EDITING AND IMPUTATION FOR THE EIA WEEKLY PETROLEUM SURVEYS

Eugene M. Burns, Energy Information Administration*

1. INTRODUCTION

This paper discusses the major statistical features of the editing and imputation system used for the Energy Information Administration (EIA) weekly petroleum surveys. The objectives of this discussion are (1) to present the main ideas behind the overall design of this system and (2) to show how various statistical problems which arose in implementing this design were analyzed and resolved. Additional details on these surveys can be found in The Weekly Petroleum Status Report: Technical Background (Energy Information Administration 1983, Ch. 2).

The six weekly petroleum surveys collect data on petroleum refinery operations, and on imports and inventories of crude oil and selected petroleum products. Reports of activity for the week ending 7:00 am Friday are due at the EIA by 5:00 pm on the following Monday. Approximately 450 survey forms, submitted by a sample of the larger U.S. oil companies, must be processed by noon Wednesday to allow publication of the Weekly Petroleum Status Report on Thursday morning. Given this tight production schedule, automated editing procedures were needed to screen the incoming survey data, and imputation procedures were required to impute both for respondents missing the reporting deadline and for reported values rejected during the automated editing.

The typical weekly survey form is a one-page grid, in which the rows are petroleum products and the columns are geographic regions. Respondents are required to enter their volumes for the week into the appropriate cells of the forms. These data are all quantitative, and present few opportunities to verify the internal consistency of the forms, except for checking that the U.S. total line is indeed the sum of its parts. However, the same units report to the survey each week, thereby providing an excellent opportunity to match incoming data against company-specific historical reporting patterns. If the reporting patterns of the companies could be adequately summarized, then the summary statistics could form the basis for both data editing and imputation. Assuming that satisfactory summary statistics for the historical data could be found, the next problem was deciding how to use these summary statistics to develop an operational editing and imputation system.

Thus, two main statistical issues emerged in the design of the editing and imputation system for the weekly petroleum surveys:

1. a data modeling issue, i.e., how best to use the available historical weekly data to build summary statistics of company reporting patterns; and
2. a data comparison problem, i.e., how to use the summary statistics to recognize improbable reports and to determine what to impute for nonresponse or faulty data.

The rest of this paper will discuss these two issues.

2. MODELING THE WEEKLY DATA

One notable feature of completed weekly data collection forms is that most cells remain blank. For example, only a few bulk terminal companies have storage facilities in all regions of the United States, and few report all products surveyed. As a result, most cells on a completed bulk terminal form are empty.

However, companies tend to report the same products in the same districts from week to week. Figure 1 contains several typical frequency distributions of the proportion of weekly nonzero responses over the years 1981 and 1982. These distributions are strongly bimodal. It therefore seemed reasonable to start the modeling process by predicting whether or not an item would be reported in a given week, i.e., by predicting which cells of the forms would be filled. The usefulness of this idea was borne out by an earlier EIA study of outlier detection procedures (Burns 1980), which showed that the various procedures tested were more effective when a distinction was made between zero and nonzero reports. A good predictor of the incidence of nonzero reports could serve both in editing to detect whether data were entered in the wrong cell of the form, and in imputation to determine when a nonzero quantity need be imputed.

After the prediction of a nonzero report, the next logical step becomes the prediction of the magnitude of the report. This prediction could be based either on all reports (including zeroes), or just on the nonzero reports. In other words, the choice is between an unconditional prediction and a prediction conditional on a prior prediction of a nonzero report. For many series, there would be no difference between the two options, since nearly all reports are nonzero. However, in series which show both zero and nonzero reports, the modeling of nonzero reports seemed preferable. For instance, if a refinery shuts down for a period of time, it will submit zeroes for inputs and production during these weeks. When the refinery resumes operations, it will probably resume reporting at about the pre-shutdown level. In another case, importers may not import every week, but the volumes reported when there is an import will often be similar. If the predicted value were based on both zero and nonzero imports, then the frequency and the volume would be confounded. As a result of the above considerations, the

*The author wishes to thank James Diehl and T.C. Swann, both in the Energy Information Administration, for their comments on earlier drafts of this paper.

data modeling issue was separated into two sub-issues: (1) modeling the magnitude of the nonzero reports and (2) modeling the incidence of the nonzero reports.

Modeling Magnitude. In addition to the requirement that the nonzero reports be modeled adequately, the modeling technique had to satisfy three other requirements if it were to be implemented: (1) the technique had to be (relatively) easy to program, (2) the technique could not require excessive amounts of computer resources (time or space), and (3) the amount of intervention required by skilled technical personnel had to be minimized. Several alternatives were considered and tested (see Burns 1980), but the effectiveness of the procedures was found to vary by product. The eventual choice for a procedure to model the nonzero reports was exponential smoothing (Granger and Newbold 1977, pp.163-176). The equation for exponential smoothing is

$$\bar{Y}(t) = \underline{a}\, y(t) + (1 - \underline{a})\, \bar{Y}(t-1),$$

where $\bar{Y}(t)$ is the exponentially smoothed mean updated with data through time t, y(t) is the value observed in time t, and $\underline{a}$ is the smoothing constant, which can take values between zero and one. Exponential smoothing was chosen to model the nonzero reports for two reasons. First, although other models may be optimal for particular series, exponential smoothing usually gives reasonable short term forecasts over a wide range of applications. By setting the smoothing constant close to one, more weight is given to the most recent observation, so that series for which the best forecast is the last observed value can be modeled. On the other hand, as the constant approaches zero, more weight is given to the historical data, as would be desireable if the best forecast for the data were the historical mean. Over the short run (from week to week), seasonal and trend effects are negligible. Thus, the exponentially smoothed mean is able to mimic a wide range of models over the short run.

Second, exponentially smoothed means are very simple to update. This reason is important in an automated data quality control system which requires updated forecasts for several thousand series on a weekly basis. An exponentially smoothed mean of nonzero reports is calculated for each cell of each company's form, and is updated weekly using the above equation.

Since an ARIMA(0,1,1) model (i.e., a first order moving average of first differences) with parameter theta is equivalent to an exponential smoothing model with parameter $\underline{a}$ equal to one minus theta, ARIMA modeling techniques could be applied to the problem of estimating the exponential smoothing constants. Using standard software (SAS's PROCEDURE ARIMA), models were fit to the weekly U.S. totals (over all companies) for each form and product over a period of ninety weeks.

Data editing requires the ability to form ranges of acceptable values, not just point estimates. To provide these ranges, another series was also modeled, the mean absolute deviations from the forecast of nonzero reports. These mean absolute deviations are the absolute values of the one-step forecast errors. Exponential smoothing was also chosen to model the deviation series. The parameter values were obtained by fitting an ARIMA(0,1,1) model to the absolute values of the residuals from the fit of the weekly U.S. totals for each form and product. The exponentially smoothed mean absolute deviation is also calculated and updated weekly for each cell of each company's form.

Modeling Incidence. It was less obvious how to model the incidence of nonzero reports. For the majority of data series, especially the inventory series, most companies either always report nonzero values or always report zeroes. However, for some data series, especially imports, the problem is not so trivial.

Although a binary pattern of zero and nonzero responses is observed for any series, it was assumed that there was an underlying probability of a nonzero response. If this probability could be estimated, then it could form the basis for predictions of zero or nonzero response. This prediction could be based on a statistic like the proportion of nonzero responses over some time period (as presented in Figure 1). However, such a statistic would not be adaptive to short-term changes in company reporting patterns, such as caused by refinery shutdowns. The modeling approach chosen was to calculate an exponentially smoothed frequency of nonzero reports. For each cell of each form, a binary variable is defined to be 1 if the report is nonzero and 0 if the report is zero. These binary data are then used in conjunction with the exponential smoothing equation given above.

The decision to use an exponentially smoothed quantity to estimate the probability of a nonzero response raised two further issues. First, a method needed to be chosen for modeling the binary time series. Standard ARIMA model-fitting techniques could be employed. In the literature on clipped time series, Kedem (1980a,1980b) presents a method for fitting autoregressive models to binary time series formed from an observed continuous series. He advocates clipping as a computational shortcut. If it is possible to model continuous series satisfactorily after reducing it to binary form, then it should also be possible to use binary data to represent a hypothetical underlying continuous series. Alternatively, the exponential smoothing constant could be fit directly by performing a grid search to find the best (i.e., minimum mean square error) value.

The second issue revolved around finding the best way of using the weekly data to fit the model. For the exponentially smoothed mean of nonzero reports, the weekly data were aggregated to form U.S. totals for each form and product. For binary data, such aggregation would produce weekly counts of nonzero responses. More appropriately, the respondent-level data could be used directly. The practice of using respondent-level data to estimate parameters of time series models has been suggested in articles by Scott et al. (1977) and Smith (1978).

Taking the above two considerations into account, the following approach was adopted to fit the binary time series:

1. For each form and product, all the non-trivial respondent-level time series (i.e., series in which the proportion nonzero was greater than zero and less than one) were identified.

2. Using nine values for the smoothing constant (.1 to .9 by .1), each individual series was passed through the smoothing expression, and the squared two-step forecast errors were computed.

3. For each value of the smoothing constant, the mean square error was calculated as the mean of the squared two-step forecast errors over all non-trivial series for a given product.

4. To refine the parameter estimate, four additional values were evaluated in the vicinity of the best estimates from step 3.

The final parameter estimate was taken to be the value associated with the minimum mean square error at step 4. The chosen parameter values were then tested on some of the original primary time series. Two of these test runs are shown in Figure 2. Note that the resulting mean behaves as one would intuitively expect an underlying nonzero response probability to behave given the observed data.

As for the exponentially smoothed mean and mean absolute deviation, a value of the exponentially smoothed frequency of a nonzero response is maintained for each cell of each form.

3. EDITING AND IMPUTATION

The preceding section showed how the data modeling issue was resolved by developing profiles, based on historical data, for each cell of each form. The profiles consist of three summary statistics: an exponentially smoothed mean, an exponentially smoothed mean absolute deviation, and an exponentially smoothed frequency of a nonzero response. To use these summary statistics in editing and imputation, cutoff limits were developed to define tolerances for editing and to determine what to impute. The development of these cutoff limits is the topic of this section.

The weekly data processing system is depicted schematically in Figure 3. At time of system initialization, the historical data were used to create the three summary statistics. Incoming data are compared with the summary statistics during editing, and the data (with accompanying edit flags) are passed on to the transaction file. When estimates are required, both the transaction file and the summary statistics file are employed. If a reported data element is unacceptable or a company has not responded, then values based on the summary statistics are used.

As implemented in the weekly processing system, the three exponentially smoothed values are not updated with the most recent data. Reported data for a given week are not used to update the three summary statistics until two weeks after the end of the reference period. By waiting two weeks, additional time is allowed for the receipt of late or revised reports, and for the resolution of problems. At the end of the second week, the data are as clean as possible, since resubmissions are not entered after the second week.

Using two-week-old data to update the means does not affect the means of nonzero reports because the two-step forecast from an exponential smoothing model is the same as the one-step forecast. The mean absolute deviation of a two-step forecast is larger than that of a one-step forecast. However, either could serve as the dispersion measure for data editing. The values of the smoothing constants for the frequency mean were chosen on the basis of two-step forecast errors.

Editing. Each form is edited twice, once during on-line data entry and once in batch mode. The on-line edits are more tolerant, and cause a critical flag to be raised if failed. Failure of the more stringent batch edits results in a warning flag. Items with critical flags are replaced with imputed values during estimation unless verified with the respondents.

Three types of edits are performed on the weekly data: (1) consistency checks (not discussed in this paper), followed by (2) frequency checks, using the exponentially smoothed frequency of a nonzero response, and finally, (3) the outlier checks, using both the exponentially smoothed mean of the nonzero responses and the exponentially smoothed mean absolute deviation.

The frequency check requires a prediction as to whether each cell will be zero or not. Actually, a three-way prediction is made. If the value of the frequency mean is above a certain limit, then a nonzero report is expected. If the frequency mean is below a certain limit, then a zero report is expected. If the value of the frequency mean lies between the lower and upper limits, then either a zero or nonzero report is acceptable.

Thus, to make the frequency prediction two limits are required. These limits were determined for each form and product. The upper limit was chosen by examining the empirical distribution of the frequency mean for items actually reported as zero. The limit for on-line edits was chosen with the goal that fewer than 1 percent of the values would have been rejected, and the batch edit limit was chosen so that fewer than 5 percent would have been rejected. The lower limits were chosen in a similar fashion by examining the distribution of the frequency mean for nonzero reports.

The outlier test is only performed on nonzero items passing the frequency check. A reported item is flagged as an outlier if it varies from the nonzero mean by more than a certain number of mean absolute deviations, and, in addition varies by more than a certain absolute amount (a "fuzz" value). For the on-line outlier checks, the fuzz cutoff was set at the median value. The batch fuzz cutoff was set at the 25th percentile. The number of acceptable deviations was set by examining the distribution of standardized deviations from the mean exceeding the fuzz level for actually reported data, and setting rejection rate targets at 1 percent for on-line edits and 5 percent for batch edits.

Imputation. Imputation is performed for nonresponse and for each data element with unverified critical flags. The determination of an imputed value is a two-step process: (1) predict whether a zero or nonzero value should have been reported and (2) if nonzero, predict a value. For imputation, unlike editing, either a zero or nonzero value must be predicted, and so only one cutoff is required. If the frequency mean for an item requiring imputation is below the cutoff value, then a zero is imputed; otherwise, the exponentially smoothed mean nonzero report becomes the imputed value.

Imputation limits were set at .40, .50, and .60, and were tested by imputing for the entire sample. The weekly totals obtained by full-sample imputation were then compared with the totals of the actual submissions. For most stocks series, there was very little difference between the three imputation limits. However, some series, particularly imports series, were sensitive to the choice of cutoff limit, and for these series the cutoff was revised to the value which gave the closest approximation to the actual data.

4. PRELIMINARY EVALUATION

The weekly processing system incorporating these edit and imputation procedures became operational in January 1983. In the months since then, there have been several adjustments of updating parameters and cutoff limits, as well as a major revision to the weekly sample. Updating parameters for the frequency mean, fit by the methods described in this paper, were incorporated in August.

Due to these changes in the system, definitive evaluation of the editing and imputation system is not yet possible. However, preliminary results indicate that the system is working well. No serious data errors have occured since the system became operational. In March, operating personnel noted that there seemed to be an excessive number of critical outlier flags being raised. The problem was traced back to a lack of sufficient digits in the calculation of the mean absolute deviations. Truncation was causing deviations to approach zero for smaller items.

The imputation procedures have also been successful. Publication of the Weekly Petroleum Status Report was advanced from Fridays to Thursdays in July, due both to company cooperation in reporting earlier and to the performance of the imputation procedures for nonrespondents.

The editing and imputation system will continue to be monitored and evaluated, and a more definitive evalution will be completed early next year.

5. REFERENCES

Burns, Eugene M. (1980). "Procedures for the Detection of Outliers in Weekly Time Series." Proceedings of the Business and Economic Statistics Section of the American Statistical Association.

Energy Information Administration, U.S. Department of Energy (1983). The Weekly Petroleum Status Report: Technical Background. Prepared by Eugene M. Burns. DOE/EIA-0414, Washington, D.C.

Granger, C.W.J. and P. Newbold (1977). Forecasting Economic Time Series. New York: Academic Press.

Kedem, Benjamin (1980a). Binary Time Series. New York: Marcel Dekker.

_____ (1980b). "Estimation of the Parameters in Stationary Autoregressive Processes After Hard Limiting." Journal of the American Statistical Association, 75:146-153.

Scott, A.J., T.M.F. Smith, and R.G. Jones (1977). "The Application of Time Series Methods to the Analysis of Repeated Surveys." International Statistical Review, 45:13-28.

Smith, T.M.F. (1978). "Principles and Problems in the Analysis of Repeated Surveys." In N.K. Namboodiri (Ed.), Survey Sampling and Measurement. New York: Academic Press.

Figure 1. Distribution of Proportion of Nonzero Weekly Reports, 1981-1982

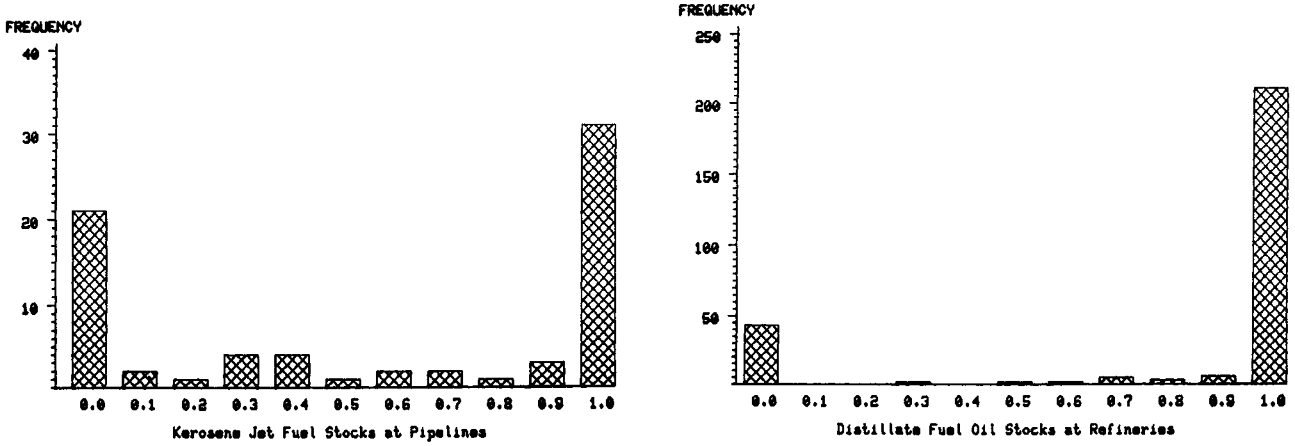


Figure 2. Binary Series and Estimated Individual Reporting Units Frequency Mean for Selected Individual Reporting Units

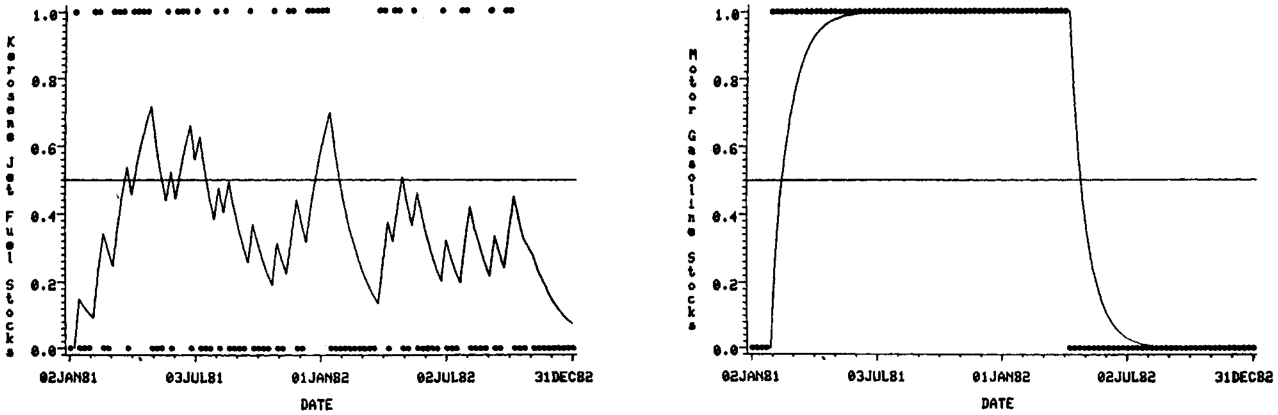
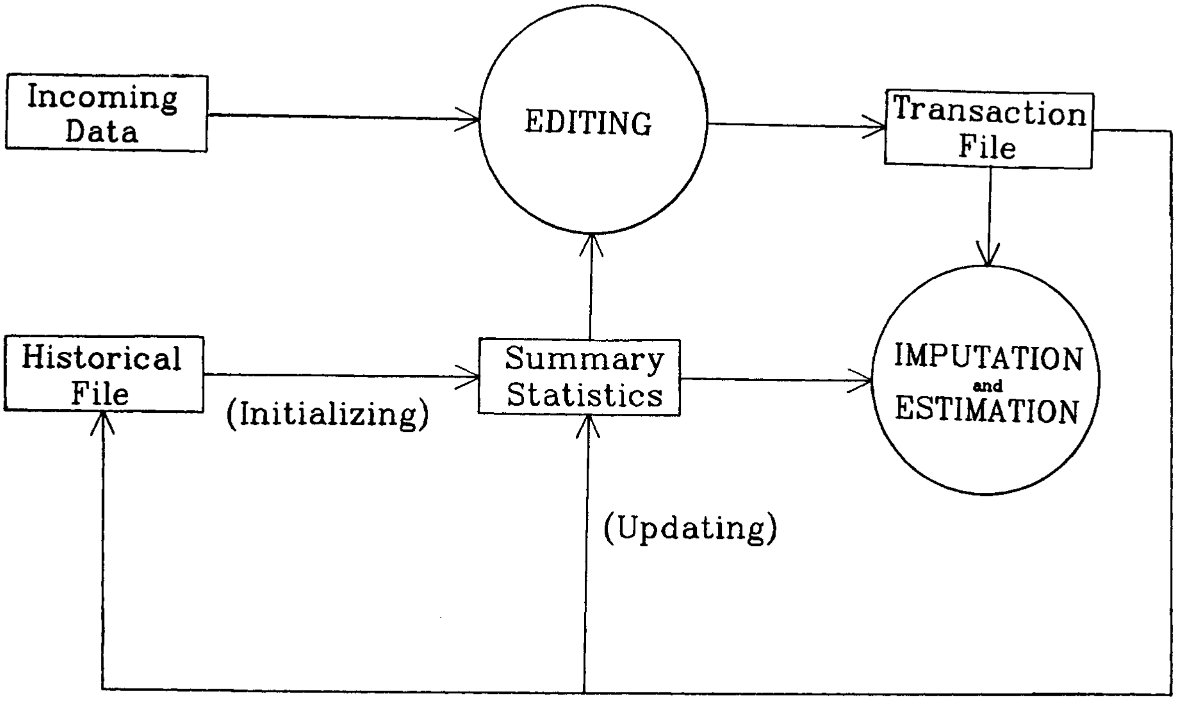


Figure 3. Flow of Data Through the Weekly Survey Processing System


Incoming Data
EDITING
Transaction File
Historical File
(Initializing)
Summary Statistics
IMPUTATION and ESTIMATION
(Updating)