ENERGY-DATA VALIDATION: AN OVERVIEW AND SOME CONCEPTS[1]

A. S. Loebl and S. Cantor, Energy Division, Oak Ridge National Laboratory

ABSTRACT

Energy-data validation can be viewed operationally as a three-fold assessment process: (1) a determination of the quality of the data collected, i.e., an assessment of accuracy; (2) an analysis of the relevance and usefulness of the data so as to assess how closely the data collected is meeting the requirements of its users; (3) an assessment of measures that can be taken to enhance the effectiveness of the data system under study. Assessment (1) is akin to the process of "critical" evaluation of data in the physical sciences. Assessments (2) and (3), the more distinctive features of validation, underlie two of its important goals. In brief, the analyst validates both the information and the requirements for the information, and as deficiencies in these two aspects are uncovered, the validation analyst formulates and evaluates the means for correcting these deficiencies.

This paper focuses upon the use of an error model for systematizing the assessment of accuracy in a data-system validation study. Other topics discussed briefly are: a capsule history of energy-data validation; ORNL studies of outlier detection methods, automatic data editing and sampling theory in support of the energy-data validation program; reviews of energy-data requirements.

Introduction

This paper provides an introduction to this session and a very brief background to the subject of energy-data validation. We highlight the statistical and mathematical research carried out at Oak Ridge National Laboratory in support of energy-data validation, then discuss certain statistical aspects of energy-data system validation. These aspects represent, however, but one set of facets to a profoundly intriguing and practical question: is our energy data trustworthy? Certainly, this is a statistical question, but it is also a question involving a host of other considerations—compliance with regulations, phenomenological and economic implications, and ultimately, a philosophical question: how does the act of checking the data alter their content and magnitude?

[The session will continue with three papers on methodological aspects of post-survey energy-data validation: (1) preliminary internal data screening, (2) outlier detection, and (3) error localization.]

The validation of energy data is concerned with more than assessing data quality; it also seeks to determine the needs of data users—especially users who develop, implement, or evaluate energy legislation and energy policy. The strong linkage between usage and assessment may well be the distinguishing feature of data validation as compared with more traditional data assessment efforts, such as auditing (e.g., tax returns), or evaluating scientific data. Data auditing may or may not presume duplicity, but it does assume that there is "a right answer" and that the good of the whole is best served by discovering each element of accurate information. Data evaluation, on the other hand, is broader in scope; its goal is to assign best estimates to a set of data values together with estimates of uncertainty of those data values. However, common to all data assessment efforts is the idea that, if the investigation is careful and rigorous, the means for improving data quality will suggest themselves.

In essence, data validation is a process for determining the strengths and weaknesses of a data base with respect to the needs of its users, the ultimate goal being correction of the weaknesses. Energy-data validation can be viewed operationally as a three-fold assessment: (1) a determination of the accuracy of the data collected, (2) an analysis of the relevance, usefulness, and completeness of available data so as to assess how well presently collected data are meeting the requirements of users, and (3) an assessment of measures that can be taken to eliminate (or ameliorate the effects of) errors in the data system under study. This three-fold assessment is permeated with statistical considerations as well as with issues amenable to the rigors of inferential analysis.

But, the predominate foci of this session are more narrowly defined: first, to an overview of energy-data validation, and then for the balance of this session, to the frontiers of data-base management, to statistical issues that deal with efficient analysis and correction of a large volume of data.

Capsule History

A greatly increased sensitivity to the validity of energy data was one by-product of the widespread concern about national energy needs that followed the energy "crisis" in the winter of 1973-74. The Federal Energy Administration Act of 1974 created a new agency (the FEA) to deal with the Nation's energy shortages, giving its Administrator wide ranging information-gathering power. In 1976, by amendment of the FEA Act, Congress extended the agency's information activities by establishing the Office of Energy Information and Analysis within FEA, charging its Director with establishing and maintaining the "scientific, engineering, statistical, or other technical capability to perform analysis of energy information to: (1) verify the accuracy of items of energy information submitted to the Director; and (2) insure the coordination and comparability of the energy information in possession of the Office and other Federal agencies." (Public Law 94-385, August 14, 1976); the word "validation" is actually used, without defining it, in Section

57(b) of Public Law 94-383. This legislative beginning of validation increased substantially when the Energy Information Administration was established in October 1977 (Public Law 95-91, Aug. 4, 1977) in the same law that created the Department of Energy. This 1977 legislation also transferred the functions of FEA's Office of Energy Information and Analysis to the Energy Information Administration (EIA). Within EIA, the Office of Energy Information Validation (OEIV) was created to engage "in checking and improving on existing data systems and models and participation in the design and installation of new ones—always with an eye to the quality of the process and of the results it produces." This last quotation is taken from a paper (Moses, 1979) by the first Administrator of EIA.

Since early 1978, Oak Ridge National Laboratory (ORNL) has conducted an energy-data validation program that supported the Office of Energy Information Validation (OEIV) in its mission of determining the quality of EIA's statistical and analytical information. As this paper was being prepared, EIA underwent a reorganization in which OEIV was abolished. After July 1981, the missions and functions of its data validation program, we understand, will be continued in other departments of EIA. As part of the data validation responsibilities of this program, ORNL's support has had two broad aspects:

(1) Conducting independent analyses as well as providing leadership of studies which formulate the concepts (i.e., the methods and materials) of assessment, and which develop or adapt mathematical and statistical techniques for validation processes;

(2) Using studies of energy-data systems to uncover needs for new methodologies as well as to test the efficacy of patterns and procedures —e.g., reviews of information requirements, data-element standardization, user identification, and serial interviewing rules.

Some Highlights of ORNL Research in Statistical Techniques

ORNL's statistical studies in support of energy-data validation has covered a broad spectrum from the very basic (e.g., Chernick, 1981) to the very applied. Of the areas studied, those of greatest applicability to this meeting are: sampling theory, outlier detection, and automatic data editing.

The study of sampling theory is germane if one hopes to draw reliable inferences about a large population by examining a representative subset of that large population. For data validation studies, sampling theory is very helpful in defining and selecting appropriate samples and in interpreting the results obtained from the selected samples. Several aspects of sampling theory methodology applicable to data validation studies have been reviewed by Chernick (1980). In an investigation of multiway stratification (Chernick and Wright, 1980), a simple technique was introduced for systematically allocating (rather than randomly allocating) a sample to the strata formed by two-way stratification. The authors showed that, in several instances, systematic allocation yields a smaller variance than random allocation.

Outlier detection methods are needed to identify, in large volumes of data, records which are unusual and which, therefore, should be examined further. One relatively new method, the influence function method, was used to examine 36 months' data on electricity generated and fuel consumed at 25 power plants (Chernick and Downing, 1980). Outliers were detected for two plants; subsequent contacts with the respondents led to correction of the data. In the same investigation, the subset of cogeneration facilities was identified as "outliers." This points to the power of the influence-function method (a) in detecting outliers in presumably homogeneous data and (b) in its robustness in the presence of nonhomogeneous data. As an analysis of the power-plant characteristics later showed, their data were drawn from two separate, but related populations: power plants which produce steam solely for generating electricity, and plants with cogeneration facilities, which also produce steam for other purposes, e.g., commercial heating. In another recent outlier-research study (Downing and Pierce, 1980), simulated data were compared using six detection methods chosen because they are amenable to large data sets and to relatively short computing times. These six methods were: (1) adjusted discriminant function, (2) discriminant function, (3) first principal component, (4) difference-in-fit statistic, (5) studentized residual, and (6) influence function for estimated correlation. The effectiveness of these methods, when applied to simulated bivariate data with purposely introduced outliers, was found to be in the order listed above. The six methods were also applied to bivariate energy data in which automobile-engine displacement was paired with gasoline consumption, with the result that all six methods proved satisfactory.

The study of automatic (or computerized) data editing is conveniently considered in terms of three processes: (1) identification of erroneous records, (2) localization within the erroneous records, and (3) imputation methods for "adjusting" the errors.

The first process, identification, denotes that a record has failed a specification of constraints or a set of consistency conditions. For energy data, these constraints or conditions are usually based on accounting identities, business patterns, economic principles, and/or physical principles. By way of a very simple illustration, assume that a record has the constraint that the sum of quarterly sales must equal the annual sales total: $Q_1 + Q_2 + Q_3 + Q_4 = A$. Then, the record $A = 39725 = (Q_1, Q_2, Q_3, Q_4) = (11121, 12687, 9422, 9100)$ has failed to obey the constraint and, accordingly, has an "identified" error.

The next process, localization, involves pinpointing the specific data element within a

record which has caused the record to fail the edit. Localization is a more difficult process than identification. As may be seen from the illustration above, there is no additional information in the edit to "localize" the error further. The obvious step would be to trace the data back to its source. Unfortunately, with more complex data, this simple follow-up is ordinarily impractical or impossible owing to time, to cost, or to the unavailability of higher-quality data. Error localization, therefore, becomes an inferential process. (From the illustration above, an inference might be that the datum, Q_4 = 9100, is suspiciously even.) It is beyond the scope of this paper to detail the difficult issues associated with inferential methods of error localization. The interested reader is referred to the papers of G. E. Liepins, especially his most recent review (Liepins, 1981).

Computerized imputation, the third process of automatic data editing, is also a technically difficult research area, but progress has been made especially in understanding the limitations of applying data-matching techniques. More details about these techniques as well as other imputation methods, including discussion of their difficulties and limitations, are also given in Liepins' review (op. cit.).

We now turn our attention to the two major types of studies carried out in the EIA/ORNL data-validation program. These two types are called: (1) reviews of energy-data requirements, (2) system validation studies.

Review of Energy-Data Requirements

The primary objective of this type of study is to define a set of data elements that best serve generally recurring information needs for an energy topic area. This "ideal" set of required data should be current, consistent, and complete, but must be neither excessively costly for the Government to collect nor excessively burdensome for respondents to supply.

A review of energy-data requirements encompasses the following six components:

- o a comprehensive description of the transactions and information flows of the energy topic area. The description includes items of information that are measureable, and of these, which are presently gathered by industry or others in the course of normal business.
- o an analysis of legislation, regulations, treaties, and Presidential proclamations affecting information needs.
- o an analysis and specification of information needs of regulatory users and of other knowledgeable, concerned analysts.
- o a description of the "ideal" data set which justifies which data should be collected, with what accuracy, from which respondents and with what frequency to best serve information needs.
- o an analysis of existing data collection instruments to determine their adequacy in satisfying user needs. This analysis compares currently collected data with the ideal data set; the comparison includes order-of-magnitude assessment of costs and benefits as well as estimates of respondent burden.
- o a set of recommendations regarding the improvement of existing systems and/or the development of new systems.

A review of energy-data requirements is a subtle, intellectually demanding investigation for which the methods of operations research (esp. decision theory) are often quite applicable. EIA has published (DOE/EIA-0276, March 1981) a detailed account of how such an investigation should be conducted and interested readers are referred to this document.

System Validation Studies

These studies assess the accuracy and meaningfulness of the data collected by current systems. A validation study should follow a requirements review since the latter, in addition to providing the framework for understanding data needs, also provides criteria to apply in determining both meaningfulness and the needed accuracy. In most validation studies, however, system-specific user requirements are examined in greater detail, thereby extending the information in the requirements review. Where a requirements review has not been conducted, a first task in validating a data collection system is to conduct a limited requirements review to gain an understanding of user needs and of phenomena underlying the data currently collected.

Although understanding requirements and usage is a very important aspect, the major unifying theme of a validation study is the assessment of accuracy. This assessment is systematized by means of an error model which is both a structural tool and a heuristic device. As a structural tool, the model is used to represent errors which can occur at all stages of building the data base—from design to dissemination. As a heuristic device, the model aids in devising the means for determining the magnitude of error. The error model evolves in the course of a validation study. One may envision three overlapping stages—identification, elaboration, and application. In the first stage, an initial framework is constructed to classify errors in each of the data elements that the system collects; thus, the first stage is concerned with identification of sources of error. In the second stage, the framework is modified and built upon as knowledge of the system increases; hence, the label "elaboration." In the third stage, the error model is applied to devising the means for redetermining data and for assessing (actually estimating) error in all data elements.

In identifying sources of error in energy-data systems, experience has shown that four categories—(1) specification, (2) coverage and selection, (3) respondent, and (4) processing errors—should always be considered.

1. Specification Error. This type of error refers to the difference between the data specified as necessary and the data that is actually collected. It is diagnosed by comparing: (a) the data requirements of the primary users, (b) the definitions and instructions in the collection form, and (c) the capability of respondents to furnish the requested data. Diagnosing specification error may not be very difficult, but estimating its magnitude and its effect on the data element of interest is often formidable; in many instances, rough estimates are all that can be achieved. In reaching an estimate of its magnitude, it is usually unproductive to subdivide specification error into components of variance and bias since its "causes" are due to misjudgements of design or to misunderstanding of the needs and goals of users. The chief reason for seeking specification errors is to eliminate them from future versions of the survey instrument.

2. Coverage and Selection Errors. This category encompasses errors of coverage arising from differences between the frame and the target population. Such errors may arise from: (a) omitted units, (b) duplicated units, e.g., on account of change of name or because a subsidiary of the unit is not recognized, (c) units that are not part of the population, (d) erroneous information on the size of units, (e) information not current to permit location of units. When data are obtained from a sample survey, the quality of the frame may take on added importance; for instance, if a subset of units is under-represented in the frame, then an otherwise good sample design can yield data of poor quality. The category includes other errors introduced in sample selection, e.g., assigning the wrong-sized unit (because of misinformation) to a stratum, or else selecting units subjectively for a probability (random) sample. Also covered in this category are errors due to non-response, including those arising from imputation for non-response.

3. Respondent Error. This general category arises from respondent practices and, accordingly, may require considerable analytic effort to identify all significant sources of error. This part of the error model will probably not be completed until well into the fieldwork step of a typical system validation study. The following six sub-classes of respondent error may be helpful in constructing the initial error framework:

(a) measurement errors—faulty instruments or inaccurate laboratory analyses, biased estimates made when instruments are inoperative or unavailable, systematic error in the measurement (e.g., reporting the actual weight of rain-soaked coal rather than the dry weight requested),

(b) misunderstanding or misinterpreting instructions for filling out survey forms,

(c) carelessness with units, e.g., reporting volume in gallons instead of in barrels,

(d) errors in recording and transcribing data,

(e) errors of omission, e.g., neglecting to report retroactive adjustments to the contract price of fuel delivered to electric-power plants,

(f) misreporting, e.g., reporting the highest cost of fuel delivered to a central storage facility (coal stockpile, tank farm) as the cost of fuel delivered to the power-plant location.

4. Processing Error. This category covers those errors that occur after the respondent population (or sample) has submitted the data. The magnitude of processing errors is estimated from an audit of the processing system which is ordinarily an important task for every system validation. The results from nearly all the processing-system audits conducted in validation studies to date have indicated very small, essentially insignificant, processing errors. But, since the processing of energy data is under EIA's control, processing errors are usually the most directly correctable errors. The category includes:

(a) transcription (e.g., keypunch) errors,

(b) errors due to the loss or misplacement of responses,

(c) errors relating to the treatment of item non-response,

(d) coding error (if responses are coded)—assigning an incorrect code to a survey response,

(e) certain systematic errors—in programming, in tallying for publication purposes, in adjusting the data to achieve correct accounting balances.

The sources of error outlined above help in formulating an initial set of working hypotheses about the nature of errors in the data. Hypotheses of specification errors are developed as the validation analyst interacts with users and suppliers of the data; preliminary hypotheses of respondent and of processing errors are modified after interaction with the system operators; ideas about selection error are advanced as the analyst explores the target universe, the frame, the sample, and the response pattern. In short, an initial framework of error sources, by which accuracy of the data is assessed, is developed at the outset of a validation study and the framework is altered as knowledge of the system accrues.

The error model is refined further as the data base is analyzed for internal consistency and for external comparability. Also, as the validation analyst deepens his study of the operating procedures used to gather and to process the data, his ideas mature regarding processing error. This particular activity sets the stage for auditing a sample of data already in the data base for the purpose of estimating the magnitude of the processing errors.

The field survey of respondents is the chief means by which the analyst verifies accuracy. The essence of accuracy, as measured in the field survey, lies in a comparison between an originally (routinely) reported value for a variable and a redetermined, more accurate

value. The method for obtaining the redetermined value is based on an awareness (embodied in the error model) of sources of error in the originally reported value. A pretest with a pilot sample provides an initial test-bed for determining how correct the hypothesized sources of error are.

In the pretest and in the field survey, the validation analyst must go beyond auditing (or replicating the original measurement) because auditing, per se, will only yield a measure of precision. Precision would be measured simply by retracing the same steps (i.e. by using the same method exactly) for redetermining the originally reported value. The validation analyst should devise an alternative method free (or more free) of systematic error for redetermining a value, the method being more accurate rather than just being more precise. Thus, the error model serves an inventive purpose: to help the validation analyst design an alternative, more accurate procedure for redetermining the variable of interest. (A corollary of the above is that a value whose accuracy has been checked by an alternative procedure is more expensive to obtain than if the value had been merely audited.)

The application of the error model in estimating the accuracy of a variable is a difficult enterprise. There will always be uncertainty about how individual components of error interact with each other. The simplest approach —summing error components algebraically—may not be as correct as adding them by some other means. For applying his error model, the analyst should have an understanding of the relationship between all error components. At the very least, he must possess a good grasp of major components of error and how these components interact.

Perhaps the most important purpose for devising error models is to stimulate thought processes directed at eliminating or reducing errors in the system under study. Specification errors, although difficult to quantify, may be easy to eliminate. The biases that fall into the category of respondent error or of selection error will, no doubt, always be present in a survey. Understanding these biases is the important step in reducing their impact.

We conclude this paper by noting 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. The data bases for domestic (non-fuel) minerals production and reserves are sufficiently similar to those for fossil-fuel production and reserves so as to permit fairly straightforward application of the validation methodology. For contemplated validation study of data bases that serve regulatory purposes (EPA, OSHA, etc.), there are analogous completed studies of data bases that mainly serve the needs of energy regulatory agencies (FERC, ERA). And finally, for determining which data should be collected, the methods developed for conducting Reviews of Data Requirements would seem to be generally applicable to the broad spectrum of government information-gathering activities.

Acknowledgments

The authors gratefully acknowledge the help and encouragement of our sponsors, the Energy Information Administration, especially John Shewmaker, Jerry Eyster, and Charles Smith; the discussion of error sources given in this paper reflects greatly their ideas. A large measure of thanks is due also to ORNL subcontractors (and EIA contractors) who, by doing most of the work in conducting requirements reviews and system validation systems, provide the experiences from which the concepts of data validation evolve. Lastly, we wish to thank our ORNL colleagues: Tommy Wright, Darryl Downing, George Dailey, Edith Halbert, and How Tsao for their helpful suggestions after reviewing this manuscript or its predecessors.

REFERENCES

Chernick, M.R. (1980). "Sampling Theory Methodology Applicable to Data Validation Studies." Oak Ridge National Laboratory Report ORNL/TM-7084.

Chernick, M. R. (1981). "A Limit Theorem for the Maximum of Autoregressive Processes with Uniform Marginal Distributions." The Annals of Probability, 9(1), pp. 145-149.

Chernick, M.R. and Downing, D.J. (1980). "The Influence Function Method Applied to Energy Time Series Data." Proceedings of the 1980 DOE Statistical Symposium, Berkeley, California, CONF-801045, pp. 102-111.

Chernick, M.R. and Wright, T. (1980). "Multi-Way Stratification Problems." Oak Ridge National Laboratory Report ORNL/CSD-58.

Downing, D.J. and Pierce, J.E. (1980). "Simulation Analysis of Some Outlier Detection Methods." Oak Ridge National Laboratory Report ORNL/CSD/TM-143.

DOE/EIA-0276, March 1981. "Guidelines and Procedures for the Conduct of a Review of Data Requirements." Report prepared by Office of Energy Information Validation, Energy Information Administration, U.S. Department of Energy.

Liepins, G.E. (1981). "Towards the Quantitative Assessment of Energy Data." Proceedings of the International Conference on Energy Use Management, Berlin, Germany, October, 1981, in press.

Moses, Lincoln E. (1979). "Early Steps Toward a National Energy Information System." American Statistician, 33(3), p. 97.

Public Law 94-385, Aug. 14, 1976. Quotation is from Section 54 which was added to the Federal Energy Administration Act of 1974.

Public Law 95-91, Aug. 4, 1977, The Department of Energy Organization Act, Section 205.

---

[1]Research sponsored by the Energy Information Administration, U.S. Department of Energy, under contract W-7405-eng-26 with the Union Carbide Corporation.