DISCUSSION

Tommy Wright, Union Carbide Corporation, Nuclear Division

It pleases me that my colleagues have permitted me to serve as a discussant in this session. The true test of our friendship after this discussion will come when we see each other in Oak Ridge on Friday of this week. I make a distinction between a discussant and a referee. If I were a referee in this session, this would mean that I would need to know something about the areas discussed in each of the papers and be able to evaluate their worth. However, I see the role of a discussant requiring less qualifications. The discussant is permitted to give personal comments and raise questions. While I claim to be no expert, I hope that some benefit will be realized by this effort.

Being a discussant today reminds me of some comments made by D. R. Cox before discussing a collection of papers on a similar occasion. He said, "Discussants generally start off by praising the papers and their authors only to be followed quickly by 'HOWEVER' ... "

Not wanting to break with tradition, I would like to applaud the presenters of the papers today and encourage continued efforts by them and others towards seeking solutions to the many problems surrounding the quality of data collected by data collection systems, and in particular those related to energy data. (It is important to note that the efforts at Oak Ridge National Laboratory have been underway for less than two years. More time is required before significant results are realized.)

The paper by Loebl and Cantor gives an overview of the Energy-Data Validation Program which can be viewed as a joint effort by Oak Ridge National Laboratory and the Energy Information Administration. As the authors point out, this program was initiated largely because someone raised the question, "Is our energy data trustworthy?" I am not sure that I agree with the authors when they say that this is a statistical question. It may be more appropriate to say that statistics is a tool which might help one in the assessment of the quality of the energy data or which might suggest techniques for collecting data which may lead to improvements. At least through assessment, we will be able to know approximately to what degree we can trust the energy data.

I agree with the authors when they say that the quality of the data is not to be considered without an awareness of the needs of data users and the cost those users are willing to pay for improved data. Whether we say audit, evaluation, or validation, the important things are the lessons learned and the improvements which are initiated for future data collection.

The statistical areas of outlier detection, automatic data editing, exploratory data analysis, and sampling have been major methodology tools which have been considered in the validation process. Other areas which seem applicable include: time series, multivariate analysis, and pattern recognition.

Many of the data collection systems take censuses mainly because they support federal regulations. I am not convinced that federal regulations should imply always that censuses should be taken. It seems to me that even compliance with regulations could be monitored by considering use of various types of sampling schemes. Indeed more accurate and timely data can be obtained by sampling. The opponent to sampling would argue that in addition to the need to monitor the activities of all members of the target universe, there would be a loss of detailed information for small domains of interest. But there are methods of controlled selection, including multi-way stratification, which can yield reasonable results, not to mention the techniques for small area estimation. One major problem which Oak Ridge is considering but was not mentioned which seems to be a constant worry has to do with the imperfect frame problem. How well does the frame match the target population, and what techniques are useful when frame and target population are believed to be different?

The goal of assessing the accuracy of a given data collection system is indeed noble. In so doing, it is not necessarily the case that the validation analyst must use an alternative method for determining the value that should have been reported by a given respondent to a particular system. If the respondent used the correct method, then it should be the same method as that to be used by the validation analyst.

Loebl and Cantor make mention of an error model approach consisting of several components. I am not sure that I agree with them when they say that "the analyst needs to understand the relationship of error components." It is true that this understanding would lead to major forward steps, but for the immediate future, it may be a bit ambitious. It seems to me that it would be significant if one could locate a reasonable number of the individual sources of error and determine their impact. Those sources where the impacts are greatest would be those areas where one might use his resources initially to reduce the impact.

I agree with Loebl and Cantor when they say that the concepts and methods designed for the validation of energy data are sufficiently broad in scope to apply to many government - mandated data collection activities". Such agencies are constantly seeking ways to improve the quality of their data. In fact, a Panel of the National Academy of Sciences is currently reviewing the Statistical Program of the Bureau of the Mines at that agency's request with a focus on data quality and better ways of collecting the data.

The paper by Pack on preliminary internal data screening is very appropriate. Too often one is quick to provide a correct solution for the wrong problem. With that thought in mind, it seems fitting therefore in evaluating a data base to use the preliminary internal data screening as a means of letting the data speak for itself rather than attempting to force the fit of a model without knowledge of its appropriateness.

Pack mentions three dimensions for data bases: variable, cross-section, and time. There might be another dimension one may want to consider which is - Type of Respondent.

Though the preliminary internal screening may

pick these up, one has to beware of the problems of different target populations over time, different definitons of the same variable over time, preliminary data vs final data, etc.

I agree that there are some benefits to be realized from external data screening, or the comparison with other data sources; however, as Pack notes, the methodology for doing it is not clear. Indeed one should beware that close agreement between two different data bases estimates of the same parameter does not necessarily mean that the estimates are accurate, just as disagreement between two estimates does not imply that at least one of the estimates is in error, for a closer examination might reveal two different targets.

It seems possible that a sampling inspection type procedure would be another type of classical confirmatory type of statistical test that can also be used when centering upon quantification and summarizaation or probable nonsampling errors and error patterns.

The variable Stem-and-Leaf Plot is attractive because it not only summarizes the data, but it also has the ability of preserving the original data. I am concerned however about its use in very large data bases and for highly variable data. In such cases, its use might lead to Vine-Stem-and-Leaf Plots. Further investigation is needed to determine the application of the approaches in Table 2 and others to large data bases.

I am a believer in Bayesian type procedures, and I believe that the thoughts presented in the paper by Liepins and Pack on maximal posterior probability can possibly be useful in imputing an observed vector $\underline{y}$ which fails certain edits. The idea of replacing $\underline{y}$ by that $\underline{x}_0$ such that

$$p(\underline{x}_0 | \underline{y}) = \max_{\underline{x}} p(\underline{x} | \underline{y})$$

is appealing. However in practice, Bayesians know that it is a task to choose an appropriate prior distribution $p(\underline{x})$. This seems especially true for energy data. It is not clear that the use of a uniform prior is appropriate as suggested in equation (18) even though one often thinks of it as a noninformative prior. Studies will show that if the prior is incorrectly chosen, then one may be doing worse than he would had he not considered a formal Bayesian approach.

It is also a task in practice to choose a prior distribution that is meaningful and at the same time that leads to a posterior distribution in a manner that is tractible. If the mathematics is not tractible, then one is forced into making approximations for which he may have no feeling of their goodness.

Of course, similar comments apply towards identifying the set of fields to impute, i.e. error localization.

While independence of errors between fields was assumed for simplicity, for application, one will also want to consider the more general and realistic case of dependence.

I agree that further work is needed before applications are possible.

The discussion in the paper by Downing and Pierce on a comparison of the outlier detection methods appears quite adequate. The authors are very clear in their discussion and do a good job of indicating areas for further research. It is not clear however why these specific six methods were selected for this study. Are there other known methods for multivariate data? Are the authors aware of any analytical results for comparing multivariate outlier detection methods?

In closing, I would like to note that exploratory data analysis, outlier detection, and error localization are all POST-SURVEY energy data validation techniques. As Oak Ridge National Laboratory realizes, one can not hope to achieve much toward controlling the quality of data only AFTER it has been collected. At best it seems that one can assess the extent of the damage done, identify sources of trouble, and seek ways to diminish the effects of the sources of trouble. For an effective quality control program, more attention needs to be given to PRE-SURVEY considerations, that is, an organized effort promoting PRE, DURING and POST Considerations is needed.

Such a comprehensive program of Quality Control for Data Collection Systems would include the following considerations:

i) A clear statement of the problems of the data collection systems and an understanding of the subject matter,

ii) A classification of errors and an error profile,

iii) Studies to determine the usefulness of various types of error models,

iv) Statements supporting the design of the data collection systems with special attention on the need to have a census, a survey, or a combination of both,

v) Statements on the adequacy of the frame (including plans for updating)

vi) A thorough review of the survey form (or questionnaire)

vii) An application of exploratory data analysis and pattern recogniton techniques, and

viii) An application of outlier and automatic data editing techniques.