IMPUTATION, REVISION, AND SEASONAL ADJUSTMENT

Yvonne M. M. Bishop, Department of Energy

This talk will address the four questions:

What do we mean by imputation, revision and seasonal adjustment?
Why does an agency that publishes statistical series based on periodic surveys need such techniques?
What specific techniques and procedures are we currently using?
What sort of procedures do we anticipate we will be following in the future?

I. Definitions pertinent to EIA

In general usage the word "imputation" has a somewhat negative connotation, but its Latin origin can be translated as "to bring into the reckoning," which is closer to our usage. We have broadened the term to mean any statistical procedure that takes raw data from a subset of the universe and estimates a statistic to describe the complete universe. The subset can arise from: (a) missing values from a census, (b) sample design, (c) incomplete frame, or (d) the deletion of outliers.

In our publications we label with "R" for "revision" any statistic for which a different value has previously been published. We do not distinguish between a correction that occurs because an unanticipated error has been detected, or other revisions that represent improved information and can often be anticipated e.g. filing of late responses.

Seasonal adjustment we interpret as any method that presents time series information in such a way that period to period change can be assessed in the context of the "normal" amount of change for the season of the year, whether this be an index computation, other deseasonalized statistic or a graphical presentation.

II. The need for imputation, revision and seasonal adjustment

In many series, perodic surveys are based on samples and the national estimates must be obtained by weighting, regression or other techniques. In both the surveys based on samples and those where all known members of the universe are solicited for response, non-respondents and late respondents are encountered. To ensure timely publications many deadlines must be met.

1) The end of reporting period
2) Receipt of forms
3) Screening and data entry
4) Non-respondent listing
5) Edit
6) Receipt of revisions and corrections
7) Edit
8) Aggregate and produce tables
9) Produce graphs
10) Write interpretation
11) Inhouse review
12) To printer

Many publications present results from several different surveys, and an important item is the "energy balance" which provides summary information on the national supply of energy products during the preceding period. To compute such balances, current information is needed from each of the different surveys. Thus estimates must be made for late respondents or for values that the edit detects as being unlikely and for which there has not been time to obtain a correction. These events occur in spite of vigorous non-response follow-up by telephone.

Another form of missing data arises when statistical estimates are made from a series designed for regulatory purposes, and the regulations do not apply to all members of the universe because, for example, they exclude small companies.

Whenever an imputation procedure is used for missing respondents, and the full information is subsequently obtained from all targeted respondents, a revised statistic is published. In other instances the respondents who are prompt submit estimates that they subsequently revise, this is particularly true for quantities of crude oil and products imported. These revisions may be submitted several months after the event occurred. Thus for a monthly publication that presents the last month and several preceding months, if emphasis is placed on providing the most accurate information available the statistics will change in each issue. This is clearly undesirable, and it is necessary to develop imputation procedures such that subsequent changes in the first three digits at the national level are rare, and only occur at scheduled intervals.

Occasionally a correction is needed because an error is detected after publication. This may be due to misreporting, e.g. reporting in barrels instead of thousands of barrels, or due to a processing error. Such errors require sophisticated edit techniques so that they are detected before publication. A further problem that occasionally arises is double-counting, where two respondents both report on the same physical quantity, e.g. when a business is changing hands both buyer and seller report. These problems are more difficult to detect, and require bench-mark comparisons.

The annual consumption of energy at the national level shows a strong seasonal pattern. This pattern occurs in many data series, although it is more pronounced in consumption for heating than consumption for other purposes. It is thus necessary for month by month data to be presented in such a manner that the reader observing a change is able to immediately perceive whether this change is usual for the time of year.

III. Examples of procedures in current use

Much effort is currently being devoted to developing sophisticated edits, and appropriate imputation techniques, and methods of presenting seasonal data. The following examples show the approaches being taken:

1) Adjusting for incomplete frames in crude oil production
2) Editing electric power data
3) Using small weekly samples of oil stocks in conjunction with monthly census
4) Adjusting for non-respondent solar manufacturers

Other approaches to editing and the use of graphical methods to relate current oil stocks to seasonal patterns are described elsewhere (see references).

Example 1: Domestic Crude Oil Production

The Problem: If we look at the monthly statistics for domestic crude oil production as published in June 1979, we find an apparent drop in January, February and March and then a rise in April (Table 1). This pattern was repeated in successive issues and was the result of using different data sources for months of varied recency.

Table 1. Domestic Crude Oil Production in thousands of barrels per day (MB/day) (June 1979, Monthly Energy Review)

Date	Production	Data Source
1978 Oct.	8,830	State Records
Nov.	8,728	" "
Dec.	8,651	" "
1979 Jan.	8,346	First Purchase Series
Feb.	8,286	" " "
Mar.	8,369	" " "
April	8,618	American Petroleum Institute

The oldest statistics were based on information obtained from the States, the next three months from a regulatory data series, and the most recent months from the American Petroleum Institute publications. The State data are not available for 3 - 4 months. The regulatory series were collected in conjunction with the entitlements program and required that all persons purchasing more than a certain amount of crude oil report their purchase within 30 days - with the option to submit corrections up to 90 days. Compared with past State data, the past First Purchaser values were consistently lower. Nationally the discrepancy averaged 208 thousand barrels a day over the period examined. The discrepancy occurred because oil consumed on the lease site and small purchases were not required to be reported; an incomplete frame problem.

The Solution: To remedy this wide discrepancy between series a three-step strategy has been developed. The preliminary step was to use the First Purchase data, incorporate ratio-estimates for non-respondents and add 208 thousand barrels a day to the total.

The second solution was to develop Box-Jenkins forecasts, individually for the large States and for the remainder jointly, and use these in conjunction with the early returns from States who were able to report rapidly - Alaska information in particular is obtained by telephone. These forecasts were usually close (Table 2).

Table 2. Example of Box-Jenkins forecasts for domestic crude oil production. Forecasts in MB/day are for January 1980 and are based on data from Jan. 1973 through Dec. 1979

State	Standard Error of Forecast	Forecast	Final Value
California	11.31	968	947
Louisiana	66.03	1359	1354
Texas	23.64	2708	2692
Other States	59.29	2055	2019
TOTAL lower 48 States	92.53	7108	7014

The final solution was to change the reporting requirements for the First Purchase System to include small purchases and lease use, and return to this system as soon as it was established.

Example 2: Edit Procedures Electric Utilities

The Problem: Data are collected monthly from 861 electric utilities on 3059 plants. A plant may have several units, each with different capacities and consuming different fuels, but the information received is on a plant basis. The variables are generation in megawatt hours, consumption and stocks of coal in short tons, oil in barrels, and gas in cubic feet. These must be broken down by unit to compare with capacity, and converted to kilowatt-hour equivalents to compute efficiency, (defined as the ratio of electricity generated to fuel consumed.)

A number of manual checks had customarily been used to verify the raw data. These hand computations were time consuming. They have been evaluated by computerizing and determining which yield best results. By analyzing past data improved tests have been implemented.

The Solution: A battery of computer edits has been devised for each of the variables of interest: generation, consumption, stock level and efficiency. This battery is applied to the time series for each plant and, for some variables, to the first order differences between months.

The elements of the battery that are appropriate for a particular variable are chosen by a branching process. This branching process relies on prior analysis of the usual patterns as exhibiting white noise, linear trend, and/or seasonal pattern and the relevant parameters are stored.

The battery currently has three basic approaches and specific examples are:

1) Control chart approach - z-test
   The historic mean and variance are stored. The new data point is checked to fall within a preset number of standardized deviations of the historic mean, currently within 2.5 units.
2) Forecasting approach - exponential lag tests
   An auto-regressive forecast is made that takes a weighted average of the two preceding months, appropriately adjusted for seasonality. The new point is compared with the forecast and must be close to it for acceptance.
3) Multivariate approach - influence function tests
   The particular influence function test used is the weighted difference between two correlation coefficients between variables, where one coefficient is computed from past data only and the other includes the new data.

An important attribute of the battery of tests is that, once each plant has been analyzed, the parameters for each test are stored. If a new data point passes the test, then the parameters are automatically updated. Currently work is underway to evaluate each component of the battery in terms of false positives and false negatives, and adjust the fail criteria to detect all large errors while minimizing the number of followup procedures needed. In a test batch of 667 plants the battery detected 158 errors in 22.91 seconds execution time. The influence function test alone isolated 16 data points, all of which were found to be computational errors.

In addition it is planned to classify the potential errors generated by each test, so that in emergency conditions a particular fail criteria can be overridden e.g. - during a coal strike the pattern of consumption of coal would be reduced compared with normal. Bypassing is only possible by means of special passwords, known only to the data base administrator.

Example 3: Relating a sample to a census - weekly stocks

The Problem: Until January 1980 we were publishing statistics from the American Petroleum Institute in our Weekly Status Report. Last year we began to collect weekly stock data from a subset of the refineries (170 out of 315) and bulk terminals (71 out of 155) and from the 57 pipelines. These data are received on Monday for the preceding week and printed by Friday. Thus they are more timely than the monthly statistics collected from the universe, where the time-lag is about two months. We wished to obtain weekly regional estimates, divided into Petroleum Administration for Defense (PAD) districts. Although the samples were selected to cover the major companies, within a district the proportion of total stocks represented varied considerably. It was necessary to devise a weighting strategy that would give a representation of the total stocks in each district. Fixed weights would be unsatisfactory if the pattern of behavior differed between large and small companies.

The Solution: The most recent monthly data are used on an establishment basis. The weekly data are summed to give a total w_{ijk} for each district i, product j and type of company, k. The corresponding sum for the identical establishments is obtained for the most recent month, m_{ijk}, and the sum for all establishments for the month is available, M_{ijk}. Then the weekly estimate for all companies of this type is derived as a ratio estimate

$$W_{ijk} = w_{ijk} \frac{M_{ijk}}{m_{ijk}} .$$

The PAD district estimate for district i and stocks of product j is

$$W_{ij+} = \sum_{k} W_{ijk} .$$

Prior to implementing this procedure, weekly values specific to both product and type of company were compared graphically with monthly values both by region and nationally. In most instances there was good agreement after the weekly data had been edited. On one occasion a sudden departure of the weekly estimates pointed up an error in the monthly data. This was found to be a single report where units were given in barrels instead of thousands of barrels. This example pointed out the need for computerized editing of both monthly and weekly data and analyses are underway to determine acceptance regions for each new entry.

Example 4: Adjusting for non-response - solar collectors

The Problem: Information is obtained from manufacturers and importers of solar collectors every six months. Interest lies in whether this industry is growing, which types of collectors are being produced, for what purpose and in which parts of the country. For the last six months of 1978, there were 54 nonrespondents of the 250 who had responded in the previous six-month period. To publish results based only on responders would indicate a decline in square feet of production instead of an increase.

The Solution: The strategy used to adjust for non-response was a ratio-estimate, specific for the type of collector. A manufacturer often produces more than one type of collector and the change over time in quantity produced varies considerably by type of collector. Of the 54 non-respondents at time of publication all except two subsequently submitted reports. The unadjusted total was 3,654 thousand square feet, the adjusted total 4,584. With only two non-respondents the total was 4,109. (See Table 3) Thus the estimate was 475 thousand square feet above the reported total with two non-respondents. These two non-respondents accounted for almost half of this amount in the preceding period, thus unless they doubled their production the overall estimate tended to be high. However, when we look at the individual categories it is apparent that for some the estimates were low, e.g. for low temperature nonmetallic collectors the later raw data were larger then the estimate.

Table 3. Example of adjusting for non-response: Solar collector manufacturers and importers last 6 months 1978 - production measured in thousands square feet

Type of Collector	Early Raw Data	Estimates	Later Raw Data
LT non-metallic	1,834	1,856	1,899
LT metallic	151	421	339
MT air	360	446	380
MT liquid	1,204	1,727	1,384
Special concentrator	73	75	78
Special Evac. tube	27	53	27
Other	5	5	5
Total	3,654	4,584	4,109
Non-respondents	54	NA	2

LT = low temperature
MT = Medium temperature

This imputation procedure could probably be improved. Possible strategies include: estimates based on more than one preceding time period; estimates based on disaggregation by another variable; or estimates based on multiple time periods and multiple variables. Fortunately the non-response rate improved in subsequent periods but when resources are available we hope to improve on the imputation procedures for this series.

IV. What are our future plans?

In order to produce timely, credible statistics we have a great deal more work to do in the areas of developing series-specific edit procedures. The pattern-recognition techniques with self-adjusting algorithms currently being used for the electric generation series hold promise of wider applicability for the detection of errors.

The improvement of timely response rates requires a many-pronged approach. One of the aspects will be development of universe frames with updating mechanisms. With such frames in place, strategies for periodic bench marks combined with small rotating samples will reduce respondent burden and may help with responses. It is however unlikely that the problem of missing values will entirely disappear, and the need to develop imputation procedures will remain. As many of our series vary over time it is likely that time-series analysis will be invaluable in developing the imputation procedures, as well as aiding in the presentation of our statistics.

I believe that our use of time series analysis and pattern recognition approaches will provide the desired results - a credible information system that will guide national energy policy and keep the public informed of the current energy situation.

References

Yahia Z. Ahmed "Use of First Purchase Data as Surrogate for Production Data" (mss)

Eugene M. Burns, "Proceedings for the Detection of Outliers in Weekly Time Series" Proceedings of the Business and Economics Statistics Section, 1980

Edward Hill, Jr. "Data Base Quality Assurance" (mss)

Nancy J. Kirkendall "Seasonal Analysis and Forecasting of Petroleum Inventory Time Series" Proceedings of the Business and Economics Statistics Section, 1980

Jerald S. Schindler "An evaluation of the Crude Oil Production Estimation System" (mss)

Manuscripts can be obtained from the authors who are all in Energy Data Operations, Energy Information Administration.