residuals into random and nonrandom factors.

Rao, P.S.R.S. (1992 – influential in 1994) was subsequently applied to revenue per kilowatthour variance estimation. As an adjunct EIA employee at the University of Rochester, Dr. Rao corresponded with solutions he derived for model covariance estimation, which were used for more than six years, beginning *circa* 1994.

Knaub (1997), "Weighting in Regression for Use in Survey Methodology," is a study of the impact of regression weights on survey results. It includes consideration of the accuracy with which heteroscedasticity might be determined, and suggested 'default' solutions. The need for this was apparent when Sweet and Sigman (1995) pointed out that Knaub (1993) did not "…present specific criteria …" for all situations.

Knaub (*circa* 1998), "Model-Based Sampling, Inference and Imputation" is a very simple explanation of the concept of regression modeling, placed on the EIA website as a response to telephone calls asking for clarification of this methodology.

Knaub (1999), “Using Prediction-Oriented Software for Survey Estimation,” promotes a new concept in which the author proposes that any commercial software that can be used for predictions, and allows the programmer to save and reuse certain statistics, can be used to very simply and effectively apply regression to small area estimation, other model-based sampling or general imputation. It integrates well with scatterplot (graphical) editing. The simplicity with which data may be handled is a tremendous advantage when nonsampling error and processing problems plague an organization short on resources. Also, data may be grouped for estimation purposes in a different manner than they will be grouped for publication. This allows for a form of “borrowed strength,” as well as to avoid using a single model on data heterogeneous under that model. By seeking homogeneous “estimation groups,” nonignorable nonresponse may be converted to essentially ignorable nonresponse.

Knaub (2000), “Using Prediction-Oriented Software for Survey Estimation - Part II: Ratios of Totals,” deals with ratios of totals. An example from the electric power industry would be the estimation of revenue per kilowatthour and its associated variance estimate.

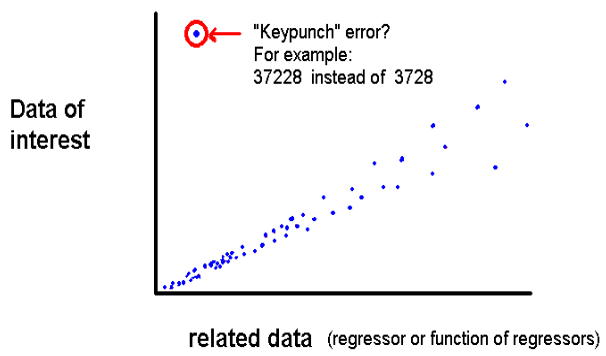
Knaub (2001), “Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias.” A succinct description of this method is found on pages 12 through 15. Bias does not appear to be a substantial problem.

Ancillary sources:
- In addition, the author benefited by correspondence over a number of years with K.R.W. Brewer. His book, Brewer (2002) contains an interesting discussion of heteroscedasticity for establishment surveys.
- Also, it is known that Dr. Roger L. Wright has conducted load and other research, also finding uses for regression modeling.
- Carroll and Ruppert (1988) and Valliant, Dorfman and Royall (2000) are also valuable resources.

Note that a member of the ASA Committee on Energy Statistics, Dr. F Jay Breidt, has suggested calculating variances more rigorously than in Knaub (1999, 2000, 2001), but the author is of the opinion that numerous processing problems call for more simplicity, wherever possible.

Scatterplot (Graphical) Data Editing as related to Modeling:
Perhaps the most useful possible graphical edit is the scatterplot, which has been used at CNEAF for at least six years on the sales and revenue monthly survey, simply by using SAS PROC PLOT and matching points on the plot to data in tables. Currently, scatterplots are also in place for editing generation, fuel consumption and stocks data. Data managers have shown enthusiasm for this type of editing, and a seminar conducted by the author on this topic was well received. However, resource shortages have slowed the implementation of scatterplot editing procedures. Further, using paper copies of plots and tables can be cumbersome. A ‘point-and-click’ version of these edits for the personal computer (PC) has been requested. This would make identification of suspect data points and links to respondent contact reports (RCRs) easy to accomplish. In the meantime, Dr. Orhan Yildiz, an analyst working for a contractor at the EIA, improved the paper scatterplot edits, using the capability of SAS to determine a weighted least squares confidence band, and then list those responses, along with respondent identification, that fall outside of the prescribed confidence band. Because points nearer the origin have larger weights, they can greatly influence regression lines that may then be used to impute for missing data. (The missing data can be either the result of nonresponse, or if a sample is used, they could be due to establishments that are not in the sample.) It is reasonable then that some points near the origin may have an influence on results greater than one might expect from appearance alone. This has proved to be the case, and to a greater extent than the author previously expected. Therefore, once a ‘point-and-click’ PC version of scatterplot editing is made available, a list of responses that fail this edit may still be desirable. Another possibility favored by Dr. Yildiz would be the ability to isolate and expand the PC view of the portion of the graph nearest to the origin. Data revision activity may be supplemented by a change in regression weight(s), such as discussed in “Thermometer Effect” below.

When a single regressor is used, then at least one choice for a plot is obvious. One would have the data of interest on the y-axis and the regressor data on the x-axis. For example, one could plot monthly sample sales data on the y-axis and corresponding data from the most recent annual census of sales could be plotted on the x-axis. The same could be done for revenue data. One would also want a plot of sales *vs.* revenue to guard against cases where either data element was consistently reported in incorrect units. For a multiple regression case, separate plots could be done for each regressor, or the x-axis could be a function of regressors, such as a rough estimate of the y-values. When scatterplot edits include the relationships used for imputation, one ensures that data irregularities that impact most on the imputation process will be addressed.


Scatterplot form of Graphical Editing

Scatterplot graphs can help data managers learn more about the relationships between their data sets. Some scatterplots may show much weaker relationships than others. Data may not appear to be obviously linear. However, the most important/flagrant data problems are generally easy to recognize. Commonly, because of incorrect keystrokes or reporting in incorrect units, data may be in error by one, three or even six orders of magnitude. When these errors are not corrected, there can be a very large, upward bias. (Note that 2000, 5000, 4000 and 3000 incorrectly reported as 2, 5, 4 and 3,000,000 will overestimate, although three numbers were underestimated by three orders of magnitude and only one was overestimated to that same degree.) To discover and correct these errors should be the goal when editing data. Trying to edit more finely may bias the data and waste resources that should be spent on other areas. It is important to strive for error-free data collection, rather than trying to repair the damage later, as to a large extent, this will not be possible. (See Data Quality, at http://www.dataquality.com/.)

Notes on Data Collection and Related Topics:

Data managers may not always realize the impact of mishandling data, and the statistician should be on the alert for such circumstances. The statistician should be aware of the kinds of mistakes that can be made and to try to prevent them, or at least recognize them when they present themselves or someone else presents them. Having the right tools (graphical edits, for example) does not guarantee proper use of them. The statistician needs to communicate well with the data managers.

With regard to data quality, there is no substitute for careful data collection. Otherwise, nonsampling error can have devastating effects. As an extended view, this starts with forms designs that take into account the level at which data are to be collected and how these data may be merged from different surveys. A comprehensive data quality control program would also limit opportunities for computer file corruptions and make data processing as simple as possible. To this end, the method of Knaub (1999, 2000 and 2001) contributes by providing either an observed or an imputed number for every member of the population, for any given data element. Thus data managers may more easily determine that numbers are reasonable, that there are no duplicate or missing records, and data may be more easily handled and stored.

The volume of data to be processed is of concern. The Office of Management and Budget (OMB) must periodically approve burden to the industry. Further, it is not practical for CNEAF, given resource restrictions, to process volumes of data that are too huge, on too frequent of a basis. Many respondents may not even be capable of supplying good quality data within the time frame needed, and any attempt by CNEAF to process such data may be counterproductive. Thus, in the case of monthly samples, simply increasing a sample size may actually decrease already flagging accuracy. Also, there may often be a tendency to attempt to publish too many numbers. If, for example, resources and data collection conditions make collecting more than approximately 3000 of the largest responses impractical, yet 2000 'aggregate' level numbers are to be published due to a wide variety of data customer interests, then accuracy, timeliness or both must suffer.

Special circumstances may arise, which are sometimes handled by special adjustments to the data. This may often be detrimental to data integrity, and a statistician should at least insist that a record be kept of what was done, and footnotes made to published tables so that data customers are made aware of what processes resulted in the data presented. Unfortunately, data revisions/adjustments, changes in priorities, changes in types of fuel to be reported, and file maintenance and software problems, may all tend to obscure results. Human resource shortages and trying to meet frantic customer demands can also be detrimental to data quality.

"Thermometer Effect":

When examining scatterplot edits, it has often been apparent that variance has been unusually large near the origin, in these regressions through the origin. Reference here is not to errors that are one or more orders of magnitude, which may occur anywhere on the graph (forcing the regression line to be near one of the axes, when a number has been reported in incorrect units). Nor is this a reference to ordinary heteroscedasticity, through which we expect larger variances with greater distances from the origin. Here, a linear regression through the origin may seem very appropriate, and variance may generally increase with increasing regressor values, but there may be a 'bulge' about the origin that provides what might be described as a "thermometer effect." That is, the points on the graph may resemble an old thermometer with the reservoir at or near the origin. This appears to be due to three possible sources: 1) nonsampling error, 2) outliers such as data that should have been add-ons, and 3) special cases that cannot be placed into their own estimation group because they probably appear throughout the observed and unobserved cases in an unknown manner. For the first two sources of this problem, careful data management would be most helpful. In all three cases, however, attention to regression weights might be used to alleviate the problem to some extent.

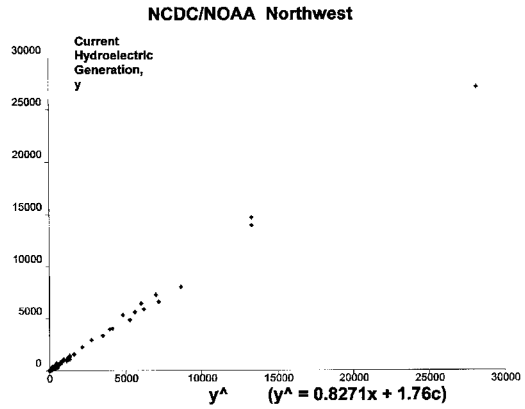
An illustration of the "thermometer effect" can be seen in Knaub (1996) on the Internet. It may be that a short term solution could be to use ordinary least squares (OLS) regression as opposed to weighted least squares (WLS) regression in an attempt to avoid problems near the origin. Probably the best solution, however, is to better address the data problems.

Experiments with Estimation Groups:

In Knaub (1999), Knaub (2000) and Knaub (2001), an imputation and estimation methodology is described that allows for very flexible data handling and storage. (See particularly, Knaub (2001), pages 12 through 15.) Data may be collected and grouped one way for purposes of imputing missing data, and then published under a different method of grouping for publishing aggregate numbers. For example, data on hydroelectric plants may be collected according to US Standard Regions for Temperature and Precipitation, as determined by the National Climatic Data Center (NCDC) of the National Oceanic and Atmospheric Administration (NOAA), thus determining the estimation groups. Aggregate data could then be published, for example, by North American Electric Reliability Council (NERC) regions. (See Knaub (1999).)

Note in the following graph that "x" represents hydroelectric generation from a census that was previously collected, and "c" represents nameplate generation capacity.

An Example of what is Probably a Good Estimation Group from Page 24 of Knaub (1999):



Relative Standard Error under a Superpopulation (RSESP):

A very substantial remaining topic is that of measuring nonsampling error. Total survey error, including nonsampling error, which is all that remains if the survey is a census with no nonresponse, may be the subject of a great deal of effort in an attempt to control it, but there will always be nonsampling error remaining. If it is 'known' to be negligible, that would be fine, but this is not often the case, and how would one know if it were the case?

Revisions to data may sometimes provide some indication of the level of nonsampling (measurement) error. A table of such information is published in the appendices to each issue of the ***Electric Power Monthly*** (***EPM***), published by CNEAF. These revisions are often referred to as "revision errors." However, obviously not all errors will be revised; some numbers will be revised to numbers that are less accurate; no values are known exactly; so, the term "revision errors" is something of a misnomer.

In Knaub (1999), on pages 8 and 9, there is a suggestion that could be used for a sensitivity analysis when investigating nonsampling error. Near the end of the ASA CD version of Knaub (2001), another suggestion is made regarding an 'average' error, but that would need to be applied to the total. Further consideration has therefore lead to the following**:**

An estimated relative standard error (RSE) measures the damage to accuracy of an estimated total due to the fact that some data are not observed. This estimate is impacted by nonsampling error in the observed data. However, if all data are observed, then the RSE is zero, because it only applies to the part of the universe that is not observed. Suppose that in design-based sampling, the sampling variance were applied to all data, thus treating the population under study as a part of an infinite superpopulation from which any data observed are actually taken. Thus we would drop the finite population correction (fpc) factor. This reference to a superpopulation is reasonable if we consider each observation as having a measurement error associated with it with some degree of randomness. We will have made one observation from each infinite set of possible observations in each case.

The same can be done for model-based sampling, or for model-based (regression) imputation for nonresponse in a census, by considering predicted values for the entire universe, even where there are observed values. In the case of a complete census, we would consider that the data observed were still only a sample from the infinite population of all possible observations that include measurement error. It would be a special collection of such observations in that there would be one corresponding to each of a specific set of respondents. Each respondent could, however, have provided a different observation with different measurement error.

Thus, consider the relative standard error under a superpopulation (RSESP), a measure that would make use of design-based sample variances, or, for modeling, residual and model coefficient variances. See Knaub(2002). The RSESP partially reflects total survey accuracy. The RSE, however, estimates loss in accuracy due to the fact that not all members of the finite population respond, either by design (a sample), or failure to respond when queried (nonresponse). The RSE assumes all observations are without measurement error, although large estimated RSEs have been used to identify the presence of large nonsampling/measurement error. The RSESP, therefore, estimates overall loss in accuracy due to inherent variance, which can be largely due to nonsampling/measurement error, even when a census is taken. However, it can also be influenced by other factors, such as less than optimal grouping of data. In Knaub (1999), Knaub (2000) and Knaub (2001), the estimation groups described are flexible, and one challenge is to attempt to pick them optimally. Failing to include all data that reasonably can be described under a single model needlessly increases variance. However, mixing data that are not reasonably described under a single model also damages accuracy. When the estimation groups are picked well, nonresponse may be "ignorable" as opposed to "nonignorable."

Thus, like the RSE, the RSESP is influenced by factors other than those intended. However, they can be treated as somewhat indirect, but fairly useful measures for sampling error and total survey error, respectively. RSE has perhaps a wider range of very meaningful application. When sample sizes are small, it may be all that is needed to convey overall accuracy if bias is not a substantial problem. (See Knaub (2001).) However, when dealing with a census or near census, RSESP becomes much more important than the RSE, as an indicator (indirect measure) of accuracy.

Think again of the sample or census as being from a larger superpopulation, and after it is used to establish model parameters and variances, predictions are made in place of all observations. If total variance in such a case is low, nonsampling error is not likely to be a problem.

Conclusions:

In the introduction it was stated that "Prior to the late 1980s … The primary function in the electric power area had been to track data for certain key utilities." Even today, there seems to be an emphasis on this. However, there has been some concern with estimating for the data that are not collected. Considering the relationships that can be found between various sets of electric data, regression has become very useful and should continue to take on a larger role. This does not preclude the use of design-based sampling and inference in some instances as part of an overall system, but simplicity has advantages in this production environment where customers want a lot of data available in a short time frame. A strictly model-based regression approach has been quite successful, in spite of data collection problems and limited resources due to budgetary cutbacks. Randomization is not helpful when the smallest respondents cannot be relied upon for data of acceptable quality on a frequent basis. What is being done at CNEAF is **mass single value regression imputation, primarily for the smallest members of highly skewed populations**, for which, as in sampling, data are not sought on a frequent basis. However, a census is sought on a less frequent basis, and that can be used as regressor data. Research continues to be done to improve on this system.

Further, regression modeling is strongly related to scatterplot (graphical) editing, which is a highly efficient manner of editing, which can also be used to help train data managers who need to learn about relationships between data sets. Finally, an emphasis is being placed on measuring nonsampling as well as sampling error.

Acknowledgements:

Thanks to Orhan Yildiz and other contractor and government personnel at the Energy Information Administration for helpful conversations.

References (See expanded list in Knaub(2002)**:**

Brewer, K.R.W. (1995), "Combining Design-Based and Model-Based Inference," Business Survey Methods, ed. by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, John Wiley & Sons, pp. 589-606.

Brewer, KRW (2002), Combined Survey Sampling Inference: *Weighing of Basu's Elephants*, Arnold: London.

Carroll, R.J., and Ruppert, D. (1988), Transformation and Weighting in Regression, Chapman & Hall.

Chaudhuri, A. and Stenger, H. (1992), Survey Sampling: Theory and Methods, Marcel Dekker, Inc.

Kirkendall, *et.al.* (1990), "Sampling and Estimation: Making Best Use of Available Data," seminar at the EIA, September 1990.

Knaub, J.R., Jr. (1989), "Ratio Estimation and Approximate Optimum Stratification in Electric Power Surveys," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 848-853.

Knaub, J.R., Jr. (1990), "Some Theoretical and Applied Investigations of Model and Unequal Probability Sampling for Electric Power Generation and Cost," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 748-753.

Knaub, J.R., Jr. (1991), "Some Applications of Model Sampling to Electric Power Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 773-778.

Knaub, J.R., Jr. (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, pp. 520-525.

Knaub, J.R., Jr. (1996), "Weighted Multiple Regression Estimation for Survey Model Sampling," InterStat, May 1996, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1996.)

Knaub, J.R., Jr. (1997), "Weighting in Regression for Use in Survey Methodology," InterStat, April 1997, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1997.)

Knaub, J.R., Jr. (*circa* 1998), "Model-Based Sampling, Inference and Imputation," found on the EIA web site under http://www.eia.doe.gov/cneaf/electricity/page/forms.html.

Knaub, J.R., Jr. (1999), "Using Prediction-Oriented Software for Survey Estimation," InterStat, August 1999, http://interstat.stat.vt.edu, partially covered in "Using Prediction-Oriented Software for Model-Based and Small Area Estimation," in ASA Survey Research Methods Section proceedings, 1999, and partially covered in "Using Prediction-Oriented Software for Estimation in the Presence of Nonresponse," presented at the International Conference on Survey Nonresponse, 1999.

Knaub, J.R., Jr. (2000), "Using Prediction-Oriented Software for Survey Estimation - Part II: Ratios of Totals," InterStat, June 2000, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 2000.)

Knaub, J.R., Jr. (2001), "Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias," InterStat, June 2001, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 2000.)

Knaub, J.R., Jr. (2002), "Practical Methods for Electric Power Survey Data," detailed version, InterStat, July 2002, http://interstat.stat.vt.edu.

Kovar, John G., and Whitridge, Patricia J. (1995), "Imputation of Business Survey Data," Business Survey Methods, ed. by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, John Wiley & Sons, pp. 403-423.

Lee, H., Rancourt, E., and Saerndal, C.-E. (1999), "Variance Estimation from Survey Data Under Single Value Imputation," presented at the International Conference on Survey Nonresponse, Oct. 1999, published in Survey Nonresponse, ed by Groves, Dillman, Eltinge and Little, 2002, John Wiley & Sons, Inc., pp 315-328.

Rao, Poduri, S.R.S. (1992), unpublished correspondence, Aug. - Oct. 1992, on covariances associated with three Royall and Cumberland model sampling variance estimators.

Saerndal, C.-E., Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling, Springer-Verlag.

Sweet, E.M. and Sigman, R.S. (1995), "Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data," Proceedings of the Section on Survey Research Methods, Vol. I, American Statistical Association, pp. 491-496.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000), Finite Population Sampling and Inference: A Predictive Approach, John Wiley & Sons.