WP. 30
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (v): Editing based on results

THE DEVELOPMENT OF A MACRO EDITING APPROACH

Invited Paper

Prepared by J. Meyer, SAIC, J. Shore, P. Weir, and J. Zyren, U.S. Energy Information Administration, United States[1]

I. INTRODUCTION

1. The U.S. Energy Information Agency (EIA) collects and disseminates state, regional, and national petroleum market data. Through a family of surveys, the EIA collects supply data on a monthly basis that allows the agency to estimate demand. This demand estimate, called product supplied, for a given region is calculated from the components of supply as reported in the surveys, i.e., production within the region plus foreign net imports plus net receipts from other regions plus the change in inventory. These data are collected from a census of the "primary level" population of the supply chain, which is defined as bulk storage of 50,000 barrels or more. Some of these components of product supplied can vary widely from month to month, even though actual consumption may not vary in that manner. Respondent level data, such as that reported by pipeline companies, may legitimately vary significantly, making it hard to detect misreporting. However, it is expected that when the data across surveys and respondents are combined, some of this variation will be smoothed. That is not always the case. As a result, EIA is trying to develop methods for checking aggregate data for outlier results to supplement its respondent-level or micro-editing. This paper explores current work at EIA in developing a new approach for macro-editing survey data and describes its use in particular for verifying monthly aggregate distillate fuel oil product supply (demand) survey data, which is published in the report *Petroleum Supply Monthly*.

2. The tools being developed provide point and interval forecasts of next-month distillate demand using econometric time series or ARMA models based on the fit to historical data. Three models are developed for each area (nation-wide and regional) of interest: a base model using trend and seasonal indicator variables; an ARMA model; and an explanatory model which is the base model plus exogenous market variables (e.g. precipitation, heating degree days, etc.) The forecast intervals derived from these models are then compared with preliminary survey aggregate estimates to help detect potential outlier data.

3. This paper briefly discuss the data survey process, the procedures used in model development, a description of typical models, the forecast variability of the models, and the usefulness of these confidence intervals for macro-editing of survey data.

[1] This report is released to inform parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily of the U.S. Energy Information Administration

II. THE PROBLEMS IN THE MONTHLY SUPPLY REPORTING SYSTEM

A. The Monthly Petroleum Supply Reporting System

4. The Petroleum Supply Reporting System (PSRS) represents a family of data surveys, data processing systems, and publication systems. It comprises three parts: the Weekly Petroleum Supply Reporting System (WPSRS); the Monthly Supply Reporting System (MPSRS); and the revised monthly estimates made available in the Petroleum Supply Annual. Data for the WPSRS are collected through six survey forms reported by a sample of all petroleum companies who report crude oil and petroleum product stocks, refinery inputs and production, and crude oil and petroleum product imports.

5. The Monthly Supply Reporting System in comparison provides a more comprehensive set of information from a census of the population. Nine surveys comprise the MPSRS which collect detailed refinery/blender, natural gas plant, and oxygenate operations data; refinery/blender, bulk terminal, natural gas plant, oxygenate producers and pipeline stocks data; crude oil and petroleum product imports data; and data on movements of petroleum products and crude oil between Petroleum Administration for Defense (PAD) Districts.

B. The Limitation of Micro Editing and the Need for an Edit Approach at the Aggregate Level

6. Data collected on these surveys are subjected to numerous edits and imputation is performed for failed unresolved data items for all but two of the surveys. Imputation is not performed for the survey that collects data on the imports of crude oil and petroleum products, nor is it performed for the survey that collects data on tanker and barge movements of crude oil and petroleum products among the PAD Districts. The data on these two surveys are not imputed because of their highly variable nature.

7. Efforts are currently underway to improve the edits in the MPSRS, as well as the imputation methodology. However, because of the legitimate variability of the data at the respondent-cell level, it is still expected that misreporting in some situations will be difficult to identify even with the improved edits and tools. It is also reasonable to expect that when the data across surveys and respondents are combined, some of this micro-variation would be smoothed. That, however, has not always been the case and it is believed that this is only partially due to the limitations of micro edits. The other area of concern is that even though the survey respondent set is intended to represent a census of the population, the respondent set for particular products or components may fall short of the population at particular times during the data collection. The petroleum industry and associated population is constantly evolving, adjusting its business solutions and approaches with customer needs and regulatory requirements. For example, blenders and importers may enter and exit the market abruptly. While attempts are made to track the births and deaths, the respondent set may lag behind the industry. As a result of all of these shortcomings, EIA is trying to develop methods for checking aggregate data for outliers or macro editing to supplement its micro editing and identify population shortfalls.

III. THE APPROACH CONSIDERED FOR MACRO EDITING

A. General Specification of Functional Form

8. The approach to macro editing that EIA is developing to identify potential aggregate outliers consists of comparing the survey result against a model-generated forecast interval. This interval is calculated based on the model-based point estimate bounded on either side by the forecast error. A deviation of the survey estimate outside the forecast interval would prompt further examination of the underlying survey data in the last stages of survey processing. In order to obtain some indication of model reliability, particularly during model development, three separate monthly models were estimated for each fuel type, identified as the Base, ARMA, and Supplemental Models, and are defined below.

9. The underlying principle upon which these models are based is that demand elasticity for distillate is very small; in other words, demand is basically a long-term upward trend with seasonality imposed. This supposition is the basis for the first two model types, the Base Model which uses a deterministic trend and seasonal variables to model demand, and the ARMA model which uses Box-Jenkins methodology.

10. The **Base Model**, which estimates demand based on trends and seasonal factors, is expressed in general terms as:

$$Demand_t = \alpha_0 + \beta_1 Trend + \sum_{k=2}^{12} \beta_k D_k + \gamma_j Shift_j + AR\ or\ MA\ \ terms \qquad \textbf{(1)}$$

where $Demand_t$ is the reported product supply of distillate, in thousands of barrels per day for month *t*;
Trend is a linear trend beginning in the first month of the estimation period;
D_k, $k = 2, 3, \ldots, 12$ are 11 monthly seasonal dummy variables;
$Shift_j$, $j = 1, \ldots, n$ are dummy variables indicating a shift in either level or trend;
AR or MA terms are either autoregressive (AR) or moving average (MA) terms, used as necessary to make the estimation residuals random.
$\alpha, \beta, \ldots,$ are coefficients to be estimated by the model.

11. Alternatively, the **ARMA Model** is the Box-Jenkins approach of utilizing AR and MA to capture the variation and seasonal pattern of demand. The form of this model is:

$$Demand_t = \alpha_0 + \beta_1 Trend + \gamma_j Shift_j + \sum_{m=1}^{n} \delta_m ARMA_m \qquad \textbf{(2)}$$

where ARMA are autoregressive (AR) and/or moving average (MA) terms and the other variables are as defined in Equation 1 above.

12. The **Supplemental Model** is an enhancement of the Base Model utilizing explanatory (exogenous) variables such as weather, employment, industrial productivity, and fuel price to improve model fit and forecasting ability. The effect of these additional variables on demand is generally small since any normal seasonal or long-term pattern variation is captured by the deterministic trend and seasonal variables. The general form of this model is:

$$Demand_t = \alpha_0 + \beta_1 Trend + \sum_{k=2}^{12} \beta_k D_k + \gamma_j Shift_j + \sum_i \phi_i Exog_i + AR\ or\ MA\ terms \qquad \textbf{(3)}$$

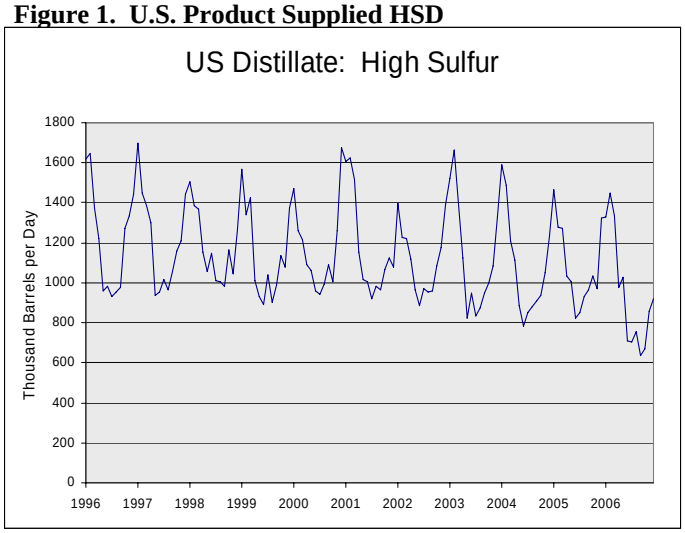
where $Exog_i$ represents explanatory variables important in explaining demand, and the other variables are as defined in Equation 1. A description of the exogenous variables used in these models is provided in the Appendix. The demand and exogenous variables used in this study were tested for stationarity; ADF test results could not reject stationarity for all data series.

13. The modeling effort first examined total US consumption of No. 2 distillate fuel oil, testing the alternative forecasting approaches to high sulfur, low sulfur, and total distillate demand. This was followed by a more detailed examination of consumption within four regions of the country (PADD's 1, 2, 3, & 5). The estimation period chosen for all of the models was January 1996 – December 2006. The discussion in this paper will focus on the National level models; details on the regional models are available on request.

B National Level Models

i. High Sulfur Distillate (HSD)

Figure 1. U.S. Product Supplied HSD



14. Monthly supply figures for National level HSD displayed in Figure 1 show there is clearly a strong seasonal influence, driven in large part by residential heating demand in the Northeast, and indicate a modest negative trend, particularly in recent years. HSD demand was estimated using the three modeling approaches described above; the estimation results for the best of each of these models are provided in Table 1. In the Base model, the *Trend* variable loses significance when a trend-shift dummy variable beginning in February 2003 is introduced (*T_FEB03*). Although the monthly dummies for the summer months (May – September) are not significant, they are retained in the model for the sake of completeness. In comparison, the best ARMA model includes a level-shift in July of 2001 (L_*JUL01*) in addition to a trend-shift. The best Supplemental model includes, along with trend- and level-shifts, two exogenous variables that proved significant: the Heating Degree-Day deviation from population-weighted long-term normal for entire US (*HDD_DEV*) (based on data released by the Climate Prediction Center of the National Weather Service), and the ratio of the average monthly spot-price of No.2 fuel oil to that of natural gas (*PR_RAT)* (based on data released by EIA).

Table 1: US Total HSD Models

Base				ARMA				Supplemental			
Var.	**Coef.**	**SE**	**Prob.**	**Var.**	**Coef.**	**SE**	**Prob.**	**Var.**	**Coef.**	**SE**	**Prob.**
C	959.85	33.55	0.000	C	1207.18	37.55	0.000	C	1047.59	45.16	0.000
JAN	588.34	39.89	0.000	T_JAN03	-3.53	1.60	0.029	T_JAN03	-2.78	0.76	0.000
FEB	501.77	39.12	0.000	L_JUL01	-98.16	45.93	0.035	L_OCT01	-60.83	21.06	0.005
MAR	408.51	34.69	0.000	AR(12)	0.38	0.07	0.000	JAN	633.79	40.39	0.000
APR	193.77	34.12	0.000	AR(1)	0.49	0.06	0.000	FEB	557.48	36.91	0.000
MAY	50.40	34.27	0.144	AR(4)	-0.23	0.05	0.000	MAR	435.15	35.60	0.000
JUL	1.72	34.27	0.960					APR	225.53	31.13	0.000
AUG	15.45	34.13	0.652					MAY	61.09	31.67	0.056
SEP	43.55	34.74	0.212					JUL	-2.29	31.70	0.943
OCT	141.89	39.15	0.000					AUG	12.23	30.87	0.693
NOV	205.67	39.89	0.000					SEP	49.31	35.52	0.168
DEC	412.84	40.72	0.000					OCT	143.46	36.76	0.000
T_FEB03	-5.65	1.35	0.000					NOV	227.08	39.87	0.000
AR(1)	0.24	0.09	0.010					DEC	429.93	37.16	0.000
AR(2)	0.27	0.09	0.004					HDD_DEV	0.99	0.13	0.000
MA(3)	0.23	0.10	0.017					PR_RAT(-1)	-48.48	23.84	0.044
								AR(7)	-0.22	0.10	0.022
								AR(1)	0.16	0.09	0.092
								MA(2)	0.19	0.09	0.043
	Adj. R^2	**AIC**	**D-W**		**Adj. R^2**	**AIC**	**D-W**		**Adj. R^2**	**AIC**	**D-W**
	0.863	11.9294	1.961		0.784	12.3185	2.091		0.898	11.6574	1.975

Estimation Period: January 1996 – December 2006 (132 observations)

15. The coefficients of the additional explanatory variables in the supplemental model demonstrate the expected relationships with high-sulfur distillate demand: larger deviations from U.S. normal temperatures, as indicated by the HDD variable, result in an increase in fuel oil demand; conversely, high spot oil prices relative to natural gas in the preceding month result in a reduction in demand, as industrial users may take advantage of fuel-switching capabilities. Additionally, a comparison of the Adjusted R^2 and the AIC of each of the considered models suggests that the inclusion of the exogenous variables significantly adds to the explanatory power of the Supplemental model.

Table 2: US Total LSD Models

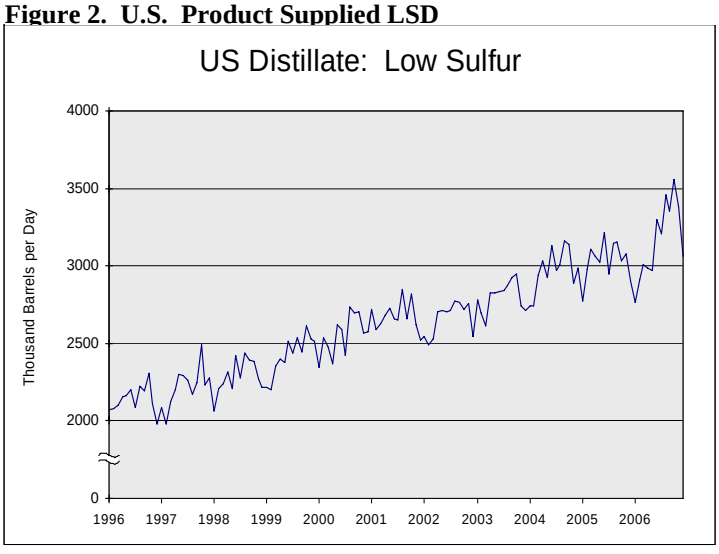
Base				ARMA				Supplemental			
Var.	**Coef.**	**SE**	**Prob.**	**Var.**	**Coef.**	**SE**	**Prob.**	**Var.**	**Coef.**	**SE**	**Prob.**
C	2006.83	49.32	0.000	C	1877.54	198.81	0.000	C	2066.32	63.13	0.000
@TREND	7.07	0.96	0.000	@TREND	9.12	1.68	0.000	@TREND	5.69	1.31	0.000
JAN	-195.89	38.18	0.000	L_NOV01	-132.16	46.36	0.005	JAN	-191.08	38.21	0.000
FEB	-178.70	35.53	0.000	T_DEC05	11.94	6.19	0.056	FEB	-169.79	35.19	0.000
MAR	-120.08	33.09	0.000	AR(12)	0.57	0.08	0.000	MAR	-113.07	33.15	0.001
APR	-81.60	43.00	0.060	AR(2)	0.22	0.08	0.007	APR	-76.09	42.72	0.078
MAY	-83.87	38.60	0.032	MA(3)	0.20	0.09	0.032	MAY	-77.41	38.09	0.045
JUL	-103.50	38.56	0.008					JUL	-100.67	37.95	0.009
AUG	2.56	43.15	0.953					AUG	8.36	42.76	0.845
SEP	-19.56	33.03	0.555					SEP	-15.32	32.74	0.641
OCT	41.28	35.54	0.248					OCT	51.95	35.12	0.142
NOV	-103.21	38.22	0.008					NOV	-90.30	38.16	0.020
DEC	-190.46	36.88	0.000					DEC	-178.90	38.51	0.000
L_NOV01	-117.27	31.01	0.000					L_NOV01	-113.44	31.99	0.001
T_MAY98	3.25	1.25	0.010					T_MAY98	4.66	1.56	0.004
AR(3)	0.21	0.09	0.020					TSI_F(-1)	10.52	6.46	0.106
AR(10)	-0.36	0.10	0.000					AR(10)	-0.32	0.10	0.001
								MA(3)	0.25	0.10	0.009
	Adj. R^2	**AIC**	**D-W**		**Adj. R^2**	**AIC**	**D-W**		**Adj. R^2**	**AIC**	**D-W**
	0.948	11.7234	1.704		0.923	12.0484	1.647		0.949	11.7168	1.791

Estimation Period: January 1996 – December 2006 (132 observations)

ii. Low Sulfur Distillate (LSD)

16. Figure 2 depicts US monthly supply figures for LSD between January 1996 and December 2006. Unlike the HSD demand discussed above, there is a pronounced long-term trend. As with HSD, there is also an indication of seasonality, but much less pronounced. The estimates of the three best LSD models are provided in Table 2. In addition to the *Trend* and seasonal variables of the Base model, the Supplemental model includes a measure of the demand for freight transport (*TSI_F*) represented by the Transportation Services Index for

Figure 2. U.S. Product Supplied LSD



freight (published monthly by the Department of Transportation). The coefficient of this (lagged and de-trended) exogenous variable is consistent with expectations, indicating that an increased demand for freight transport translates into an increase in demand for LSD.

C. Model & Forecast Evaluation

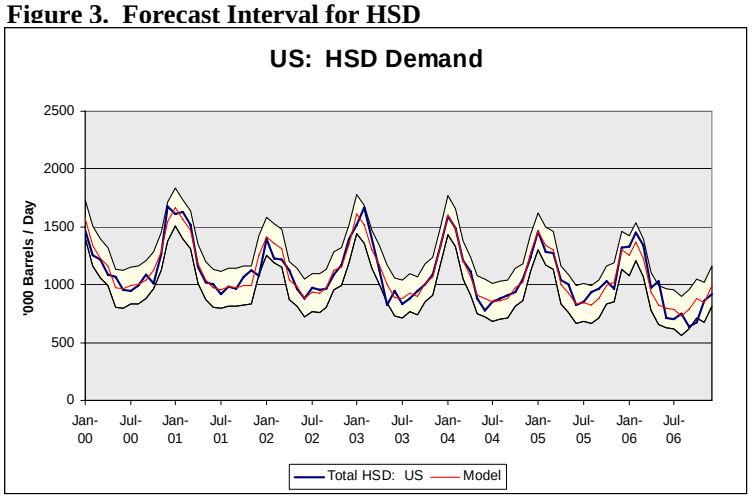
17. In-sample forecast comparisons for the three HSD models can be found in Table 3, which summarizes the in-sample forecast root mean square (RMSE), mean absolute deviation (MAE), mean absolute percent error (MAPE) statistics, the Theil inequality coefficient, and the bias, variance and covariance proportions for the three models in the estimation sample period. By all of these statistics, the Supplemental model has the best in-sample forecasting ability. The scale-sensitive statistics (RMSE, MAE, and MAPE) show that the ARMA model has the largest forecast errors while the Supplemental model has the smallest; the scale insensitive Theil U statistic confirms the same. Decomposition of the Theil U into bias, variance and covariance shares shows that none of the models have significant forecast bias. Typical of one-month-out forecasts, all of the models show a very high covariance proportion, indicating that there is little systematic forecast error.

Table 3: In-Sample One-Month-Out Forecast Evaluation Statistics

	HSD Models			LSD Models		
	Base	ARMA	Supplemental	Base	ARMA	Supplemental
RMSE	83.4753	109.3852	71.2241	74.7377	94.8447	74.3591
MAE	63.6661	84.9583	54.9918	60.2242	76.4177	59.8518
MAPE	5.7559	7.5916	5.1258	2.2761	2.9328	2.2610
Theil U	0.0362	0.0474	0.0308	0.0141	0.0179	0.0140
Bias P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Var P	0.0313	0.0539	0.0269	0.0118	0.0179	0.0131
Covar P	0.9687	0.9461	0.9731	0.9882	0.9821	0.9869

18. Table 3 also provides the corresponding statistics for the three LSD models. The forecast statistics of the ARMA model are significantly worse than those of either the Base or Supplemental model, while the Supplemental model appears to have marginally better statistics than the Base model. As Mod1 has the best in-sample forecasting ability, it is selected as the model upon which the LSD forecast intervals are based. As the primary goal of this study is to determine whether a given month's survey estimate is consistent with expected market conditions, it is necessary to construct an uncertainty interval about the estimate. This interval is defined as ±2 forecast standard errors from the estimate generated by the best model, identified above as the Supplemental model. The in-sample performance of this model is depicted in the Figures 3 and 4, in which actual HSD and LSD demands are plotted against the forecast intervals generated from the Supplemental Models. The results demonstrate a fairly close correspondence between the survey data and the model-generated estimate over the course of the estimation period, with few deviations beyond the forecast and 95 percent confidence interval.

Figure 3. Forecast Interval for HSD



D. Out-of-Sample forecasts

19. Out-of-sample behavior of the HSD model is displayed in Figure 5, where the Supplemental model was used to generate a series of one-month-ahead forecasts from January through November 2007. The model tends to overestimate demand particularly during the latter half of 2007, suggesting the occurrence of a structural shift in HSD demand too close to the end of the estimation period to be adequately captured by the model. Out-of-sample performance of the LSD model is illustrated in Figure 6. This model underestimates LSD demand, in the latter part of 2007 supporting the likelihood of a market shift between HSD and LSD demand in recent months.

20. In the United States, as a result of environmental regulations, several market segments (marine, locomotive, and non road such as construction equipment) were required to shift from high sulfur distillate fuel use to low sulfur distillate fuel in 2007. It is likely that this market shift is behind the out-of-sample performance discussed in this section.

Figure 4. Forecast Interval LSD

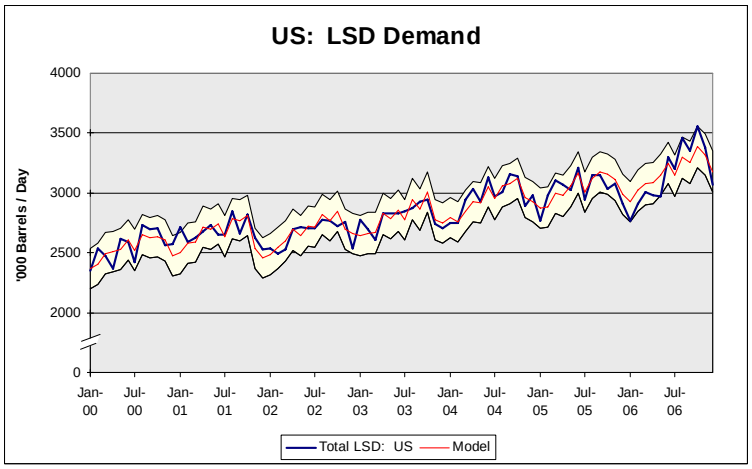
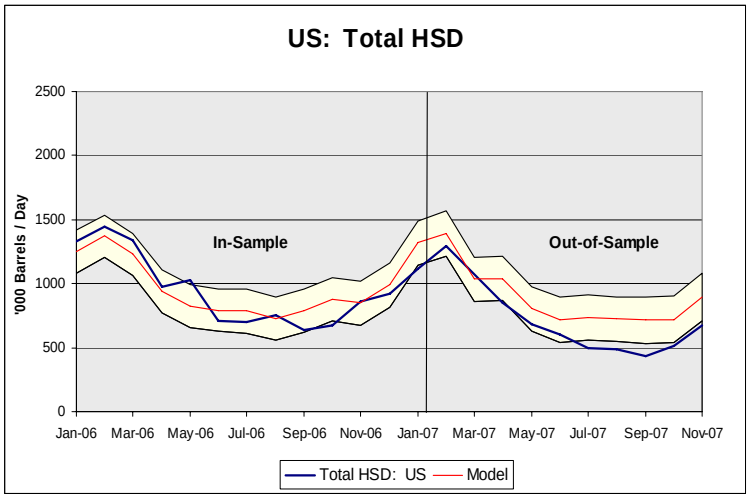
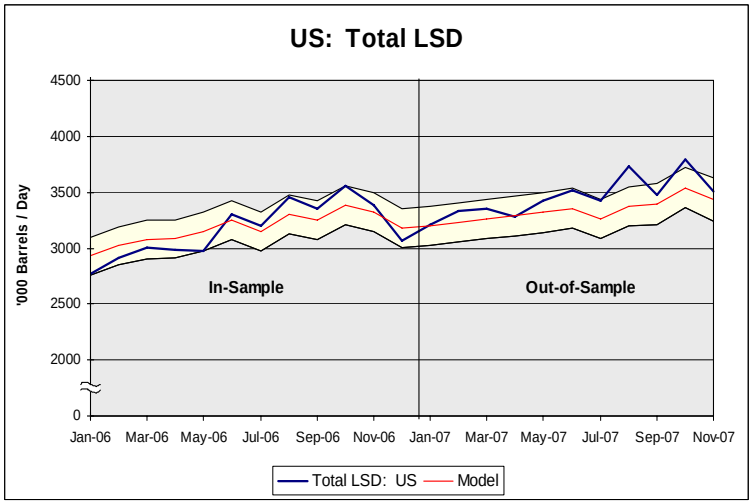


Figure 5. Out-of-sample HSD Forecast Performance



21. By way of confirmation, Figures 7 and 8 depict the performance of another National-level model, estimated separately on the aggregate (HSD +LSD) demand for distillate over the same period. In this model, the significant explanatory variables include both the weather-related HDD_DEV factor and the market-driven TSI_F measure that had proven significant in the models of the constituent components. The survey data stays fairly consistently within the confidence band about the forecast, suggesting that the model is able to provide the analysts with a reasonable range of expectations against which the respondent data can be compared.

Figure 6. Out-of-sample LSD Forecast Performance

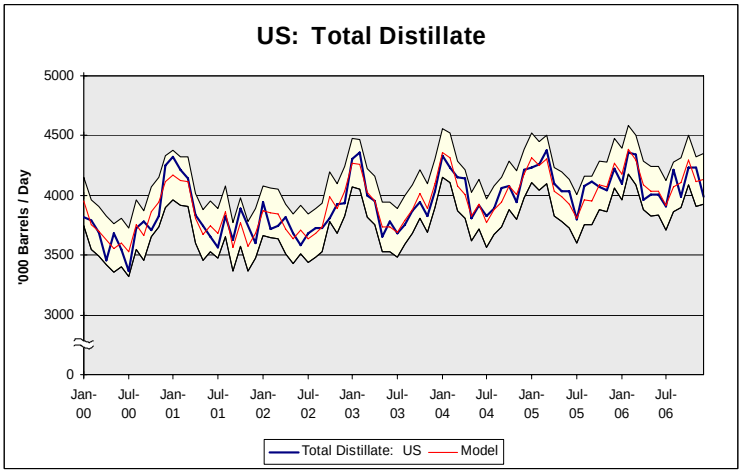


E. Extensions beyond the National Level

22. The goal of this research has been to identify and apply short term

forecasting methodologies to aggregate survey data at different geographic levels in order to detect, *a priori,* the possible presence of outliers in respondent data. To that end, over 40 separate models, covering five regions (four PADD's + National), three distillate types (HSD, LSD, + Total), and three modeling specifications (Base, ARMA, and Supplemental) were also estimated and evaluated. The behavior of the models expressed at the National level is mirrored, to a degree, in the regional models. These specific geographic area models demonstrated sensitivity to exogenous factors that are consistent with expectations; for example, PADD 1 (Northeast) demand for distillate is strongly influenced by residential heating requirements; and PADD 2 (Midwest) demand is influenced by precipitation levels, consistent with the predominance of agricultural uses in this region. The Appendix provides a tabular summary of the explanatory variables that proved significant in each of the geographic area models.

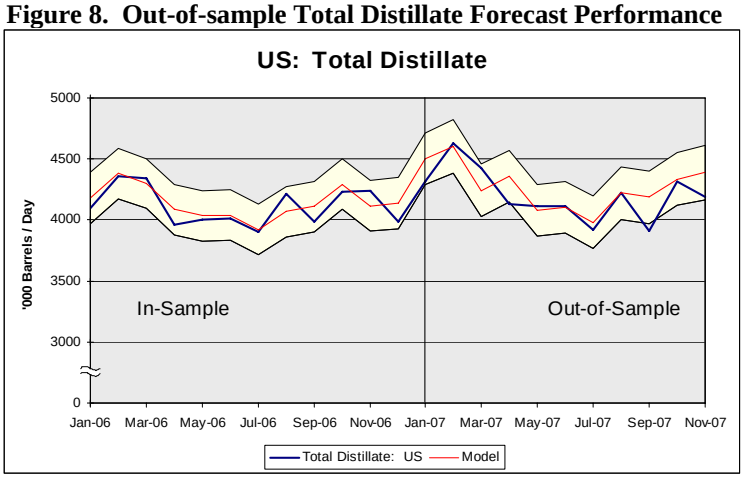
Figure 7. U.S. Product Supplied Total Distillate



IV. Summary & Conclusions

23. The underlying purpose of this study has been to develop a simple and generally applicable macro editing tool for petroleum product supplied survey data that are used to estimate demand. The U.S. level models described here represent an initial effort to identify possible reporting issues or frame deficiencies in the aggregate survey data and to guide subsequent editing and imputation activities. Models were also developed at lower geographic levels which were shown sensitive to exogenous factors consistent with expectations for the specific geographic area.

Figure 8. Out-of-sample Total Distillate Forecast Performance



24. The models generate a one-month-out estimate of expected levels of high-sulfur and low-sulfur distillate demand, bounded by a confidence interval. If reported survey data fall outside this confidence interval, this may be an early indication of reporting errors or other data issues that would motivate further investigation and adjustments. Alternatively, a persistent over- or under-estimation of results may be an indication of recent structural or market shifts that were not captured in the estimation process and that may be confirmed by comparing the behavior of the HSD and LSD model results.

25. The intention of future work is to refine these models and expand our capabilities to address other petroleum products over the coming months. These analytical tools will enhance the accuracy and timeliness of our data products, increase our process efficiency, and reduce the need for *post-hoc* revisions of published data. While the results are a promising first step to our effort to preemptively asses the quality of reported data from a macro viewpoint, recent shifts within distillate between high and low sulfur, illustrated by the out-of-sample forecast performance, indicate that additional work is needed to adequately account for current and potential future structural changes.

V. **Appendix**

	Exogenous Variables Used in Supplemental Distillate Models														
	PADD 1			PADD 2			PADD 3			PADD 5			NATIONAL		
	HSD	**LSD**	**TOT**	**HSD**	**LSD**	**TOT**	**HSD**	**LSD**	**TOTL**	**HSD***	**LSD**	**TOT**	**HSD**	**LSD**	**TOT**
HDD_DEV	*X*	*X*	*X*										*X*		*X*
PRECIP_DEV				*X*		*X*					*X*	*X*			
EMP_TRN								*X*	*X*						
IP_MFG_D							*X*								
TSI_FREIGHT					*X*			*X*						*X*	*X*
SPOT_RATIO													*X*		

* No Supplemental Models for HSD Proved Significant in PADD 5

HDD_DEV	Population-Weighted Heating Degree-Days: Deviation from Normal
PRECIP_DEV	Area-Weighted Precipitation: Deviation from Long-Term Normal
EMP_TRN	Employment in Transportation Industries
IP_MFG_D	Index of Industrial Production for Durable Goods
TSI_FREIGHT	Transportation Services Index for Freight
SPOT_RATIO	Average monthly spot price ratio: No.2 Fuel Oil / Natural Gas

Seasonal, Trend, & Dummy Variables

- Trend: Monotonically increasing, beginning with 0 in the first observation
- Month: 11 variables, 1 for the corresponding month, 0 otherwise.
- L_MMMYY: Level shift variable: 0 until MMMYY, 1 thereafter
- T_MMMYY: Trend shift variable: 0 until MMMYY, Trend thereafter
- LX_MMMYY: Spike variable: 1 at MMMYY, 0 otherwise.