An Improved Imputation Methodology Derived through Regression Trees

Pedro J. Saavedra[1], Paula Mason[2], Benita O'Colmain[1] and Jeffrey Foarde[1]

[1]ICF Macro, 11875 Beltsville Drive, Calverton, MD 20705
[2]Energy Information Administration, DOE, 1000 Independence Ave., SW, Washington, DC 20585

Abstract

The EIA collects monthly information on the balance between supply and disposition of crude oil and petroleum products through a family of surveys. The process requires all imputed values to be available before responses are received, but uses values for only select cells. Previous analysis led to the recommendation for some surveys to implement an imputation method using historical values obtained through exponential smoothing and trend adjustments from a weekly survey. One survey was particularly difficult to resolve because of more extensive dimensions of the survey, the many cases of zero values, and fewer comparable cells in the weekly survey for trend adjustment. To group cells, a regression tree method (CART) was used to obtain groups for which the same smoothing coefficient could be used. Simulation analyses were then conducted to identify an optimal coefficient for each group.

Key Words: Petroleum, CART, trend adjustments, exponential smoothing.

1. Background

1.1 Imputation in the Monthly Petroleum Supply Reporting System (MPSRS)

The MPSRS consists of nine population monthly surveys, six of which have a corresponding weekly survey reporting for a sub-sample of the companies and products. The surveys report volumes for various products and supply types. In some surveys the reporting unit is the company, and in others it is the site. These surveys are:

- EIA-810 Monthly Refinery Report
- EIA-811 Monthly Bulk Terminal Report
- EIA-812 Monthly Product Pipeline Report
- EIA-813 Monthly Crude Oil Report
- EIA-814 Monthly Imports Report
- EIA-815 Monthly Terminal Blenders Report
- EIA-816 Monthly Natural Gas Liquids Report
- EIA-817 Monthly Tanker and Barge Report
- EIA-819 Monthly Oxygenate Report

This paper is concerned primarily with two of these surveys: the EIA-810 and the EIA-812. The MPSRS requires an imputation procedure which can be implemented as each reporting form comes in. That means that the imputation cannot wait for all the forms to come in to arrive at the imputed values. This then, means relying primarily on the historical data reported by the company for that product, supply type and Petroleum Administration for Defense District (PADD) across the months. However, if one obtains a historical value for every cell, there is still a possibility of obtaining adjusted trends.

Each of the monthly surveys has a corresponding weekly survey which collects data for a period two to three months ahead of the monthly. Hence, on can obtain ratios for various combinations of products, supply types and PADDs and use the ratios of these aggregated volumes for one month to the next in order to adjust the estimates for trend.

Obtaining historical estimates requires decisions as to whether to use exponential smoothing or moving averages, and what parameter to use in the case of exponential smoothing. However, using a different parameter for each combination of company, product, supply type and PADD is problematic. In simpler surveys, such as the EIA-812, there are natural groupings of products for which a single parameter can be obtained. In other surveys, such as the EIA-810, an empirical procedure described in this paper, was used to group cells for which the same parameter or type of adjustment was made.

2. The EIA-812 Analysis

2.1 The EIA-812 and its Roster File

Form EIA-812 collects data on end-of-month stock levels and movements of petroleum products (reformulated and conventional finished motor gasoline, motor gasoline blending components, oxygenates, finished aviation gasoline, kerosene, kerosene-type jet fuel, distillate fuel oil by sulfur content, residual fuel oil, liquefied petroleum and refinery gases, pentanes plus, and miscellaneous products) transported by pipeline. Data include stocks of products in pipelines and working tanks, as well as movements of products between PADDs. The resulting statistics are used by public and private analysts. Data are reported on a custody basis by all product pipeline companies.

The EIA-802 is a weekly survey which reports a subset of the products reported by the EIA-812. A "Monthly from Weekly" (MFW) file is produced from the EIA-802 to obtain data that should be comparable with the EIA-812 when the weeks corresponding to a month (proportionally dividing weeks that cut across months) are aggregated.

For surveys such as those conducted for the EIA-782 Monthly Petroleum Product Sales Reports, the main reason for imputation is non-response. This means, that for any given data point, if a company has missing data it will have missing data for all of its responses. However, it seems that for the EIA-812 there are situations where a company has provided a response for several items, but one of them fails an edit. For this reason each item was treated separately in the preliminary investigations.

Another issue pertaining to the EIA-812 is that there has been a change in product codes. In particular, the Motor Gasoline Blending Components has been split into six different products as of January 2004. That means that if we tried to use the products as they are collected, we would not be able to implement exponential smoothing further back than 2004. After due consideration of the changes it was decided that exponential smoothing should be analyzed using data from 2004 on, that a predictive equation including a moving average and various lags would be generated using data from 2005 only and that the evaluation of these results would use data from 2006. In addition, only months beyond twelve months after the first report would be used in any analysis. This would allow the analysis to have a twelve month lag for seasonal adjustments if necessary.

This study used the EIA-812 roster file. This file has many records with blanks and with zeros, and many combinations of company ID, product and state that are present for some

months and missing for others. It is unclear when the data is simply missing and when it should be interpreted as zero. This is particularly a problem when a company/site/product/state combination appears for only one month. One would assume that the appearance was a mistake, but it is not certain. On the other hand, if a site stops reporting in a given state for a given product, it is not clear if this means the volume was zero for that combination of product, state, company and month. Data was available from January 2001 to December 2006, but only data from 2004 on was used. For purposes of this analysis it was assumed the site was not in operation for a product before its first report or after its last.

The determination of whether a missing value should be regarded as a zero, or of whether a zero should be taken literally, when there were reported values for at least some months before and after was problematic. Also problematic were cases where there was a repetition of volumes. If more than 10 such repetitions of exact volumes were found the case was considered suspect and eliminated. Only company/PADD units with at least 12 reports in the three-year period and a gap of 16 months between the first and last were used in the analysis.

2.2 Methodology – Early Analysis

The earlier analyses used only EIA-812 data. We examined the following predictors:

1) *Lagged values* reported n months previously
2) *Exponentially smoothed historical values* are obtained by taking a previous historical volume (hv_j) and a current volume (cv_j) and a number k where $0 < k < 1$ and $hv_j = (k)hv_{j-1} + (1-k)cv_j$. We considered k at .1 intervals from .1 to .9.
3) *Average of last twelve months*
4) *Combinations of the above* using regression

There are several ways of evaluating estimators across estimates. These include:

1) *Absolute deviations* are obtained by averaging the absolute value of the estimate minus the amount being estimated (in this case the reported volume). The average across cells serves to evaluate the estimator.
2) *Root mean square of deviations* averages the squares of the deviations of the estimated volume and the reported volume, and then takes the square root so as to make the results meaningful. It is more sensitive to large deviations.
3) *Correlation coefficients* correlating the estimated and reported volumes have the drawback that the results will ignore a bias. In other words, if the estimates consistently fall 10,000 gallons below the reported volume, and this is true for everybody, the correlation would still be high.
4) *R-square in a regression without an intercept* can be used to avoid the problem posed by the correlation.

We used the fourth approach to identify combinations of estimators and create new ones, and in doing so, examined the correlations. To formally evaluate estimators, however, we used the first two approaches. As will be described later we also used several forms of trend adjustments using the monthly estimates from the EIA-802.

In order to include the site in the evaluation we required certain conditions:

1) The site must have reported at least 12 non-zero values for the product.
2) The difference between the first and last report must have been at least 16 months.

3) The first report must have been twelve months in the past (this eliminated all of 2004 from analysis, but not from contributing to historical values).

Historical values were set to current values for the first month in which a reported value was available (which must have been a year in the past for cells in the analysis). Stepwise regressions without intercepts were conducted using data from 2005 and 2006. The optimal equation was then programmed and compared to the optimal single historical estimator and the monthly from weekly estimator.

The first variable to enter, accounting for a large proportion of the variance, was the exponential smoother that calculated the new historical value as .1 of the previous historical value with .9 of the previous reported value. This was a uniform finding whether one included data from 2006 or not. However, the precise equation varied considerably depending on whether one included 2006 or not, and even depending on precise exclusions and assumptions. While an equation may have performed better than v9 by itself, the comparison was biased, because it was being evaluated in data used to derive the equation. When the 2006 data was removed, the coefficients changed and the equation no longer outperformed v9.

As a result it was decided that the exponentially smoothed historical value should be used as the basis of the imputation. The single parameter .9 is not necessarily optimal for every product and PADD, but there is not sufficient data to obtain differentially optimal parameters. It should be pointed out that in the early 1980s the EIA-782 used one parameter for all products, and this was the same parameter as has been identified here. Subsequent analysis suggested different parameters for different products.

The next part of the analysis was to include the monthly from weekly data obtained from the EIA-802. One difficulty is that there is no EIA-802 equivalent for some products. As a result, some EIA-812 products were associated with an EIA-812 estimate most likely to represent them, and for others, no adjusted trend was done (the historical value from the EIA-802 was set to 1 for every month). The analysis used all EIA-812 products with sufficient data, because a comparison with the equation was done at first, so it was deemed advisable to retain the estimates, even though the products with no EIA-802 equivalents would not enter into the analysis. The same results were later obtained by excluding them. The historical values that were matched were at the PADD level.

2.3 Modification of Analysis

The first analysis that was conducted used the optimal unadjusted coefficient and then used two adjustments, the chain-link (calculating an adjusted historical value every month) and the direct adjustment (calculating separate historical values from the 812 and 802 and then adjusting the 812 historical value by the ratio of the current value to the historical value in the corresponding PADD and product of the 802). The two approaches were comparable, but it was considered that the direct method would be easier to program, and to further explore with subsequent analysis.

A decision to further explore whether different exponential smoothing coefficients for different products would perform better led to a more careful examination of the data and the methodology. This examination called into question the methodology used for the first analysis, as the following observations were made:

1) Even as an exponentially smoothed coefficient may be optimal without trend adjustment, a different one could be better after adjustment.
2) The exponentially smoothed variables can be easily programmed so as to multiply them with the corresponding adjusted historical value.
3) Values could be adjusted by the trend from the previous month, regardless of the coefficients, but this was not optimal in most cases.
4) Extreme adjusted trends could negatively affect predicted values.

Selecting the optimal predictor was difficult, as the competing predictors are multi-collinear, and hence any addition or deletion of data points can alter the result. As some changes were made, the optimal results kept changing, and seemed to vary from product group to product group. Correlations without intercepts often differed on the third or fourth decimal place between the optimal coefficient and the runner-up. And a coefficient may have done better if absolute deviations were used, while another performed better under the criterion of square deviations. As a result, the following modifications were made:

1) A single predictor for volume was sought from three classes of variables:
    a) Exponentially smoothed historical values (including previous month)
    b) Exponentially smoothed historical values adjusted by the ratio of current month to previous month from the monthly from weekly.
    c) Exponentially smoothed historical values adjusted by the ratio of the monthly from weekly exponentially smoothed historical value with the same coefficient.
2) As before, the Product/PADD combination (with the changes to products noted above) from the MFW was matched with the 812 record for the same month. If there was no match, the trend was set to 1.
3) If the ratio exceeded 2, it was set to 2. If it was lower than .5, it was set to .5.
4) The years 2005 and 2006 were used in the regression (though the exponential smoothing began with 2004). A single predictor was selected for all combined, and for each of eight groups of products (i.e the products and records in a group were pooled and the regression was done for the pooled records). All regressions used no intercept term.
5) After most of the predictors (including the global one) fell in category c) the process was repeated using only trend adjustments with the same coefficient.

The optimal exponential smoothing coefficient when the trends were taken into account was .5 for the combined population. Three groups (two of which had no matches from the MTW) had a coefficient of 1.0 – equivalent to the historical value corresponding to the previous month. When those three groups were eliminated and a regression run, again the optimal coefficient was .3.

The ten historical values were calculated using historical values with coefficients from .1 to 1.0, using both the Monthly from Weekly file and the EIA-812. Then ten variables representing adjusted trends were presented as candidates for a univariate regression without an intercept. The best exponential smoothing parameter is presented and evaluated for each group of products. The regression was conducted at the same level as the groups, but the trends were calculated for products and PADDs.

Thus a comparison was made between a parameter of .5 for all groups (Estimate 1), a parameter of .3 for all groups (Estimate 3) or a group-specific parameter for each group of products (Estimate 2). The table of Root Mean Square Deviations is presented below.

Table 1. Root Mean Square Deviations for the EIA-812 Study

Key	Group	Param.	Est 1	Est 2	Est 3
0	All products	----------	144.2	136.1	145.7
1	Finished Motor Gasoline	.3	191.9	188.7	188.7
2	Gasoline Blending Components	.7	120.9	119.3	127.1
3	Blendstock and other	.4	83.3	83.6	86.3
4	Oxygenates	1.0	14.9	13.6	16.4
5	Kerosene	.2	121.3	114.1	115.7
6	Distillate	.3	154.5	151.3	151.3
7	Liquefied Petroleum, & Refinery Gases	1.0	132.4	109.8	144.0
8	Other and Miscellaneous Products	1.0	84.7	78.6	88.7

An examination of the results for all groups combined indicates that using the group specific parameter (Estimate 2) seems best, though whether it will be stable enough is yet to be seen. The coefficients may be sensitive to assumption and methodology, and to new data.

3. The EIA-810 Study

3.1 The EIA-810 System and Initial Preparation of Files

EIA-810 – "*Form EIA-810 collects information regarding the balance between the supply (beginning stocks, receipts, and production) and disposition (inputs, shipments, fuel use and losses, and ending stocks) of crude oil and refined products located at refineries. The resulting statistics are used by public and private analysts. Data are provided by all operating and idle refineries, as well as blending terminals located in the 50 States, District of Columbia, Puerto Rico, Virgin Islands, Guam, and other U.S. possessions."* [1]

The initial files for this study were extracted from the central EIA database which contained Company, Product, and Supply type data for various volumetric measurements as reported by month for the EIA-810 form. Corrective CINs from evaluation (prefix 999999) were eliminated. There was a format change in April of 2004, so all data from before that was excluded from the analysis, as it would have biased the data in ways that could not easily be corrected for. In addition, some records which had reported no data since April 2004 were removed from the sample entirely. These artifacts represent real companies, but ones which consistently fail to report on certain products for the 810, though they may report yearly stocks or weekly stocks which indicate that they produce stocks of that type. For those companies which reported stocks on the weekly 800 form, those results were analyzed as part of the Monthly-from-Weekly assessment performed later, but those stocks were *not* aggregated into Monthly stocks and used as valid data.

The original task was to impute only for certain Product/Supply Type combinations. Therefore, all combinations which were not requested for imputation were removed from

[1] http://www.eia.doe.gov/oss/forms.html#eia-810

the sample, with the exception of Ending Stocks, which were left in for all Products in which at least one Supply Type was being imputed. There were some Products which are reported on the 810 for which no Supply Types were originally examined, because the imputed value could be calculated from that of other cells. However, after it was discovered that direct imputations for some of these were better than for some of the other cells, they were restored to the file.

The supply types for each product are related based on the equation:

Ending Stocks= Beginning Stocks +Receipts+Gross Product-Inputs-Shipping-Uses and Losses.

In addition beginning stocks could be calculated from the previous month's ending stocks. Most of the time these two matched or were different by a small value.

The data from the EIA-800 Weekly Survey were evaluated by obtaining monthly data by creating a Monthly-from-Weekly volume for each company, product, and supply type combination. Similar to the data from the 810, all data from before April 2004 were removed; all companies who had not reported a particular product/supply type since April 2004 were removed. In addition, any weeks with a reported zero volume were removed from the data set, but included in the MFW volume for that month. If all weeks for a month were zero, then that month was removed entirely. These volumes were compared to the imputation estimates for each month, and the optimal method over time was chosen by product and supply type. However, the 800 only collects ending stocks and inputs/net production, depending on the product, so a straight product/supply type comparison was not possible for all EIA-810 combinations.

The EIA-810 presented particular difficulties in its imputation analysis:

1) Multiple supply types and products
2) Not all products report all supply types
3) Not all supply types are found in corresponding Weekly (EIA-800)
4) Some supply types are derived from others
5) Volumes for some supply types can be negative (particularly in EIA-800)
6) Large number of 0 or missing volumes

This led to a need for some procedures beyond those used in the EIA-812

3.2 Exponential Smoothing and Moving Averages

The initial imputation procedure was done at the cell level, where a cell was defined as a combination of company, product, and supply type, and used only the previous month's reported value from the same cell to conduct the imputation, replicating the current survey's imputation methodology. No connection was made between a company's report of a given supply type and product and that of another supply type or product. No distinction was made at this point between zero and missing volumes.

The data that were used for each cell began with the first positive report on or after January 2004 and ended with the last report up to and including June 2008. In order to be used in the analysis, a cell had to have at least eight positive values reported and at least 12 months between the first and last months with reported values. Any month between the first and last with no positive report was treated as zero reported volume.

The first month reported for any ID was treated as if the previous twelve months had been reported using the same value as that for the month itself. This is the same

procedure as used in the EIA-811 and EIA-812 imputation studies. The alternative was to assume that prior reports would have been zero. While this may be reasonable for months with missing data between the first and last reports, we decided that using the previous months equaled the current month was more robust.

Eleven different historical average values were then obtained for each cell and month by varying the weight placed on the smoothed historical data and weight on the most recent reported data up to that month. In particular, the first value (indexed k = 0 in several tables) was a moving average i.e. the simple average of the previous twelve months. The other ten values, indexed from k=1 to k=10 reflected an exponentially smoothed average. Assume that x_j was the historical value for month j and r_j was the reported value for month j. Then for index k, $x_{j+1} = x_j (1-k/10) + r_j (k/10)$. As can be seen, when k=10 then the historical value is the reported value from the previous month (i.e. the current imputation procedure) and when k=1 it is .9 times the historical value, plus .1 times the reported value of the previous month.

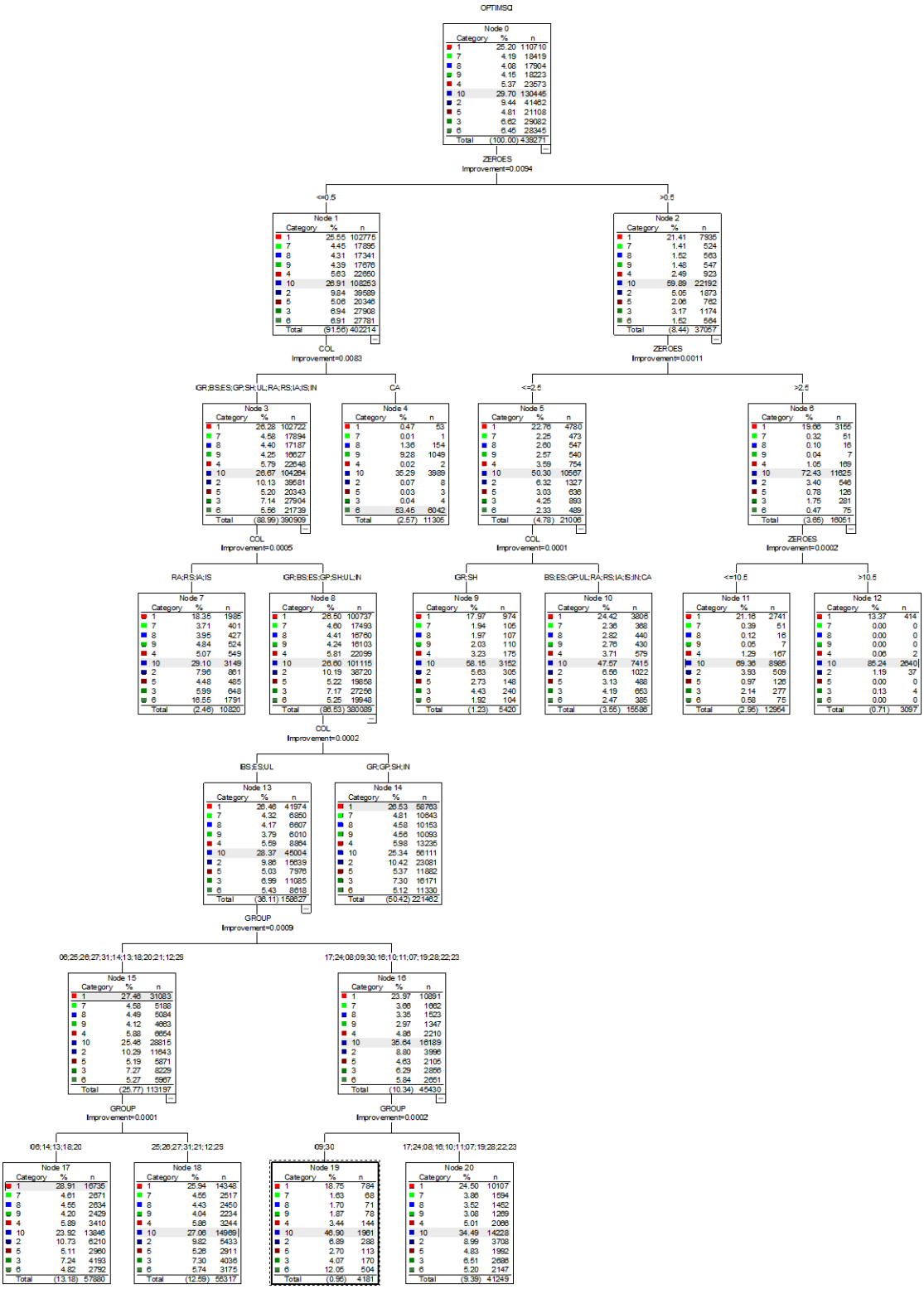
In order to minimize the bias caused by the 12 replicated artificial values before the first month of the evaluation period, the imputed values for the first 12 months in the evaluation period are excluded from the analysis. Thus the first year contributing to the tables is 2005, 12 months after the first report of actual data. The remaining months, up to the last reported month for the cell, were used to determine the correct smoothing parameter to be used in the imputation. Three criteria which were used in the prior imputation analyses for the EIA-812. These were the root mean square of the differences between the predicted value and the reported value, the mean absolute difference between the same two values, and the absolute value of the maximum difference between the same two values. Different cells yielded different optimum values, therefore, defining the domains (product, supply type) for which the optimum value should be calculated was not straight forward. In particular, a balance was needed between avoiding combining cells too many different optimum parameters across cells and having too many cells which would not result in a robust performance, particularly given some with a small number of responses. One effort at achieving homogeneity was to group the products into groups, under the assumption that members of the same group would have similar optimum parameters. But even so, the groups, each of which contained only a few products, were too numerous to achieve homogeneity and robustness.

3.3 Regression Trees

However, the analysis had indicated that the moving average was seldom optimal, and thus the problem could be reduced to finding an optimal parameter (k) from 1 to 10. A regression tree procedure, Classification and Regression Trees (C&RT) was used to find the optimal parameters. After a few attempts trying to use it on aggregated data, it was decided to use each estimate (supply type, product, company and month combination) as a data point. Both a continuous and categorical approach was tried for the dependent variable, but the categorical option was superior. So, for each cell the exponential smoothing parameter (k) resulting in the smallest absolute discrepancy from the reported value was identified.

Another key decision was what variables to use as predictors, the variables that define the possible splits in the regression tree. At first PADD, supply type, group and product were used. Soon, however, it became apparent through the analysis that the number of zeroes (or missing/blanks) preceding the current month made a difference in what the optimal parameter was as well. Therefore, the number of preceding months with zero volume was added as a predictor variable.

Figure 1. Regression Tree



The first split in the regression tree hinged on whether the previous month's reported volume was zero. Four of the terminal nodes were defined by having zero volume for one or more previous months. Nodes 9 and 10 represent cells where one or two previous months were reported zeros, with 9 having supply types (variable COL in the diagram) GR (gross receipts) or SH (shipments) and node 10 having the remaining supply types. Node 11 consists of cells where 3 to 10 consecutive previous months reported zero volumes and Node 12 of cells with at least 11 consecutive previous months reported zero volumes.

On the other side (left) of the main split in the tree, CA (operable capacity) formed Node 4. Node 7 was formed by crude oil RA (receipts average API gravity), RS (receipts average sulfur content), IA (input average API gravity) and IS (input average sulfur content). Node 14 was the largest node, and consisted of supply types GR (receipts), GP (production), IN (inputs) and SH (shipments). This node could not be split further using the predictor variables. With only one parameter for such a large group, this group generated the largest number of outliers. Nodes 17 through 20 included BS (beginning stocks), ES (ending stocks) and UL (fuel uses and losses), with the differentiation being the actual product groups represented.

We define categories according to the terminal nodes a value to be edited falls in. For each category the different exponential smoothing coefficients produced different estimates. The optimum coefficient for each category was the one that yielded the prediction closest to the reported volume. The smallest sum of the squared deviations between predicted and reported volumes (reported as a root mean square – RMS – which is a monotonic function of the sum of the squared deviations) was the criterion used for "closest". Absolute deviations and the smallest maximum deviation were also examined, but the criterion used at every stage was the minimum square deviation. Results for the terminal nodes are presented in Table 2.

Table 2. Results for Terminal Nodes

Node	Optimum RMS	Optimum Absolute Deviation	Minimax Criterion	Non-zero Records	Total Records
4	10	10	7	11,257	11,305
7	7	8	5	10,760	10,820
9	4	9	1	2,333	5,420
10	6	9	4	8,395	15,586
11	6	10	5	3,969	12,954
12	2	10	4	457	3,097
14	5	5	1	208,873	221,462
17	2	3	7	70,798	71,099
18	5	5	8	41,920	42,098
19	5	6	9	4,089	4,181
20	10	10	5	40,668	41,249

These optimum parameters from the root mean squared deviations were used as the exponential smoothing coefficients for the remaining of the study. However, it is imperative that one point be understood. Each company, product, and supply type needs to carry several historical values. The reason is that the combination will fall into different categories depending on the previous month's report or the months that precede it. Thus a cell may be in category 20 and be using a smoothing coefficient of 10. If all of

a sudden it reports zero volume, the coefficient will change. But it is not applied to the previous historical value, but to the historical value obtained if the coefficient for its new cell had been used all along. For this reason the historical values need to be carried on for all the smoothing coefficients (or at least for several) for each cell.

3.4 Adjusted Trends

A file of Monthly-from-Weekly (MFW) data was extracted from the SIS database. These values were based on data from the weekly survey (EIA-800) system. These MFW values provided early estimates for the counterpart values reported on the EIA-810. However, not all products or supply types are collected by the EIA-800 nor all companies because the EIA-800 is a cutoff sample of the EIA-810 respondents. The MFW values were used to build into the methodology an adjustment for trend. However, because the EIA-800 is a sub-sample, the trend adjustment cannot be constructed at the company level nor could it be constructed for all supply types and products. In addition, there was also the difficulty that the MFW values contained negative volumes for Input and Net Production for some products.. In particular, negative values were present for gasoline blending components.

As a result, trend adjustment ratios were constructed for three MFW supply types--ES, IN and NP (net production). Aggregate values were obtained at the national level (because the EIA-810 is a site survey, the number of reported values for each PADD can be considerably small, and hence the ration is likely to not be robust) and exponentially smoothed historical values were calculated for each supply type-product combination. In the instance of residual oil, only totals (all sulfur levels) are reported on the weekly so the one MFW value was applied to every sulfur type of residual oil reported in the EIA-810.

Ten ratios of current values to exponentially smoothed values were obtained for each MFW supply type/product/month combination possible (moving averages were not found to be optimal for any of the categories). The smoothing was done for the historical value only, and this was used as a denominator, with the current reported volume as the numerator. Note that it was necessary to calculate all ten ratios (or at least more than one) because the same company/Supply/Product combination may use different historical values in different months. If the historical value was zero the ratio was set to 1(no trend adjustment). The ratios were capped at a maximum of 2 and a minimum of .5 (including where the current volume was 0 or negative and the historical value was positive). If the historical value was negative and the current volume was positive then the ratio was set to 2. After merging the file of ratios with the file of historical values from the EIA-810, the ratio that corresponded to the optimal exponential smoothing parameter for the EIA-810 was used to adjust the exponential smoothing. In addition, analysis was also performed to determine if a different supply type MFW ratio performed better in adjusting ES. The ES ratios were estimated with all the exponentially smoothed predictors regardless of supply type. The analysis showed that use of different MFW supply type ratios (e.g. using ES to impute GP or SH) applied to monthly historical values was not useful in the general case. However, ES was useful when the preceding value was zero or when the stock was BS (though BS could be better obtained from the previous month's ES). Nor was combining NP and IN as if they represented one supply type effective. Despite the distinction of net in the NP MFW ratio, NP proved effective as a trend adjustment for GP in the EIA-810.

The results of applying the trend adjustment to the historically smoothed monthly values were evaluated by product and supply type as well as by the previous month report being

a zero. The evaluations showed mixed results; in some cases application of the trend adjustment lead to a better estimate but not always. Based on that evaluation, two approaches were identified: 1) unadjusted (historically smoothed only) regardless if an MFW smoothed ratio is available; and 2) trend adjusted using the matching MFW supply type ratio (ES, GP or IN).

4. Summary and Conclusions

The analysis showed that one can develop imputations in the MPSRS by first using exponential smoothing to obtain a historical value and then using corresponding weekly aggregate data if available to obtain adjusted trends. However, a key decision in this sort of analysis will always be how to group the data in order to identify the proper coefficients. In some surveys, such as the EIA-812, there are natural groupings which can be used and have been found reasonably homogeneous with respect to the exponential smoothing coefficient. In other surveys, such as the EIA-810, there are too many product/supply type combinations as well as too many zero volume cells. The use of regression trees provided a data driven approach to the problem, but still left a very large group that could not be split and for which the compromise coefficient of .5 had to be used. Further research and further variations on the regression tree approach are necessary.