

## Learning from the Past: Updating Data Quality Efforts

Many years ago as part of detailed reports assessing data quality EIA presented data comparisons to the public, documenting why data series that purport to measure similar concepts differ. Examples of 3 proxy measures for motor gasoline demand are: EIA data on motor gasoline sales, EIA data on product supplied of motor gasoline, and Federal Highway data on sales. The data assessment reports were discontinued in the early 1990's. The Statistics and Method Group (SMG) is now considering a new Web product to display such data comparisons. Because the Web presents an opportunity to present information as we obtain it, we can publish information without waiting for a detailed product to be completed. We are currently working on ideas for this new product and welcome suggestions from the Committee. This paper presents background on how data quality assessments were previously performed and discusses what did and did not work well. It also presents an example of what we are thinking about for the future. Past activities may serve as an impetus for other ideas for Web products and are related to the paper on survey self-assessments.

### The Past and Present

During the 1980's and early 1990's EIA presented information on data quality to the public in reports that became known as "State-of-the-Data" reports. This series of reports covered all major energy areas and described our findings on the strengths and limitations of the data. After we completed a study of each of the major energy areas we changed our focus for the next round of studies. We prepared shorter and less extensive reports for the public and developed in-house reports that contained recommendations to improve the surveys. We also briefly experimented with preparing a Quality Profile for the public, patterned after the one that the Census Bureau prepared for the Survey of Income and Program Participation. This, too, was a comprehensive report.

Ultimately, we<sup>1</sup> transitioned to working with the program offices as a team. The transition corresponded to the change in the Agency's focus from data integrity to process improvement as the agency evolved and data integrity was established. However, some information on data quality was still presented to the public by way of the feature articles that the Petroleum Division prepared. In addition, during EIA's Business re-engineering effort in the mid 1990's, we attempted to develop a summary measure of data quality.

EIA's past data quality efforts can be summarized as activities to 1) describe to the Administrator and to the public what we learned about the strengths and weaknesses of the data (State-of-the-Data reports, Quality Profile, feature articles), 2) improve the data or the process (in-house reports containing recommendations), and 3) measure data quality (the former Business re-engineering team's attempt to develop a summary measure of data quality).

### Informing the Public

State of-the-Data- Reports. State-of-the-Data reports attempted to answer the question of what we knew about the quality of EIA data, what were the strengths and limitations? Initially, the EIA Administrator was a customer as well as users of energy data. The methodology included reviewing information from previous validation studies that focused on all aspects of a data collection. The validation studies were designed during the early days of

---

<sup>1</sup> This is from the point of view of a Statistics and Methods Group (SMG) person who also worked in the predecessor office, the Office of Statistical Standards. The program offices had and continue to have ongoing data quality efforts, albeit reduced due to resource constraints.

EIA, when data integrity was an issue. These studies included a search for deficiencies in the universe list, an audit of company records to determine if they corresponded to what was reported, a check for transcription errors by comparing hardcopy to the automated data file, and many other activities. Reflecting on their extensiveness they were referred to as “cradle-to-grave” examinations.

While the validation studies produced a lot of useful information, they were generally not well-received by either the survey respondents or the survey managers. Furthermore, they were very expensive. They were discontinued when our budget was reduced in the early 1980's. One of the reasons that these studies were not popular with the survey managers was that they felt the results did not tell the whole story with respect to data quality. The audit of the company records, for example, was likely to detect errors, but survey managers pointed out that these errors often did not have much affect on the overall quality of the data as evidenced by the good correspondence of an EIA data series with other related data series.

To put the findings into perspective, the methodology of the State-of-the-Data reports included comparisons of selected data collected by EIA or other organizations. We performed the comparisons at the aggregate level and when data were available, at the respondent level. We found comparisons to be the most useful in situations where data were collected from the same respondents and records could be matched. In these situations we could identify the individual respondents with differing responses and follow up to find out why. In doing so, we gained information about how respondents interpreted our definitions and instructions.

As time went on, comparative series became less plentiful and we no longer had validation studies available. We then turned to outlier detection techniques to look for symptoms of problems in the data. We used a variety of procedures to search for anomalous values in our data series, either at the aggregate or respondent level. As anomalous values could indicate changes in the market or company structure, as well as indicating data problems, we attempted to follow up with the outlying respondents when we detected such values. See the attached paper in the PDF by Nancy Kirkendall and Renee Miller from the 1986 Census Bureau Annual Research Conference for examples of how we used comparisons and outlier detection techniques to assess data quality.

The first round of State-of-the Data reports also included detailed descriptions of how the data series were obtained. We reduced the detail associated with these descriptions for the second round on the advice of our independent expert reviewers who noted that descriptions also appeared in the Explanatory notes section of the publications. Appendix A provides a summary of what each of the State-of-the-Data reports covered.

Quality Profile. The Quality Profile also contained comparisons with other data series, but comparisons were not its focus as in the State-of-the-Data reports. Rather, the Quality Profile presented information on sampling and nonsampling errors, focusing on nonsampling errors. It systematically looked at issues associated with coverage, nonresponse, measurement error, data processing, imputation, and estimation for the Residential Energy Consumption Survey (RECS). It included information on the quality of all of the residential energy consumption surveys conducted from 1978 through 1993 to help users interpret the data for longitudinal analysis.

The Quality Profile for the Residential Energy Consumption Survey

<http://tonto.eia.doe.gov/bookshelf/XsearchResults.asp?fueltype=Consumption&title=&start=50> was written for EIA by Thomas B. Jabine and was very well-received by the statistical community. However, it was not well-publicized and so we did not get requests from users for it. Lack of exposure was also a problem for the State-of-the-Data reports as well, although since they were an ongoing product, we developed some interested users.

Feature Articles and Other Information for the Public. Although the State-of-the-Data reports have been discontinued, the Petroleum division has been presenting information in their publications on comparisons of annual EIA data with independent sources

[Comparisons of Independent Petroleum Supply Statistics \(PDF 79kb\)](#)  
[A Comparison of Selected EIA-782 Data with Other Data Sources \(HTML\)](#)

In addition, under the heading, “Featured topic,” on the Alternative Fuel Web page [http://www.eia.doe.gov/fuelalternate\\_njava.html](http://www.eia.doe.gov/fuelalternate_njava.html)

is a button, “Data Quality on Alternative Fuel Vehicles,” that leads to an article that Howard Bradsher-Fredrick of SMG prepared, “Improving the EIA Survey Data Quality on Alternative-fueled Vehicles Using Cognitive and Other Methods.” This article contains comparisons of related data series, as well as recommendations to improve the surveys, keeping with the more recent emphasis on identifying areas for improvement.

Furthermore, the Explanatory Notes section of some of EIA’s publications contains a fair amount of information on data quality. See, for example, information on the Web page for the Commercial Buildings Energy Consumption Survey.

[http://www.eia.doe.gov/emeu/cbecs/technical\\_information.html](http://www.eia.doe.gov/emeu/cbecs/technical_information.html)

#### In-house Reports on Improving the Data or the Process

When we decided to prepare shorter and less extensive State-of-the-Data reports for the public, we concentrated on developing recommendations to improve the surveys. These recommendations were presented to the program offices and to the Administrator in in-house reports, known as “pre-clearance” reports, since they coincided with the Office of Management and Budget’s clearance cycle. These in-house reports examined the findings from the State-of-the-Data reports to determine whether issues that arose in the data evaluations stemmed from the design of the survey forms. In addition, the pre-clearance reports addressed whether the survey forms were adequate to meet changes in the industry.

We later combined the pre-clearance report with a quality control audit to determine if the quality control procedures that were in place were adequate to control response, nonresponse and processing errors and whether standards on these topics were being followed. Recommendations based on the audit and the review of standards were included in our in-house reports as well. Appendix B provides a list of the in-house reports.

#### Measuring Data Quality—The Quest for a Summary Measure

During our Business re-engineering efforts in the mid-1990’s the issue of developing a summary measure of data quality arose several times. And several times we concluded it could not be done, but we attempted it nevertheless.

The Business Re-engineering team ultimately decided a summary measure would not be workable because it would require much time and judgment to obtain. It is being presented in this paper because there are aspects that may be useful in considering a new product on data quality.

To develop a summary measure, we listed dimensions of data quality:

- sampling error

- measurement error (the difference between the value collected during the survey and the true value. It included both reporting error and specification error)
- coverage
- nonresponse
- methodological consistency (pertained to breaks in the data series and whether the changes and their impact on the data are documented)

These dimensions differ in the ease with which we can quantify them. Sampling error, on the one hand, can be computed directly from the survey data. Methodological consistency, on the other hand, cannot be directly computed. Some other dimensions sound like they could be quantified, such as measurement error. But, EIA does not have information for each survey on an ongoing basis. This is the type of information that we obtained from the validation studies which have been discontinued.

Since we couldn't quantify each dimension we decided to take a qualitative approach. For each survey, the proposal was to gather together the information we had for each of the data quality dimensions. We would then rate each survey on each dimension using a 1 to 5 scale (where 5 is very satisfactory and 1 is very unsatisfactory). To ensure consistency we started to develop guidelines on what represents a "5" versus a "4" and so on. But, it got very complicated quickly. Because of the perceived complications and other issues, the Business Re-engineering team decided not to pursue this procedure. One issue was who would do the ratings. Another concern was that we could not really ensure consistency. Furthermore, there was the perception that a lot of time would be involved in performing the ratings.

The general feeling was that even if we could be precise enough to ensure consistency, we would not be giving the user much more information than is provided in the explanatory notes section of our publications. This comment raised the issue of whether the approach we should take should focus on the descriptive explanatory material, perhaps standardizing it. This idea was never really pursued, but in thinking about a Web product it may have merit. The challenge we would face is that we do not have definitive information for each dimension of data quality and so some judgment would be required in determining what we would present.

What Worked Well and What Didn't. Both the State-of-the-Data reports and the Quality Profile had some strong points. They both provided explanations of why similar series differed. They also provided information for interested users within and outside of EIA on how the data were obtained. In gathering this information for the State-of-the-Data reports, staff members gained an understanding of how particular data series were obtained. Documentation was provided that may have otherwise been lost. This was particularly true for the RECS Quality Profile which was being prepared around the time the survey manager was retiring. The State-of-the-Data reports also provided information for the EIA Administrator that would not have been available during the routine processing of the data. The Administrator was a key customer in the 1980's. By the early 1990's the Administrator's focus was process improvement and working as a team to make the improvements. The detailed assessment reports were, therefore, discontinued.

On the downside, both the State-of-the Data reports and the Quality Profile took a long time to prepare and required considerable in-house and contractor resources. Furthermore, the reports were not widely read. So we

had a lot of input, a lot of output, but were not sure of the outcome of our efforts. In addition, the review process was very lengthy and initially it was often contentious. It became less so over time as we built relationships with survey staff, but many were still concerned that they would be blamed for the limitations of their data and so they were not eager to disclose them.

The feature articles were less resource intensive, but program office staff responsible for them commented that producing them on a tight publication schedule was sometimes challenging because finding the answers to the questions that the comparisons pose requires detective work. And detective work doesn't necessarily follow a schedule.

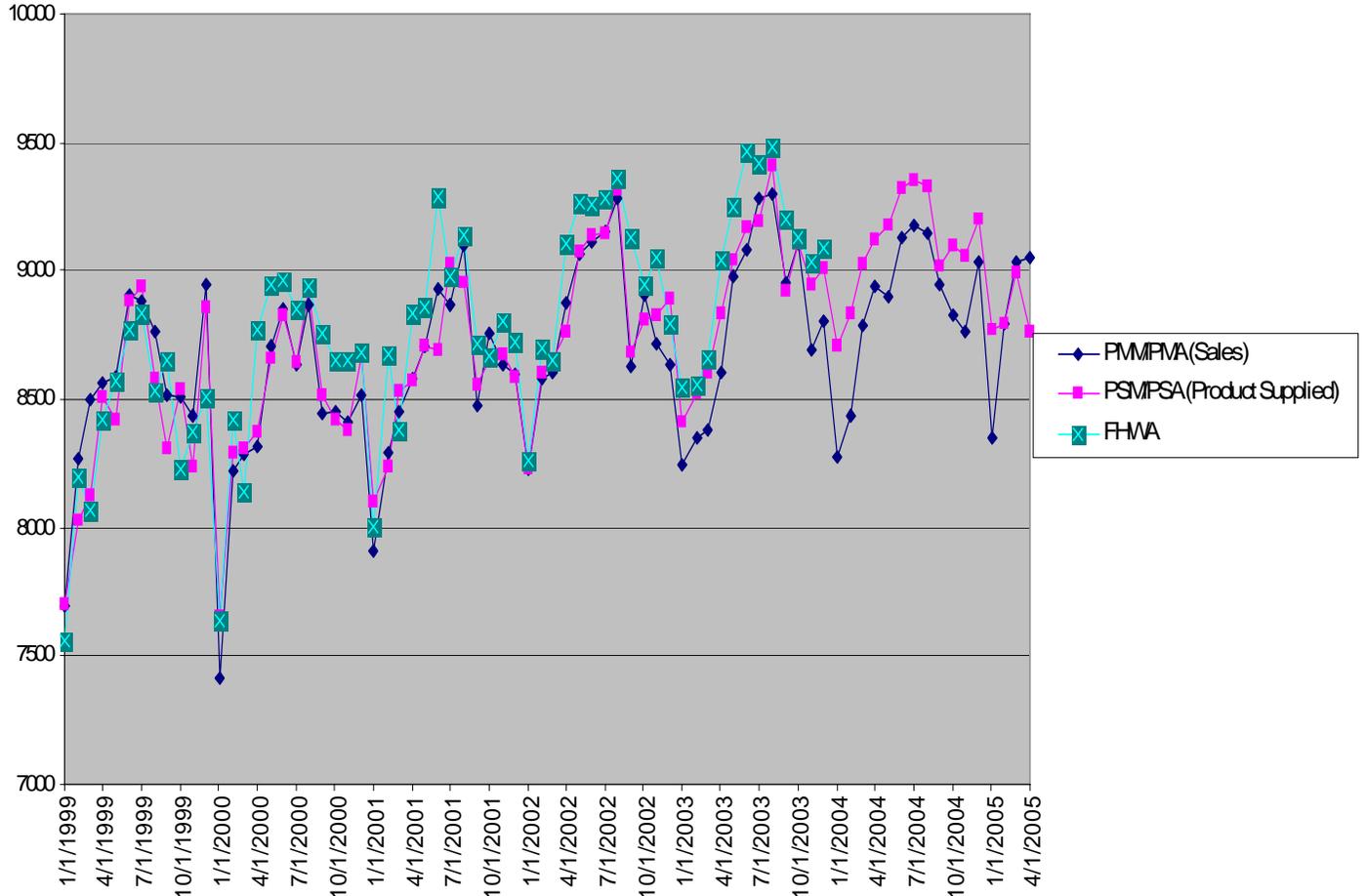
### The Future

To take advantage of what we think are the good points of our previous work, we are considering presenting data comparisons in a question and answer format on the Web. We used a question and answer format in the highlights section of the 1991 Coal State-of-the-Data report. In recently reviewing this report, it seemed that this format made it easy to follow.

An advantage of a Web product is that we could present the information as we obtain it rather than waiting for a detailed product to be completed. Using the principles of writing for the Web, we would begin with general information and then provide links for more details such as to methodology papers or to forms and instructions.

Since information is currently available on motor gasoline comparisons, we present them as an example. EIA presents information on motor gasoline product supplied and motor gasoline sales. In addition there are data available from the Federal Highway Administration. The three data series are considered as proxies for motor gasoline demand. We think that presenting comparison graphs and a series of questions and answers would provide a user-friendly product. Referring to the graph, "Demand comparisons: EIA Sales, EIA Product Supplied, Federal Highway Sales," following are questions we may pose to discuss the comparisons.

Demand Comparisons: EIA Sales, EIA Product Supplied, Federal Highway Sales



1. Both EIA and the Federal Highway Administration present data on motor gasoline sales. How do these figures differ?

We would specify the points at which the data are collected and could provide links to the EIA survey form, EIA-782C, and to its instructions. We could point out that EIA collects data from suppliers who sell to local distributors, local retailers, or end-users. Provisions are made to avoid double counting. The respondents are asked to report gasoline sales volume by category and grade.

The Federal Highway Administration, on the other hand, collects data from the States on State-fuel tax receipts from wholesale distributors. We would need more discussion on how the coverage and measurement points for the EIA and Federal Highway Administration data differ.

2. EIA also presents data on motor gasoline product supplied. Is this the same as demand and sales? What is the relationship to consumption?

We would point out that the concepts are not exactly the same. Sales and product supplied are both proxies for energy demand, which EIA defines as “The requirement for energy as an input to provide products and/or services.” Consumption is defined in terms of use. Differentiating consumption and demand will be challenging because these terms are often used interchangeably at EIA.

Motor gasoline product supplied is computed as net production plus imports minus exports plus stock change. It approximately represents consumption because it measures disappearance of products from primary sources, for example refineries. It is not quite the same as consumption because while it includes stock change at refineries it doesn’t cover secondary or tertiary storage.

3. Do the sales and product supplied series show the same pattern over time?

It appears that they generally do. Deciding how much detail to go into and explaining why the series don’t always move the same way over time will also be a challenge because we may never be able to find out why.

### Questions for the Committee

Based on our experience with data comparisons, it is likely that we will not have answers to all of the questions that the comparisons pose. From the committee’s point of view is there merit in going forward with something to show the users what we currently know? What does the committee think about what we are proposing?

Are there other ideas from our past work or the Committee’s experience that we should consider for a Web product on data quality issues?