Researching Forecast Models for Item Imputation in an EIA Survey

Joseph D. Conklin

Energy Information Administration, U. S. Department of Energy, Washington DC 20585
joseph.conklin@eia.gov

The analysis and conclusions contained in this paper are those of the author and do not represent the official position of the U.S. Energy Information Administration (EIA) or the U.S. Department of Energy (DOE).

Abstract. EIA's monthly natural gas survey collects volume and revenue data that support the estimates of state level volumes and prices published in EIA's *Natural Gas Monthly*. This paper and presentation focus on research over a nearly ten year period to improve imputations for volumes that are either not reported or of questionable quality. EIA considered (1) incorporating local heating degree data to improve predictions of distributor level volume, (2) developing multiple measures of success, (3) incorporating variables not included in earlier research, (4) identifying improvements likely to sustain themselves over the long term, and (5) testing novel ways of predicting the sales and transportation components of volume. While no procedure appeared superior after validation testing on 2011 data, a useful by-product of the research was linking the current sample selection code into the imputation code for the monthly natural gas survey.

Key words: imputation, survey, natural gas

1. Purpose of the Survey and the Forecast Approach for Imputing

EIA's monthly natural gas survey provides foundational data for EIA's publication *Natural Gas Monthly* (NGM). It is collected from a sample of local distribution companies, and the data collected shows the value chain of natural gas prices from the city gate to the end user, as well as the volumes delivered to each customer class. It provides the demand side of the natural gas balance equation which EIA reports monthly and annually. There are always some responses that are either missing or questionable that are not resolved by the deadline for the NGM.

Three sectors of interest in the NGM are commercial, industrial, and residential. The commercial sector consists of service-providing facilities and equipment of businesses; Federal, State, and local governments; and other private and public organizations, such as religious, social, or fraternal groups. The industrial sector consists of all facilities and equipment used for producing, processing, or assembling goods. The residential sector consists of living quarters for private households.

As the current NGM is readied near the end of a processing cycle, the imputation procedure for the next cycle is run. The output provides substitute values for distributors with missing or questionable responses not resolved by the next NGM. As the next month proceeds, the sum of actual responses received to date and the imputed values of distributors yet to report provides a leading indicator of the volume eventually published in the NGM.

2. Impetus for and Conclusions of the Research

Efforts to improve the procedure for the imputed values for volumes date back to at least 2003. In fall 2010, EIA adopted a new approach for calculating standard errors of NGM volume data. The improvement in the standard error approach allowed the sample size for the survey to be reduced from approximately 400 to 300 respondents, resulting in significant simplification and efficiency gains. Completion of this work stimulated interest in other ways to improve the survey.

The research effort attempted several things:

(1) extend research carried out in 2010 and 2011 that incorporated local heating degree data to improve predictions of distributor level sendout volume,

(2) develop multiple measures of success,

(3) incorporate variables not considered in earlier research,

(4) identify improvements likely to sustain themselves over the long term, and

(5) test novel ways for predicting the sales and transportation components of natural gas deliveries.

By the end of June 2012, EIA had tested fifteen categories of alternative procedures. The models were fit at the individual distributor level. No category appeared superior on an overall and long term basis compared to the current procedure. EIA decided to retain the existing procedure through 2012. Starting in 2013, as part of an effort to consolidate statistical programs for price and volume calculations, EIA converted to a new strategy for imputation. It uses a classical ratio estimator to treat missing and questionable responses as if they were not a part of the survey sample.

3. The Current Procedure and Summary of Research from 2003 to 2006

EIA researched improvements to the procedure from 2003 to 2006. The purpose of this section is to summarize research during this period. Also, it describes the current lists and the major categories of methods employed in the research.

The current procedure employs a two-step procedure: estimation of the monthly Census division totals for the residential, commercial, and industrial sectors. For this step, Alaska and Hawaii are treated as their own Census divisions. For the residential and commercial sectors, the Census divisions are predicted from a regression model in which the independent variables are monthly heating degree days and a monthly dummy variable. A regression model is also used for the industrial sector, but the independent variable is the prior month's division volume after application of a weighting factor.

The second step is allocation of the Census division estimates to individual distributors of natural gas. The allocation factor is equal to the distributor's share of the sector's total Census division volume in the most recent annual survey of natural gas. This is a census of natural gas distributors in the United States.

Reasons for trying to improve the procedure included a desire for (1) better substitutes in case of edit failures and (2) for relaxing the assumption of constant sector share for proration purposes.

The research from 2003 to 2006 ranged over five categories of options: class of model, allocation to individual distributor, change of response variable, change of predictor variable, and individually tailored models at the distributor level. Some examples are

(1) using a state's share of total monthly volume as the response variable,

(2) changing how the prior month's industrial volume is weighted during estimation,

(3) adding various lag terms to the estimation models, and

(4) a two-stage regression approach,

These various attempts yielded no improvement where gains exceeded the cost of implementation. EIA suspended research at the end of 2006 and resumed it in 2008.

4. Summary of Research in 2008

For the first half of 2008, EIA researched new alternatives. The first alternative considered was setting the imputed value equal to the median monthly volume reported by the distributor for the prior five years for the given sector and delivery mode.

EIA concluded this alternative took insufficient account of the effect of heating degree days on the demand for natural gas. It refined the alternative to adjust for heating degree days. Heating degree days is designed to reflect demand for energy for heating. Heating degree days are a function of the extent the mean daily temperature for a given location is below sixty-five degrees Fahrenheit.

EIA considered various ways including these to adjust for heating degree days:

(1) regress the monthly volume for all the survey respondents in a given state against that state's monthly heating degree days,

(2) regress the monthly volume against the state's heating degree days for each distributor separately,

EIA evaluated performance only for the calendar year 2006. The performance of the alternatives was assessed by Census division and by the months of the year. The major finding was that alternatives that produced individual adjustment models for each distributor tended to perform better than those which produced a single adjustment model for all the distributors in a state.

5. Summary of Research in 2010 and 2011

In 2010 and 2011, EIA turned its attention to an addition to the imputation procedure required by the expansion of the survey to include questions about total system sendout. Briefly, sendout refers to the total volume delivered by a distributor to its customers in a

given time period. EIA used this information to align reported volumes in the commercial and residential sectors with the calendar month. Before this change, distributors in many cases reported for these sectors based on their billing cycle, which rarely aligns with a calendar month.

EIA strove for best use of heating degree day data. EIA also looked at cooling degree day data. Cooling degree days are a function of the extent the mean daily temperature for a given location is above sixty-five degrees Fahrenheit.

In summary, EIA experimented with models involving the heating and cooling degree days at the local weather station nearest a given distributor. Other programs were written to automatically gather and update a heating and cooling degree day file. Still other programs were written to locate the weather station nearest a given distributor. EIA considered models incorporating the current month degree day data as well as others incorporating the data for the current month and the previous month.

The performance measures for these models were mean squared prediction error and the sum of absolute deviations between the actual and predicted volumes. EIA identified models most helpful for supporting the specific needs of the sendout calculations.

EIA next considered whether the promising candidates to support sendout also showed promise for a more general improvement of the imputation procedure. The main concerns arising from examining the models were:

(1) the coefficients in several models could not be uniquely estimated,

(2) local heating and cooling degree day data were available only for a slight majority of distributors, and

(3) no funding was expected in the foreseeable future to obtain local weather data for the distributors who presently lacked it, and

(4) the helpful sendout models employed a two-stage estimation strategy in which large outliers from the first stage were excluded when recalculating the model coefficients in the second stage.

Excluding the outliers from the first stage of calculation raised doubts about whether the models would perform well for data not used to estimate them. The incomplete local heating and cooling degree data made generalizing the solution difficult.

EIA sought a more comprehensive strategy that included the industrial sector, favored unsaturated models as much as possible, and made use of the local heating and cooling degree data.

Result 1. EIA's Performance Measures for 2010-11 Research

let yoi denote the observed natural gas volume for distributor i
let ypi denote the predicted natural gas volume for distributor i
let n denote total number of distributors

Mean Squared Error of Prediction (MSEP)
$MSEP = \Sigma_i (y_{oi} - y_{pi})^2 / n$

Sum of Absolute Deviations (SAD)
$SAD = \Sigma_i |y_{oi} - y_{pi}| / n$

6. Milestones in EIA Research for the Imputation Procedure for Volumes

The first major milestone was accomplished in September 2011 when EIA identified two potential replacement variables for estimating consumption in the industrial sector. One was the natural gas pipeline data published daily by an outside energy market analytics company. EIA obtained its data under license. A second variable was the state quarterly personal income data published by the Bureau of Economic Analysis.

The next major milestone was completed in the first quarter of 2012. EIA looked at the formula adopted in late 2010 to estimate the monthly standard errors for state sector volumes. A component of the formula was estimation of the state sector total. The estimation was via weighted least squares using a single predictor and no intercept. The predictor was the full year state sector volume from the census of two years previously.

EIA wanted to know if the weighted least squares estimate of the state sector volume could serve as a new way to calculate imputed values for distributor volumes. To remove seasonality as a complicating factor, EIA tested an individual model for each month of the year. A sample instance follows below.

Distributor ID	Month	Year	Sector Month Volume	Sector Annual Volume
XYZ	April	2006	116,819	1,155,448
	April	2007	137,397	1,264,307
	April	2008	130,944	1,079,707
	April	2009	60,334	1,180,528
	April	2012	73,148	1,159,577

Result 2. Weighted Least Squares Regression of Monthly Volume on Annual Volume for Example Data

y = volume for sector from monthly survey
x = volume for sector from annual census two years previously
x is also weight in weighted least square model

WLS model: $y = \beta x$; $b = 0.08882$ for example data

The third major milestone was deciding how to incorporate local heating and cooling degree day data as a research option. For the distributors who could be linked to local data, this was very easily done. After internal discussion, EIA adopted this approach for distributors missing a link to local weather data: set the imputed value equal to the prediction using local data if it is available for a given distributor. Otherwise, use the prediction from state data.

The fourth and last major milestone was designing the evaluation strategy for the alternative estimation procedures. EIA chose three performance criteria: the squared differences of the imputed value and the actual reported value, the percentiles of the relative residuals, and the maximum absolute value of the cumulative relative residual. The cumulative relative residual is calculated as (cumulative reported value – cumulative imputed value) / cumulative reported value.

Result 3. EIA's Three Performance Measures for Alternative Imputation Procedures

y_r = reported volume for sector from monthly survey
y_c = imputed volume for sector from current procedure
y_a = imputed volume for sector from alternative procedure

$F(y_{ri})$ = cumulative distribution of y_r from period 1 to period i
$F(y_{ci})$ = cumulative distribution of y_c from period 1 to period i
$F(y_{ai})$ = cumulative distribution of y_a from period 1 to period i

SS = sum of squares measure
RR= relative residual measure
CRR = cumulative relative residual measure

$SS_c = \Sigma_i (y_{ri} - y_{ci})^2$ $\quad$ $SS_a = \Sigma_i (y_{ri} - y_{ai})^2$
$RR_c = (y_{ri} - y_{ci}) / y_{ri}, y_{ri} > 0$ $\quad$ $RR_a = (y_{ri} - y_{ai}) / y_{ri}, y_{ri} > 0$

$CRR_c = (F(y_{ri}) - F(y_{ci})) / F(y_{ri}), F(y_{ri}) > 0$
$CRR_a = (F(y_{ri}) - F(y_{ai})) / F(y_{ri}), F(y_{ri}) > 0$

SS performance measure $= SS_a / SS_c$

RR performance measure = (95^{th} percentile of RR_a - 5^{th} percentile of RR_a / (95^{th} percentile of RR_c - 5^{th} percentile of RR_c)

7. Alternatives for the Current EIA-857 Imputation Protocol for Volumes

EIA selected ten alternatives for the commercial and residential sectors and ten for the industrial sector. They included the weighted least squares and heating degree day concepts discussed previously. Five alternatives in all three sectors consisted of weighting the previous two months of distributor data in various ways to generate the imputed values for volumes.

A separate model was fit for each distributor. The predictions from the models were the alternative imputed values for volume. It decided on a two-staged approach to compute and summarize the three performance criteria for each alternative:

(1) identifying superior alternatives based on data used to fit the models, and
(2) validating the superior alternatives on data for a year not used to fit the models.

Result 4. Five Alternative Imputation Procedures for the Commercial and Residential Sectors

(1) a separate weighted least squares model for each month of the year, $y = \beta_m$, where m is the m^{th} month of the year,

(2) a model predicting the logarithm of volume from heating degree days, $lny = \beta_0 + \beta_{HDD} * HDD$,

(3) a model predicting the logarithm of volume from cooling degree days, $ln(y) = \beta_0 + \beta_{CDD} * CDD$,

(4) a model predicting the logarithm of volume from both heating and cooling degree days, $ln(y) = \beta_0 + \beta_{HDD}*HDD + \beta_{CDD}*CDD$, and

(5) a composite approach using (2) during heating season and (3) during cooling season, $ln(y) = \beta_0 + \beta_{CDD}*CDD$ in cooling season and $ln(y) = \beta_0 + \beta_{HDD}*HDD$ in heating season.

Result 5. Five Alternative Imputation Procedures for the Industrial Sector

(1) a separate weighted least squares model for each month of the year, $y = \beta_m$, where m is the m^{th} month of the year,

(2) a model predicting volume from last month and the same month last year, $y = \beta_0 + \beta_{m-1} * y_{m-1} + \beta_{m-12} * y_{m-12}$,

(3) a model predicting volume from last month, same month last year, and pipeline flow, $y = \beta_0 + \beta_{m-1} * y_{m-1} + \beta_{m-12} * y_{m-12} + \beta_{pipeline} *$pipeline,

(4) a model predicting volume from last month, same month last year, and income, $y = \beta_0 + \beta_{m-1} * y_{m-1} + \beta_{m-12} * y_{m-12} + \beta_{income} *$income, and

(5) a model predicting volume from last month, same month last year, pipeline flow, and income, $y = \beta_0 + \beta_{m-1} * y_{m-1} + \beta_{m-12} * y_{m-12} + \beta_{pipeline} *$pipeline$+ \beta_{income} *$income

Result 6. Five Alternative Imputation Procedures for All Sectors

(1) a model predicting volume as the sum of (0.1 x two months ago) + (0.9 x last month), $y = 0.1 * y_{m-2} + 0.9 * y_{m-1}$,

(2) a model predicting volume as the sum of (0.2 x two months ago) + (0.8 x last month), $y = 0.2 * y_{m-2} + 0.8 * y_{m-1}$,

(3) a model predicting volume as the sum of (0.5 x two months ago) + (0.5 x last month), $y = 0.5 * y_{m-2} + 0.5 * y_{m-1}$,

(4) a model predicting volume as the sum of (0.7 x two months ago) + (0.3 x last month), $y = 0.7 * y_{m-2} + 0.3 * y_{m-1}$, and

(5) a model predicting volume as the sum of (0.9 x two months ago) + (0.1 x last month), $y = 0.9 * y_{m-2} + 0.1 * y_{m-1}$.

Result 7. Summary of Calculation Examples in This Section

SS_a = 925,669,571,602 SSc = 504,504,290,575

SS performance measure = 925,669,571,602 / 504,504,290,575,
ratio greater than one, this measure favors current procedure

95th percentile of RRa = maximum of twelve values in example = 0.1849
5th percentile of RRa = minimum of twelve values in example = -0.1139

95th percentile of RRc = maximum of twelve values in example = 0.0688
5th percentile of RRc = minimum of twelve values in example = -0.2274

RR performance measure = (0.1849 – (-0.1139)) / (0.0688 – (-0.2274))
= 0.2988 / 0.2962, ratio essentially one, this measure favors neither procedure

maximum of | CRRa | = 0.1849
maximum of | CRRc | = 0.0708

CRR performance measure = 0.1849 / 0.0708,
ratio greater than one, this measure favors current procedure

8. Calculation Examples for Evaluating Alternative Procedures

EIA computed three performance statistics for each combination of alternative and distributor. The calculations below serve to illustrate the first performance criterion, the sum of the squared differences between the reported and the imputed values. This is the SS measure.

ID	Month	Actual Value	Current Procedure	Residual	Residual^2
XYZ	1	4,531,024	4,447,114	-83,910	7,040,845,102
	2	3,373,380	3,427,313	53,933	2,908,756,714
	3	2,413,435	2,864,382	450,947	203,353,013,342
	4	1,580,450	1,927,947	347,497	120,753,967,730
	5	937,663	1,031,824	94,161	8,866,343,865
	6	718,559	814,951	96,392	9,291,387,612
	7	794,223	748,110	-46,113	2,126,386,023
	8	762,246	748,508	-13,738	188,732,821
	9	800,668	749,003	-51,665	2,669,316,436
	10	1,242,962	1,525,660	282,698	79,918,088,692
	11	1,882,512	1,820,737	-61,775	3,816,129,001
	12	3,666,364	3,414,230	-252,134	63,571,323,237
				Total	**504,504,290,575**

ID	Month	Actual Value	Alternative Procedure	Residual	Residual^2
XYZ	1	4,531,024	3,693,430	-837,594	701,562,953,605
	2	3,373,380	3,757,696	384,316	147,698,603,292
	3	2,413,435	2,594,732	181,297	32,868,754,806
	4	1,580,450	1,588,234	7,784	60,583,725
	5	937,663	920,522	-17,141	293,808,039
	6	718,559	709,094	-9,465	89,576,766
	7	794,223	736,251	-57,972	3,360,754,859
	8	762,246	698,874	-63,372	4,016,020,243
	9	800,668	764,197	-36,471	1,330,158,688
	10	1,242,962	1,073,698	-169,264	28,650,276,991
	11	1,882,512	1,957,347	74,835	5,600,290,676
	12	3,666,364	3,678,102	11,738	137,789,913
				Total	**925,669,571,602**

The sum of the squared residuals is a smaller the better measure. The ratio of the sum for the alternative procedure to the sum for the current one is an index of performance. A ratio less than one means the alternative procedure is better, and a ratio greater than one means the current procedure is favored.

The calculations below serve to illustrate the second performance criterion, the percentiles of the relative residuals.

ID	Month	Actual Value	Current Procedure	Residual	Relative Residual
XYZ	1	4,531,024	4,447,114	-83,910	0.0185
	2	3,373,380	3,427,313	53,933	-0.0160
	3	2,413,435	2,864,382	450,947	-0.1868
	4	1,580,450	1,927,947	347,497	-0.2199
	5	937,663	1,031,824	94,161	-0.1004
	6	718,559	814,951	96,392	-0.1341
	7	794,223	748,110	-46,113	0.0581
	8	762,246	748,508	-13,738	0.0180
	9	800,668	749,003	-51,665	0.0645
	10	1,242,962	1,525,660	282,698	-0.2274
	11	1,882,512	1,820,737	-61,775	0.0328
	12	3,666,364	3,414,230	-252,134	0.0688

Maximum = 0.0688 Minimum = -0.2274 Difference = 0.2962

ID	Month	Actual Value	Alternative Procedure	Residual	Relative Residual
XYZ	1	4,531,024	3,693,430	-837,594	0.1849
	2	3,373,380	3,757,696	384,316	-0.1139
	3	2,413,435	2,594,732	181,297	-0.0751
	4	1,580,450	1,588,234	7,784	-0.0049
	5	937,663	920,522	-17,141	0.0183
	6	718,559	709,094	-9,465	0.0132
	7	794,223	736,251	-57,972	0.0730
	8	762,246	698,874	-63,372	0.0831
	9	800,668	764,197	-36,471	0.0456
	10	1,242,962	1,073,698	-169,264	0.1362
	11	1,882,512	1,957,347	74,835	-0.0398
	12	3,666,364	3,678,102	11,738	-0.0032

Maximum = 0.1849 Minimum = -0.1139 Difference = 0.2988

The difference between the minimum and maximum relative residuals is a smaller the better measure. The ratio of the difference for the alternative procedure to the difference for the current one is an index of performance. A ratio less than one means the alternative procedure is better, and a ratio greater than one means the current is favored.

In this example, the 5^{th} percentile corresponds to the minimum and the 95^{th} percentile corresponds to the maximum. That was not true in the larger datasets actually used in the research. The calculations below serve to illustrate the third performance criterion, the maximum absolute value of the cumulative relative residual.

The maximum absolute value of the cumulative relative residuals serves as a fit of distribution measure—the smaller the maximum the better. The ratio of the maximum for the alternative procedure to the maximum for the current one is an index of performance. A ratio less than one means the alternative procedure is better, and a ratio greater than one means the current is favored.

ID	Month	Cumulative Actual	Cumulative Current	Absolute Cumulative Relative Residual
XYZ	1	4,531,024	4,447,114	0.0185
	2	7,904,404	7,874,427	0.0038
	3	10,317,839	10,738,809	0.0408
	4	11,898,289	12,666,756	0.0646
	5	12,835,952	13,698,580	0.0672
	6	13,554,511	14,513,531	0.0708
	7	14,348,734	15,261,641	0.0636
	8	15,110,980	16,010,149	0.0595
	9	15,911,648	16,759,152	0.0533
	10	17,154,610	18,284,811	0.0659
	11	19,037,122	20,105,549	0.0561
	12	22,703,486	23,519,779	0.0360

Maximum= 0.0708

ID	Month	Cumulative Actual	Cumulative Alternative	Absolute Cumulative Relative Residual
XYZ	1	4,531,024	3,693,430	0.1849
	2	7,904,404	7,451,126	0.0573
	3	10,317,839	10,045,859	0.0264
	4	11,898,289	11,634,092	0.0222
	5	12,835,952	12,554,614	0.0219
	6	13,554,511	13,263,709	0.0215
	7	14,348,734	13,999,960	0.0243
	8	15,110,980	14,698,834	0.0273
	9	15,911,648	15,463,030	0.0282
	10	17,154,610	16,536,728	0.0360
	11	19,037,122	18,494,076	0.0285
	12	22,703,486	22,172,178	0.0234

Maximum=0.1849

When this performance measure was calculated on the real data, the cumulative relative residual for the first time period was excluded as it was judged to be volatile and likely to not properly reflect the overall data.

9. Summarizing Performance Measures for Alternative Procedures

The examples in the preceding section show how the individual performance measures were calculated by EIA. These had to be summarized for each alternative. The results from the preceding section's examples will be used to show how EIA accomplished this for its research.

Result 8. Key for Alternative Procedures

Category	Label
Weighted least squares for each month of the year	WLS
Logarithm of volume predicted from heating degree days	NOAA1
Logarithm of volume predicted from cooling degree days	NOAA2
Logarithm of volume predicted from both degree days	NOAA3
Logarithm of volume predicted by heating or cooling season	NOAA4
Weight two months ago by 0.1, last month by 0.9	L01
Weight two months ago by 0.2, last month by 0.8	L02
Weight two months ago and last month equally	L05
Weight two months ago by 0.7, last month by 0.3	L07
Weight two months ago by 0.9, last month by 0.1	L09
Predict volume from last month, same month last year	Ind1
Predict volume from last month, same month last year, pipeline	Ind2
Predict volume from last month, same month last year, income	Ind3
Predict volume from last month, same month last year, pipeline, income	Ind4

Result 9. Summarizing Performance Measures

SS measure from example data in previous section
925,669,571,602 / 504,504,290,575 = 1.8348

RR measure from example data in previous section
0.2988 / 0.2962 = 1.0088

CRR measure from example data in previous section
0.1849 / 0.0708 = 2.6116

Summary measure for one distributor from example data in previous section
(1.8348 + 1.0088 + 2.6116) / 3 = 1.8184, favoring the current procedure overall

Summary measure for one distributor as actually calculated in research
d_{1i} = number of months of data available for SS measure for distributor i
d_{2i} = number of months of data available for RR measure for distributor i
d_{3i} = number of months of data available for CRR measure for distributor i

Summary measure for distributor i = $SM_i = (d_{1i}*SS_i + d_{2i}*RR_i + d_{3i}*CRR_i) / (d_{1i} + d_{2i} + d_{3i})$

Overall summary measure as calculated for alternative with n distributors

$$SM_a = \{[(\Sigma_i\, d_{1i} * \Sigma_i\, SS_i) + (\Sigma_i\, d_{2i} * \Sigma_i\, RR_i) + (\Sigma_i\, d_{3i} * \Sigma_i\, CRR_i)] / [\Sigma_i\, d_{1i} + \Sigma_i\, d_{2i} + \Sigma_i\, d_{3i}]\} / n$$

Substituting the results for the examples in the preceding section and taking a simple average since they are all based on twelve months, the summary performance index is (1.8348 + 1.0088 + 2.6116) / 3 = 1.8184, favoring the current procedure overall.
Since the maximum absolute value of the cumulative relative residual excluded the earliest month, EIA did not take a simple average of the three measures in actual practice.

Result 10. Overall Summary Measures for Alternative Procedures

Alternative	Commercial Sector	Residential Sector	Industrial Sector
WLS	0.916	1.204	2.660
L01	5.308	9.661	1.563
L02	4.882	8.845	1.462
L05	3.785	6.735	1.220
L07	3.212	5.616	1.110
L09	2.760	4.738	1.042
NOAA1	1.210	2.540	
NOAA2	7.318	10.940	
NOAA3	3.806	8.603	
NOAA4	3.173	4.589	
Ind1			0.789
Ind2			0.779
Ind3			0.719
Ind4			0.749

Superior alternatives have overall summary measures less than 1.0. Measures highlighted in red were tested on 2011data.

The overall summary measure for an entire alternative was the weighted sum of the overall measures for all the distributors it included. The tables on the preceding page show the overall summary measures for the fourteen alternatives explored.

Alternatives superior to the current procedure, as earlier explained, score below one. The overall summary measures led to three conclusions:

(1) The WLS model for each month/distributor was superior for commercial sector.
(2) Of the alternatives tested, none were superior for residential sector.
(3) Models Ind1 through Ind4 were all superior for industrial sector.

The WLS model for the commercial sector was selected for validation on 2011 data. Ind1 and Ind3 for the industrial sector were also selected. EIA chose Ind3 because it was the best performing alternative even though its adoption would require adding state personal income data to the imputation procedure for volumes.

Result 11. Overall Summary Measures for Alternative Procedures Results for 2011 Data

Alternative	Commercial Sector	Industrial Sector
WLS	1.625	
Ind1		2.141
Ind3		2.232

When the WLS, Ind1, and Ind3 alternatives were tested on 2011 survey data, the overall performance measures respectively were 1.625, 2.141, and 2.232. Hence, none of the three alternatives proved superior in validation. The biggest reason in all three cases was failure on the SS criterion, the sum of the squared differences between the reported and the imputed values.

10. EIA's Options for Recommendations

The results described in the previous section were obtained in May 2012. With one month remaining in the research schedule, EIA debated what to recommend. The options were:

(1) maintaining the current procedure and moving on to a new project or
(2) maintaining the current procedure while extending the alternative procedure research

During June 2012, EIA saw two replacement projects it might implement under option (1):

(1) it could replace the procedure research with a project to better integrate the existing price and volume procedures for the survey, or

(2) it could replace the procedure research with a project to integrate the survey sample selection code with the current procedure for the imputed values for volumes.

It also saw several ways it might implement option (2):

(1) it could try to find out why the preferred procedures failed the SS criterion,
(2) it could ask the other parts of EIA for advice on other alternative procedures,
(3) it could explore cold deck imputation as an alternative procedure, or
(4) it could explore a new volume editing strategy for the survey.

In weighing the two basic options, EIA considered the following points:

(1) four distinct research efforts since 2003 meant the search for an alternative procedure has received a fair share of attention,

(2) corrections to the SS issues raised in the 2011 cross validation might introduce other problems,

(3) how likely was it other parts of EIA, who were not as experienced with the current procedure as the one carrying out the research, could propose a new alternative, and

(4) the integration ideas offered a strong prospect of sure and early pay off.

EIA decided to recommend maintaining the current procedure. During the summer of 2012, EIA also implemented one of the integration options, the one to integrate the survey sample selection code with the current imputation procedure for volumes.

Early in 2013, EIA investigated (1) the potential of small area estimation to increase the precision of state sector volume estimates, (2) ways to achieve further reductions in sample size for the survey, and (3) treating missing or questionable responses as if they were out of sample by using a classical ratio estimator to impute for them. The last option proved the most promising and is being coded and tested. It allows EIA to unify existing modules for volume and price calculations.

Another key advantage of the classical ratio estimator for imputation is that it provides a compact and convenient calculation of how the imputation affects the overall variance of the published volumes. Since the classical ratio estimator cannot be applied until the majority of the reported values for the month of interest are available, it is no longer possible to have a leading indicator of the eventual monthly volume from early on in a given reporting month.

EIA is presently researching how best to combine the classical ratio imputation approach with the forecasting approach so future surveys can retain some version of a leading indicator feature. EIA believes the search for an improved imputation procedure affirms the following enduring lessons of research in general:

(1) we should rarely consider the story finished,
(2) we should build on the best of the past,
(3) we should use many measures of success,
(4) we should be firm in resolve and flexible in pursuit, and
(5) we should consider the cost of implementation before declaring success.

11. Postscript

In May 2013, as part of evaluating progress on research into how best to preserve the leading indicator feature in imputation procedures, the EIA staff pointed out the indexes computed during the 2012 research were equally weighted regardless of the volume reported by the distributor. It recommended recalculation based on weighting by volume. EIA did so.

The marginally better alternative in the commercial sector in the unweighted case was marginally worse in the weighted one. There were no superior alternatives for the residential sector in either the weighted or unweighted cases. While the unweighted case for the industrial sector suggested some superior alternatives, in the weighted case the alternatives were all worse, from marginally so to clearly so.

The weighted results would have led to the same conclusion: retain the current procedure until the out of sample approach based on the classical ratio estimator can be implemented.