Separating the Wheat from the Chaff: The Search for the Best Imputation Methodology

Paula Weir and Pedro J. Saavedra
Paula Weir, Energy Information Administration
Dr. Pedro Saavedro, ORC Macro

Abstract

Developing an imputation methodology is often wrought with basic data issues. Yet, the interpretation and treatment of the data have bearing on the methods to be considered and the performance of competing estimators. In this study, the data suffer from a recent change in both the data elements collected and the processing system, confusion over truly missing versus zero values, and reliability of edit failed data elements. In addition, the implemented imputation must be performed without access to concurrent reports, as the system requires immediate imputation before data from other respondents are available. Exogenous data from a related survey are considered, and a number of different estimators are compared through an exploratory approach to determining an imputation model that is compatible with both processing requirements and data characteristics.

Keywords: survey processing, trend, regression, exponential smoothing

1. Background

The Monthly Supply Reporting System (MSRS) represents a family of nine data collection survey forms that are used to collect detailed operations data, stocks, imports and movements of crude oil and petroleum product from refiner/blenders, bulk terminals, natural gas plants, oxygenate products and pipelines as needed to meet EIA's mandates and energy data users' needs for credible, reliable, and timely energy information. The integration of these data also allow calculation of products supplied. Products supplied is used as an approximation of consumption of petroleum products because it measures the removal of these products from primary sources, i.e., refineries, natural gas processing plants, blending plants, pipelines, and bulk terminals. In general, product supplied for each product in any given period is computed as follows: field production, plus refinery production, plus imports, plus unaccounted for crude oil, (plus net receipts when calculated on a PAD District basis), minus stock change, minus crude oil losses, minus refinery inputs, minus exports.

The mandatory requirement for companies to file these forms has historically affected high response rates (98-100%). Imputation was performed for seven of the nine forms, using the previous period's value. Data were not imputed for two surveys that collect data on imports, and tanker and barge movements because of the high variability of those data. Response error, the difference between the true value and the value reported on a survey form, was therefore considered to be the major factor affecting the accuracy of data. To aid in detecting and minimizing reporting errors, automated editing procedures were used to check current data. These checks included verifying the current data for consistency with past data, verifying internal consistency (e.g. totals equal sum of parts), examining orders of magnitude. Data elements that failed edit criteria were flagged and reviewed manually to determine if corrections were needed.

The Monthly Supply Reporting System is complemented by the Weekly Supply Reporting System (WPSRS). The EIA weekly reporting system was designed to collect a subset of data similar to those collected monthly through six survey forms. In the WPSRS, selected petroleum companies report weekly data to EIA on crude oil and petroleum product stocks, refinery inputs and production, motor gasoline blending operations, and crude oil and petroleum product imports. The sampling procedure used for the surveys in the WPSRS is the cut-off method where companies are ranked from largest to smallest on the basis of the quantities reported during some previous period. Companies are chosen for the sample beginning with the largest and companies are added to the sample until the total sample covers about 90 percent of the total for each item and each geographic region for which weekly data are published. The estimation procedure uses a ratio estimator. First, the current week's data for a given product reported by the sampled weekly companies in a geographic region are summed to form Ws. Next, the most recent month's data for the product reported by only those same companies in the monthly are summed to form Ms. Finally, the most recent month's data for the product as reported by all companies in the monthly are summed to form Mt. The weekly estimate is calculated by multiplying Ws by the ratio of Mt to Ms.

In 2004 the processing system for the family of monthly surveys was replaced by the Standard Economic Processing System (StEPS) developed by the U.S. Bureau of the Census. In addition, a number of survey changes were also implemented that provided new product details in keeping with the changes in the industry. While attempts were made to reconstruct the previous system's

edits within the StEPS environment, not all were implemented, and no historical data were available for the detailed new products for editing. To further complicate matters, compilation of components across surveys could not be performed in StEPs, therefore dissemination-level estimates could not be examined until very late in the production process to identify potential reporting problems.

The combination of new forms and new system during a time of industry changes resulted in some unverified data. As a result, it was decided to implement imputation for not only nonresponse, but also for select data failing edit rules. In order to view the progress towards releasable aggregate estimates, it was also decided that imputed values would be required for all respondents at the beginning of the monthly reporting cycle, and these values would be used in aggregations until response was received and/or select edit failures overridden. At any point in the processing cycle, a preview of the aggregates could be provided using reported and imputed values to account for the population. In addition, a better imputation method was desired to improve on the previous period's value method. The research was targeted to start with developing a methodology for just one survey and product that was of the most concern. The survey that was chosen was the EIA-811, "Monthly Bulk Terminal Report" which collects end of month stock levels of finished petroleum products. All bulk terminal operating companies located in the 50 States, the District of Columbia, Puerto Rico, the Virgin Islands, and other U.S. possessions that have a total bulk storage capacity of 50,000 barrels or more, and/or receive petroleum products by tanker, barge, or pipeline are required to report by State for products that resides in the custody of the bulk terminal company regardless of actual ownership of the product. The specific product stock data that were to be examined to start the imputation study were gasoline products and gasoline blending components, a total of eleven data elements. Unfortunately, these products were not collected with the same detail prior to 2004, but could be summed to form three higher level products. This lack of historical data was a limitation on the methodological research.

2. The Imputation Research

The analysis conducted can be divided in three parts. The first part of the research used data from 2001 forward, but was limited to the old level of detail at which products existed on the survey forms prior to 2004. The second part of the research used data from 2004 forward and focused on cell levels corresponding to the level of the weekly survey and made use of monthly estimates derived from the weekly surveys. These derived estimates are known as the Monthly-from-Weekly or MFW. The third analysis turned back to all cells for the given set of products, using data from previous months for the responding company, but using the MFW data to adjust for trends at the aggregate level.

2.1 The First Analysis

For many surveys the main reason for imputation is to account for non-response. This means that if a respondent has missing data items it will have missing data for all of its survey items. However, in this case, the requirement for the EIA-811 imputation was to also replace data items that failed certain edits. For this reason it was decided to treat each survey item separately in the preliminary investigations.

The critical issue pertaining to the EIA-811 was the change in product codes in 2004 and the expansion of product detail. In particular, the single product Motor Gasoline Blending Components was split into six different products as of January 2004. As a result, the types of analysis on the more detailed products that could be conducted were limited. One consequence was that any analysis of exponential smoothing would have less than a full year's worth of data at that time, or have to be performed at a level higher than required for imputation in production.

In addition, the data file had 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 was unclear when the data were simply missing and when they should be interpreted as zero. This was particularly a problem when a company/site/product/state combination appeared for only one month. One could assume that the appearance was a mistake, but it was not certain. One the other hand, if a company stopped reporting in a given state for a given product, it was not clear if this meant that the volume was zero for that combination of product, state, company and month.

Given these data issues, a thorough examination of exponentially smoothed historical prices, an effective approach used in a number of EIA surveys, was discarded in the first analysis. The combination of having only eleven months where there would be sufficient data, and the many instances of blanks, zeros or missing records in the series, would have made the results questionable. As a result, the analysis started with the 2004 product levels in order to examine what worked best, but limited the study to the six motor gasoline blending components. The intent was to examine the data for the best model, even models not satisfying the requirements, even if impractical in production, in order to shed light on the nature of the data.

Three models were looked at. The first was the use of just the previous month's value. In addition, two other simple estimators were defined that made use of other respondents' data. The first was simply x_D or the average volume reported by other members of that domain p. The second estimator adjusted the previous month's response (x') by the change in the average volumes reported in that domain (excluding the number being imputed), $x'(x_D/x'_D)$. For these estimators, let x be the value of an item (e.g volume by a given company at a given site in a state and region for a given product).

The first analysis examined this second imputation estimator for seven different domains. Both estimators were examined through a stepwise regression, with predictors being the fifteen estimators so defined (two for each of the seven domains for the second estimator, plus first estimator that used the unadjusted previous month value, including and excluding zero volumes). It was hoped that this analysis would provide insight on how these estimators might work, and how they could be combined.

2.1.3 Methodology

The analysis used data from 2004 forward but used the product detail level of 2001 under the product category Motor Gasoline Blending Components. From January 2004 forward, this category had been broken into six products to reflect the different kinds of blending components. The seven domains used were as follows:

1) Company
2) Product
3) Region
4) State
5) Product and Region
6) Company and Region
7) Company and product.

A record was created for every month for every combination present in the file. Two stepwise regressions were run, one using all records and the other eliminating any records where the historical volume was zero. These records were not only eliminated from the regression, but also from the calculation of average volumes for each domain.

One of the difficulties with stepwise regression using a large sample was deciding when to stop including new variables. Stability and interpretability were two important considerations. The process here was validated by the fact that the two five-predictor equations (with and without the zeroes) had exactly the same variables (though they entered in a different order) and very similar coefficients. This was true, even though the second set (excluding zeroes) had less than half the data points of the first (1106 vs. 2277). Table 1 presents the first order correlations of the fifteen predictors with volume.

Table 1

Predictor	Zero Volumes Included	Zero Volumes not Included
Trend-Company .	0.1886	0.7436
Trend-Product	0.6978	0.6142
Trend–Region	0.8203	0.8625
Trend–State	0.8323	0.8606
Trend-Product & Region	0.3269	0.3882
Trend-Company & Region	0.1887	0.3932
Trend-Company & Product.	0.6221	0.8237
Mean-Company .	0.0732	0.0979
Mean-Product	0.0179	0.0633
Mean–Region	0.0620	0.1406
Mean–State	0.2716	0.3616
Mean-Product & Region	0.1798	0.3704
Mean-Company & Region	-0.0045	0.0535
Mean-Company & Product.	-0.0253	-0.0371
Historical Volume	0.8617	0.8732

The regressions indicated that the historical value alone contributed the bulk of the prediction. Trend was significant but added little to the R squared. Unsurprising, it was also found that exclusion of zeroes made for better predictions. While the analysis did show the importance of historical volume, the imputation methodology was required to impute values before data are received for that period, therefore, the use of the domain means from the current month for either point estimates or trend adjustments would not be possible.

2.2 Second Analysis

While the earlier analyses were conducted only for 2004, the second analysis stepped back in time and explored

data from 2001 through 2005 at the 2001 product detail levels. In addition, it examined the use of data derived from the weekly survey. After combining the months, data at the higher product level were available for 41 months. At this higher level, the longer data series qualified for exponential smoothing and examinations of lags. As a result, the second study assumed that imputation would use only a company/product/state own historical data in carrying out the imputation or available derived data from the weekly survey.

For this analysis, the following five predictors were examined:

1) *Lagged values* reported n months previously,
2) *Exponentially smoothed historical values* obtained by taking a previous historical volume (hv_j) and a current volume (cv_j) and a number k where $0 < k < 1$ and $hv_j = (k)hv_{j-1} + (1-k)cv_j$. and k was considered at .1 intervals from .1 to .9.
3) *Average of last twelve months*
4) *Estimate of the Monthly-from-Weekly derived from the weekly survey (MFW)* at the PADD level,
5) *Combinations of the above* using regression.

To evaluate the estimators across estimates, the following were examined:

1) *Absolute deviations* obtained by averaging the absolute value of the estimate minus the amount being estimated (in this case the reported volume). The average across cells served to evaluate the estimator.
2) *Root mean square of deviations* averaged the squares of the deviations of the estimated volume and the reported volume, and then took the square root to make the results more meaningful. This measure is more sensitive to large deviations.
3) *Correlation coefficients* correlated the estimated and reported volumes but had the drawback that the results ignored possible bias. (If the estimates consistently fell 10,000 gallons below the reported volume, and this was true for everybody, the correlation would still be high.
4) *R-square in a regression without an intercept* was used to avoid the problem posed by the correlation.

The fourth approach was used to identify combinations of estimators and create new ones, and in doing so, examined the correlations. For formally evaluating the estimators, however, the first two approaches were used. The product categories of Motor Gasoline Blending Components, Reformulated Finished Motor Gasoline, and Conventional Finished Motor Gasoline were used. While values were calculated beginning with 2001, in order to appear in the analysis (after the estimators were calculated) a cell had to meet the following conditions:

1) The respondent 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 18 months.
3) The first report must have been twelve months in the past (this eliminated all of 2001 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). The initial analysis took place at the cell level (company/State/global product combination), but data were eventually acquired from the weekly report at the PADD level. The imputations were conducted at the company/State level, but evaluated at the PADD level.

For this analysis, the MFW proved to be more effective than any of the historical predictors, but was further improved by combining it with historical predictors. Because MFW is only available at the PADD level, all results are presented at the PADD level.

As previously done, the missing values were treated as zeros. In addition, if a cell did not have a sufficient number of non-missing values, it was deleted. This meant that the State value for that company did not contribute to the PADD value for the company, so a difference between the monthly and the MFW could be due to the missing State. In addition, the historical values were all calculated at the cell level, and this had to be taken into account as Company/state cells appear and disappear from the survey. Zero values were then treated differently when the PADD level analysis was implemented. While these values were included as zeros in the exponential smoothing and lag values, they were not included in the regression and evaluation of analyses. For both regression and evaluation, all values where either the MFW from weekly estimate or the monthly value, but not both, was zero, were excluded. So even though zero values and missing values treated as zeros went into the historical predictors, they were not included as data points in the actual analysis.

Stepwise regressions without intercepts were conducted using data up to May 2004. The optimal equation was then compared to the optimal single historical estimator and the monthly from weekly estimator. The comparisons were done using June 2004-May 2005.

Table 2 presents the evaluation of the single estimators across the months from January 2003 to May 2004. As can be seen, the MFW from the weekly outperformed all of the historical estimators. The best single historical estimator was the exponentially smoothed historical variable that averaged the previous historical estimator with the new volume to form a historical estimator for the following month.

The regression yielded three predictor variables. First, it accepted the MFW estimator. While it was optimal for the average absolute standard deviation, it was worse for the root mean square (RMS) deviations. This suggested that the estimates from the weekly survey are closer for most cells, but had very high errors for a few cells. The exponential smoothing in general had lower RMS deviations and the .5 coefficient used to average the new reported value with the old historical value to get a new historical value, was the lowest of the group of coefficients. A stepwise regression using all the predictors was then performed in order to identify combinations of predictors that might outperform individual ones. The equation was estimated with no intercept so that a zero prediction would yield zero volume. The equation that resulted was:

.5883*MFW + .39582*V5 + .02049*Lag12

where MFW was the Monthly-from- Weekly, V5 was the exponentially smoothed estimator with a parameter of .5 and Lag 12 was the volume 12 months previously.

Table 2: Univariate Evaluation of Predictors

Predictor	Correla-tion	Abs. Dev.	RMS Dev.
MFW	0.9726	52.31	99.94
Year average	0.9440	80.10	141.77
Exponential .1	0.9424	96.29	168.56
Exponential .2	0.9590	72.24	122.93
Exponential .3	0.9626	68.71	116.17
Exponential .4	0.9637	67.85	114.64
Exponential .5	0.9638	67.78	114.55
Exponential .6	0.9633	68.27	115.41
Exponential .7	0.9624	69.18	117.08
Exponential .8	0.9609	70.54	119.51
Exponential .9	0.9588	72.24	122.73
Lag 1 month	0.9562	74.41	126.82
Lag 2 month	0.9442	85.39	143.49
Lag 3 month	0.9344	91.50	155.56
Lag 4 month	0.9255	96.22	165.14
Lag 5 month	0.9202	98.81	170.44
Lag 6 month	0.9163	101.50	174.63
Lag 7 month	0.9112	104.70	179.96
Lag 8 month	0.8990	108.79	192.30
Lag 9 month	0.8896	113.28	201.03
Lag 10 month	0.8864	114.02	203.78
Lag 11 month	0.8869	113.23	203.51
Lag 12 month	0.8862	112.10	204.24

It should be kept in mind that the cells are not independent; thus the significance tests should not be taken as indicators of probabilities, but rather strong indicators of a consistent effect.

The three estimators (the equation, the MFW and the historical smoothed) were compared across the twelve months, and all PADDs and products. Table 3 presents the three estimators and their results. Table 4 presents the comparison of the estimators.

Table 3: Differences between estimators and actual volumes

Estimator	Mean Diff.	Absolute Diff.	RMS Diff.
MFW	1.20	60.98	117.13
Historical	-7.39*	85.56	152.08
Equation	-1.82	59.69	102.42

* p<.05

Table 4: Comparison of the estimators

Comparison	Absolute Differences	Squared Differences
MFW-Historical	-8.25***	-4.03***
MFW-Equation	0.89	2.98**
Historical-Equation	12.97***	7.28***

** p<.01
*** p<.001

The MFW estimator outperformed any single estimator based on historical volumes. However, when combined with an exponentially smoothed estimator and a 12-month lag, there was a statistically significant improvement, even if not of a large magnitude. This could only be detected using squared deviations, which suggested that the equation showed improvement in instances where the MFW particularly failed. Only the historical estimator showed a bias, systematically underestimating the volumes.

An examination of those cells where the MFW failed to the greatest extent suggested that the result might be an

artifactual result. In particular the early analysis was conducted at the company/State level, but the final regression equations and evaluations used the company/PADD level. Furthermore, the early analysis began by equating the historical and the actual price, but then dropped the cells with fewer than twelve months of historical prices. Thus, only cells that had been reporting for at least twelve months were used (zero cells not followed by non-zero reports were also dropped). However, when the MFW estimator was brought in, it was defined at the PADD level, and all the State level data, including the historical estimators, were aggregated to the PADD level. The net result was that if new States were recent additions, or appeared sporadically, their volumes would not have been added to the PADD total from the monthly. This possibly affected the MFW estimator, but not the historical estimators.

2.3 Third Analysis

One difficulty with the results of the second analysis was that imputation was expected to take place at the cell level, but the MFW was obtained at the PADD level. Thus, the exponentially smoothed values would be more desirable, even if they were not as accurate as the MFW. However, the MFW could play a useful role in the imputation, even if not applied at the same reporting level where the edits take place. For example, the MFW could provide an indication of trends at the aggregate level, and these trends could be used to adjust the estimates at the reporting level.

With this thinking the third imputation analysis focused on a chain link approach. This entailed two steps:

1) the creation of a historical value for each reported value;

2) the adjustment of the historical value by trends in the monthly.

After examination of various possibilities the following general steps were taken:

1) a historical value was obtained using a weighted average between the previous month reported volume and the predicted value for the same reporting period and product;

2) the ratio of current month to previous month was obtained using a suitable cell from the MFW;

3) the historical value was then multiplied by the ratio to obtain the new predicted value.

Several parameters that had to be established in this case were:

1) the weight given to the reported and predicted values from the last period;

2) the appropriate cell level (PADD, product, combination, etc), complicated by the fact that not all products were reported in the MFW in each PADD; 3) whether the denominator in the ratio should also be exponentially smoothed;

4) whether there should be exponential smoothing of the historical unadjusted by the ratio.

The last two were consistently decided in that the best denominator seemed to be the previous month MFW value, and the exponential smoothing including adjustments at every step seemed superior. The imputation formula had the form:

$$P(t)=(bR(t-1) +(1-b)P(t-1))(MFW (t) / MFW (t-1)),$$

where P is the predicted volume for a given month and R is the reported volume for that month and b is the parameter.

The best parameter for b depended on the criteria used, but seemed to be around .6, possibly varying by product and PADD. The historical value for the MFW Weekly turned out to be the previous month. Finally, the ideal cell division for the ratio (the domain over which the ratio is taken) was unclear, and the approach seemed to work much better for some products and PADDs than others. One difficulty was that not every product was reported in the Weekly for every PADD, and in some PADDs, some products appeared some months, and the reported volume was zero for others. Thus, a method of defining a domain over which to adjust the trends was one of the difficult issues here.

The simulations were done using three clusters of gasoline products representing Reformulated Finished, Conventional Finished and Blended Components. Regions were combined in different ways, but in general PADDs 1 and 2 (East and Midwest) were combined as were 3 and 4 (South and Mountain) with PADD 5 standing on its own. Simulations were done with and without combining these three regional groups into one for purposes of calculating the trend, and with and without combining the product categories into one, as well as without the trend adjustments. This added to five different trend adjustments. For each the parameters for b were varied from 0 to 1 in increments of .01, and the optimal parameter was selected. The criterion used was the root mean square deviation over a twelve month period for all companies and products. The evaluation was done for each regional group/product group combination separately. Table 5 presents the results.

In particular, products corresponding to the old product 134 (Blended Components) had the worse fit, and those in the old product 152 (Conventional) provided the best. Various ways of collapsing PADDs were attempted without a clear-cut optimal design. As of now, the new products have only been reported since 2004, so there are

not sufficient data to fully establish the optimal procedures. Examination of the data did seem to suggest though that the domain for which one calculates the trend adjustments may matter, and the optimal procedure may be different for different product types and regions.

Table 5: Chain Link Analysis: Root Mean Square Deviations

Regions	Products	Region X Product	Region Only	Product Only	Single Domain	No Adjustment
1 & 2	**Blended**	137.294	139.721	140.971	139.351	142.085
1 & 2	**Reformulated**	87.466	88.392	86.738	88.267	90.204
1 & 2	**Conventional**	77.493	78.015	77.610	78.014	79.020
3&4	**Blended**	148.403	171.461	174.414	178.919	186.613
3&4	**Reformulated**	96.033	109.196	113.341	111.933	117.747
3&4	**Conventional**	75.616	75.382	76.485	75.992	79.133
5	**Blended**	89.561	88.952	93.623	93.852	91.653
5	**Reformulated**	49.128	47.292	51.571	48.890	49.395
5	**Conventional**	41.249	41.349	41.012	41.300	41.696

3. Summary and Future Work

In this study, the development of an imputation methodology was wrought with basic data issues. Limited historical data, zeros and blanks, and survey processing requirements for imputation resulted in the use of three related analysis. The first two analyses were exploratory. The first resulted in confirming the predictive value of the previous period's response, as well as the significance of a trend adjustment. The second analysis showed that the best single predictor was the MFW derived from the sister survey. It also showed that an equation that combined the MFW and exponential smoothing and a lag of 12 months corrected for cases where the MFW greatly deviated from the monthly. The third analysis made use of the results from the two exploratory analyses to define an equation that satisfied the survey processing requirements that impute values be available at the beginning of the survey cycle and be used for both non-respondents and select failed reported data.