Final Project for DNSC 6219 — Time Series Forecasting — Spring 2017

Instructor: Refik Soyer

Group: Kelly Berdelle, Daniel Chen, Kevin Fitzgerald, Aida Roxas, Xing Zhang

1. **Introduction and Overview**

We analyze time-series data on total monthly electricity generation by fuel in the United States from 2001 to 2016. This data is maintained by the US Energy Information Administration (EIA) and is available at <https://www.eia.gov/electricity/data/browser/>.

The EIA maintains the State Energy Data System (SEDS) that tracks energy usage in the United States. According to their website, “EIA's goal in maintaining SEDS is to create historical time series of energy production, consumption, prices, and expenditures by state that are defined as consistently as possible over time and across sectors for analysis and forecasting purposes.” The data contains monthly electricity generation by various sources of energy and the aggregate electricity generation in the U.S. These energy sources include coal, natural gas, nuclear, hydroelectric, wind, solar, liquid petroleum, geothermal, biomass, wood, and others. All values are measured in thousands of Megawatt hours (or Gigawatt-hours).

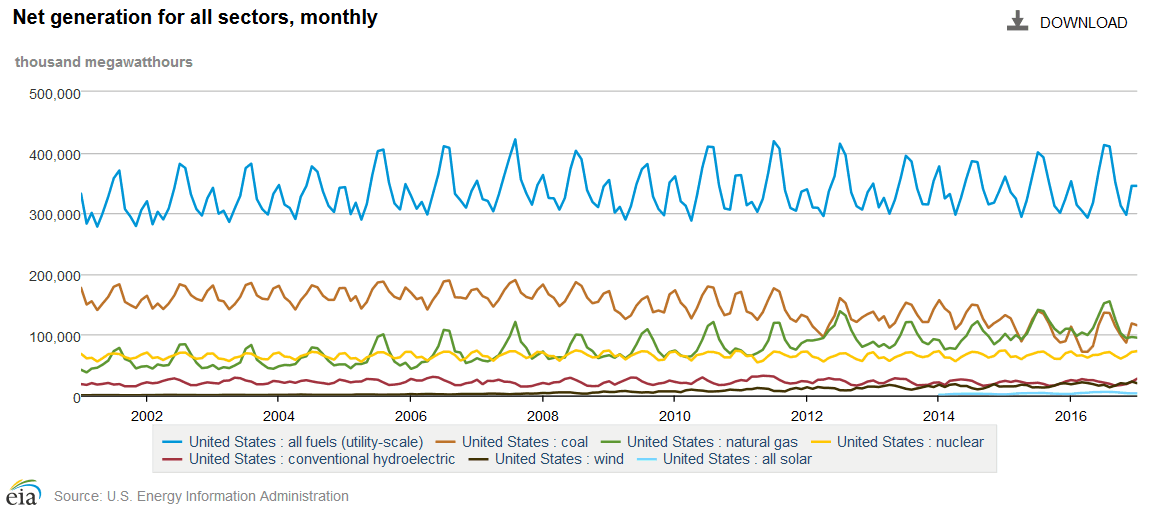


Figure US Electricity Generation by Energy Source 2001-2016 (GWh)

Net power generation by energy source is displayed in Figure 1. Coal has been the most important fuel in the power sector since the 1950s. But natural gas has risen dramatically since the new millennium and has recently overtaken coal for the top spot. Natural gas has been used as a peaking technology for decades. This is because a gas plant can be on line in minutes to meet peak demand, whereas coal and nuclear plants require hours. Moreover, natural gas has a much smaller carbon footprint than coal and has been relatively abundant and cheap since 2001. This comparative advantage (environmentally and financially) has led to a rapid expansion in natural gas generation capacity and consequently the dramatic rise of natural gas.

Net power generation routinely peaks in the summer to meet air conditioning loads. However, electricity is also used to heat buildings in the winter, and this can be seen in a distinct but smaller winter peak. The contributions of most fuels show substantial seasonality. For this project, we shall examine natural gas as a fuel for electric power generation in the United States.

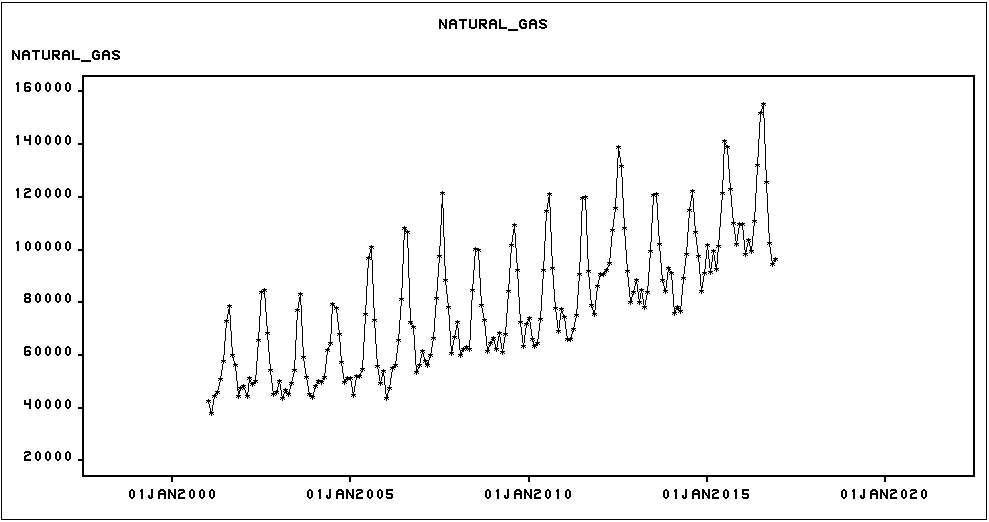


Figure Natural Gas for Electricity Generation (GWh)

The time-series plot of natural gas for power generation is shown in Figure 2. There is clearly an upward trend that may be accelerating as the variability (amplitude) seems to increase over time. If this is due to a constant growth rate, a linear trend should be a good fit to the log transform of the series.

A dominant summer peak is very clear in the seasonal box plot below (Figure 3). This is consistent with the traditional role of natural gas as a peaking technology. Before the advent of modern hydraulic fracturing (fracking) in 1999 natural gas was relatively expensive. Natural gas plants were relatively small and were located near urban centers to meet summer air conditioning loads. Since the year 2000, natural gas has become abundant and low cost, making it the fuel of choice for power generation not only for peak loads but for base loads a well. So it is not surprising that the box plot (Figure 3) shows substantial generation from natural gas in the fall, winter and spring. But for our purposes in this analysis, Figure 3 shows that this series has a distinct seasonal component.

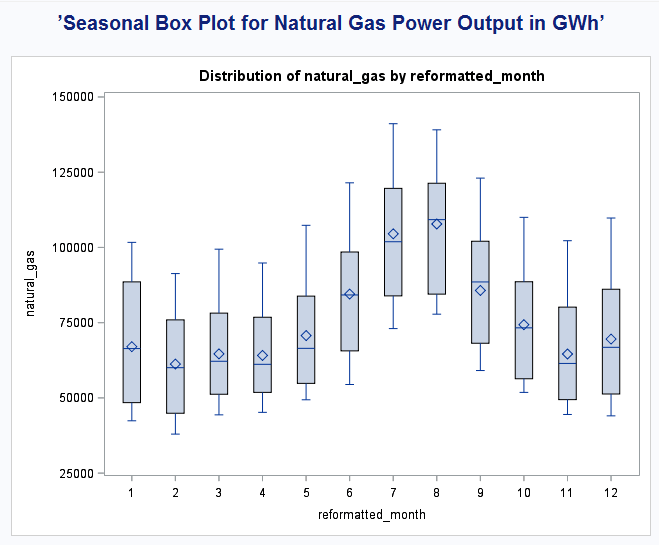


Figure Seasonal Distribution of Natural Gas for Electric Power Generation

The autocorrelation functions for this series are shown in Figure 4. The autocorrelations are decaying sinusoidally. In fact, the autocorrelation remains outside of two standard error bounds at lags up to 24 months. Thus, the natural gas time series is not stationary. As such, seasonal differencing and seasonal dummies are potential methods to deal with non-stationarity.

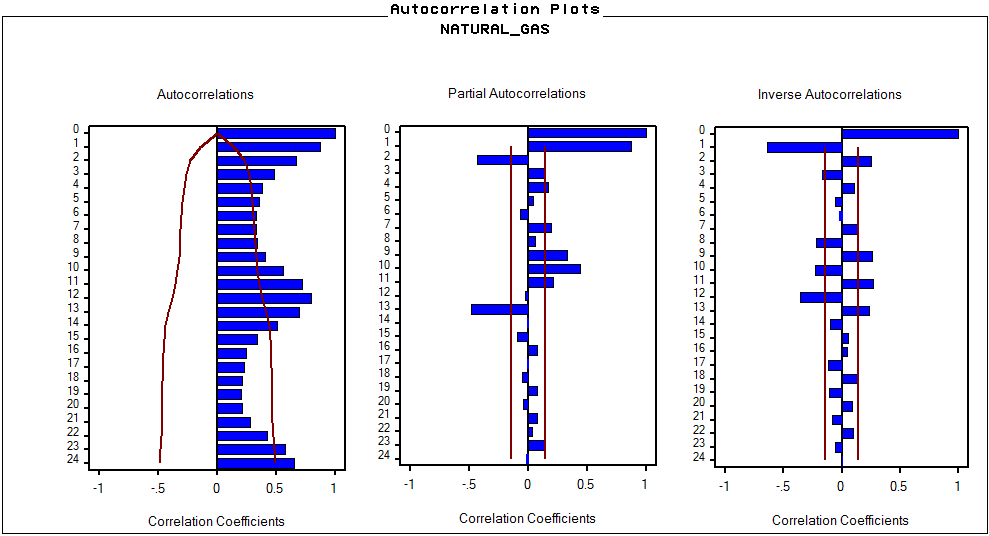


Figure Natural Gas Autocorrelation Functions

1. **Univariate Time-series models**
   1. **Deterministic Time Series Models and Error model**

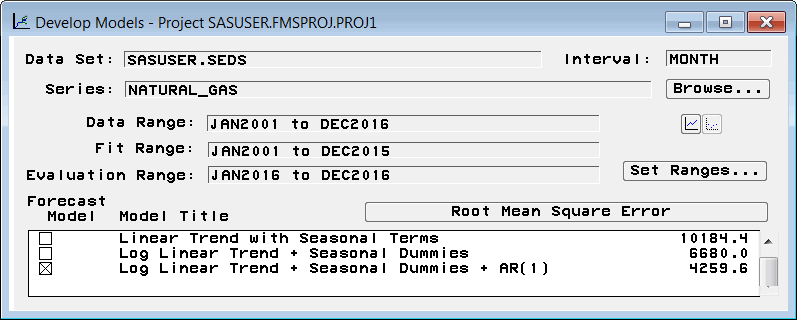


Figure Deterministic models with RMSE

In this project, we use a holdout sample of twelve periods from the overall data for evaluation: January to December 2016.

Several deterministic models were fit to the time series (Figure 5). As predicted from the curvature of the time series plot, a log transform of the series proved to best fit the linear trend. Because the series has marked seasonality, a model with seasonal terms was in order. The best fitting deterministic model without explicitly modeling the errors was a linear trend on the log transformed series with seasonal terms (Figure 6), but we can see that there exists additional autocorrelation in the residuals (Figure 7). Because the partial autocorrelation function chops off after lag 1, we use AR(1) as our error model.

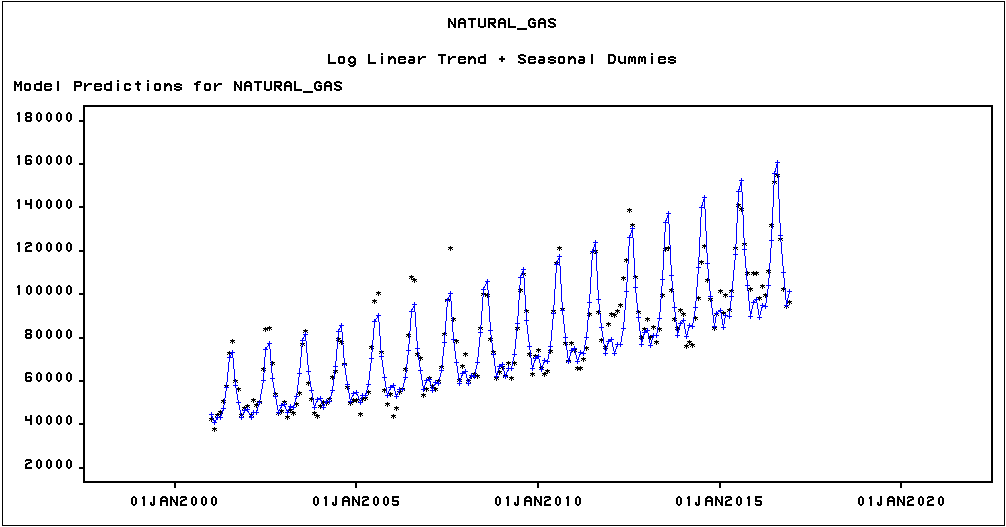


Figure 6 Log Linear Trend + Seasonal Dummies

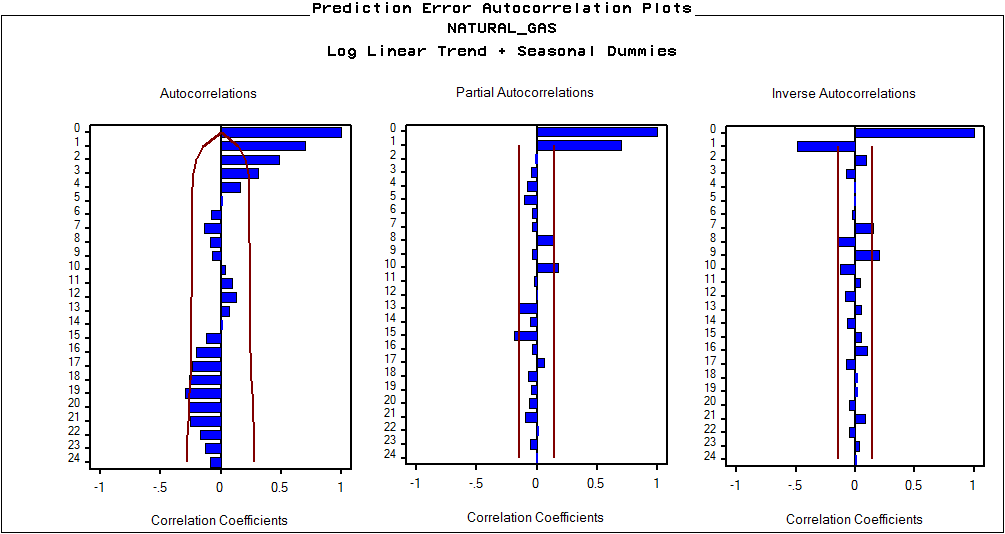


Figure 7 ACF of Log Linear Trend + Seasonal Dummies

Now, our best-fitting model is Log Linear Trend + Seasonal Dummies + AR(1). Plots of actuals vs. fitted values, ACF and White Noise/Stationarity tests for this best deterministic model are shown in Figures 8-10 below.

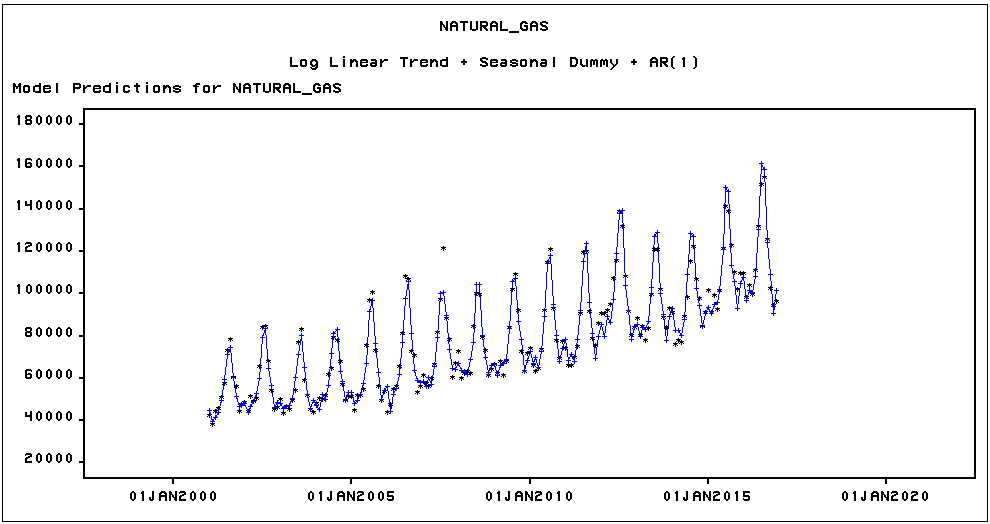


Figure 8 Actual v Predicted plot for Log Linear Trend + Seasonal Dummies

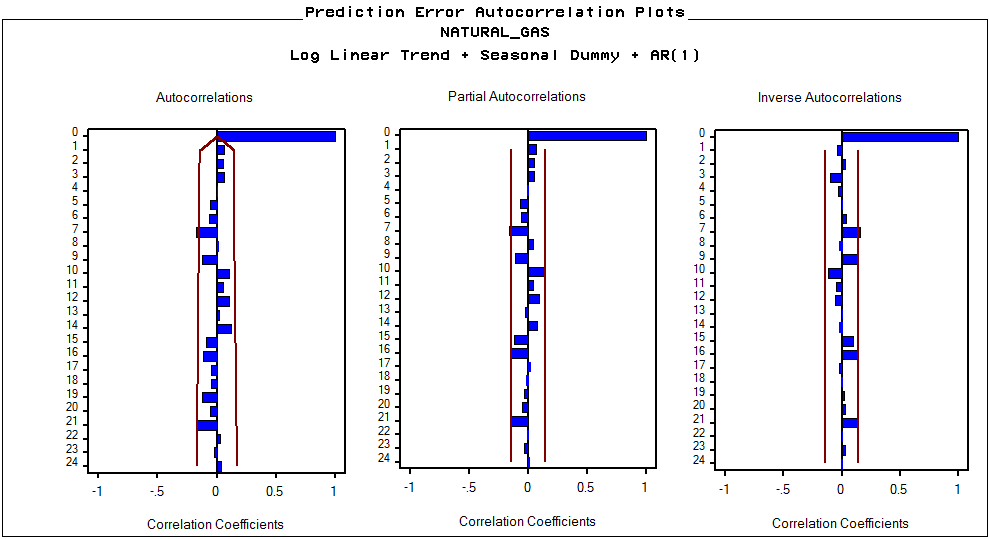


Figure 9 ACF of Log Linear Trend + Seasonal Dummies + AR(1)

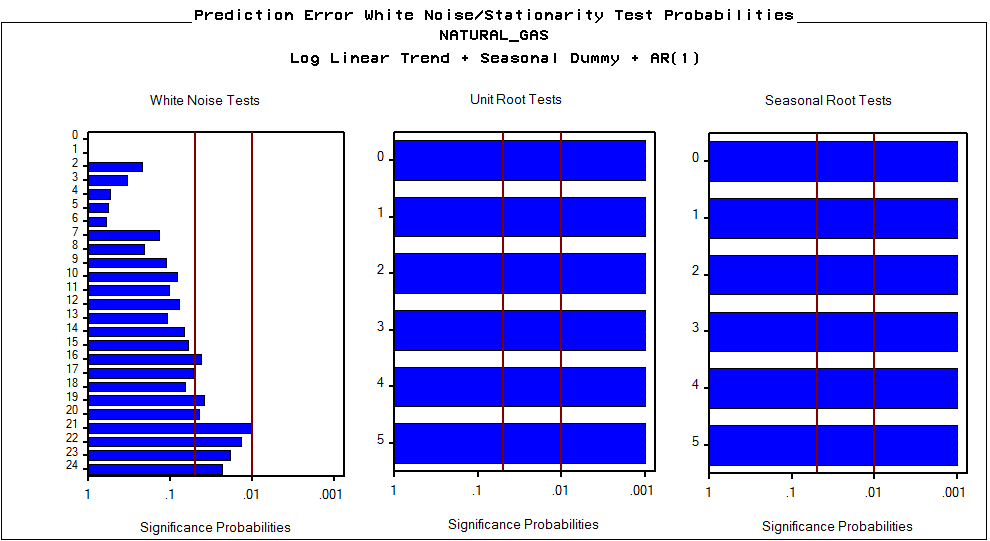


Figure 10 White Noise & Stationarity tests of Log Linear Trend + Seasonal Dummies + AR(1)

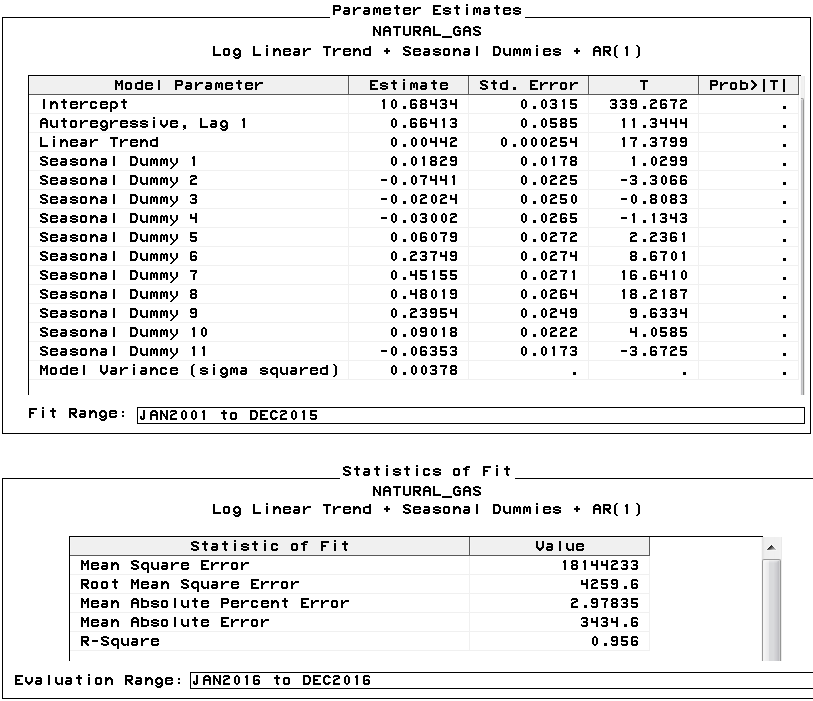


Figure 11 Parameter Estimates & Statistics of Fit for Log Linear Trend + Seasonal Dummies + AR(1)

For a deterministic time series model, our Log Linear Trend + Seasonal Dummy + AR(1) model is a very good fit. It has the lowest RMSE (4259.6) of our deterministic models and explains 95.6% of the variance in the data over the holdout period. The residuals are indistinguishable from white noise up to lag 15. The model variance over the Fit Range is a very low 0.00378. No doubt the variance is low due to our use of the log transform of the data. Nonetheless, the model fit is not perfect. There is significant autocorrelation at lag 7, and the residuals are not white noise at lags 16 and higher. We will attempt to identify a better-fitting model with more advanced modeling techniques.

* 1. **Exponential Smoothing models**

All exponential smoothing methods available in SAS 9.4 were applied to the natural gas time series in order to see if any smoothing models could produce a lower Evaluation (Holdout) Range Root Mean Square Error than the Log Linear Trend + Seasonal Dummy + AR(1) deterministic model. As seen in the Model Comparison (Figure 12), none of the smoothing models outperform the deterministic model. That said, the Seasonal Exponential Smoothing and Winters models come very close. Note that we fit the data and the log transform of the data to each model. Figure 12 reports only the best fitting model over the Evaluation Range for each method.

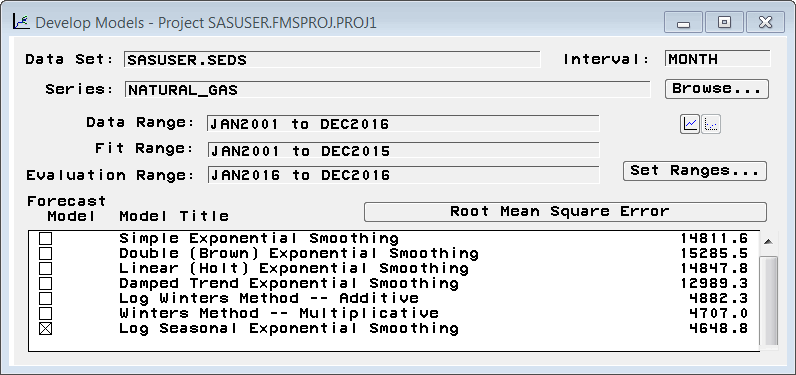


Figure 12 Smoothing Models with RMSE

Whereas the Winters models uses Level, Trend and Seasonal components the Seasonal Smoothing model has the same Level and Seasonal components but does not have any Trend smoothing. Given this difference, one might expect the more flexible Winters methods to perform better. But in this case the Trend components of both Winters models are insignificant. As a result the Log Seasonal Exponential Smoothing model is the best-fitting smoothing model (over the Evaluation Range) because it is simpler (more parsimonious).

Plots of actual vs. fitted values, autocorrelation functions, and white noise and stationarity tests for this best smoothing model—Log Seasonal Exponential Smoothing—are shown in Figures 13-15 below.

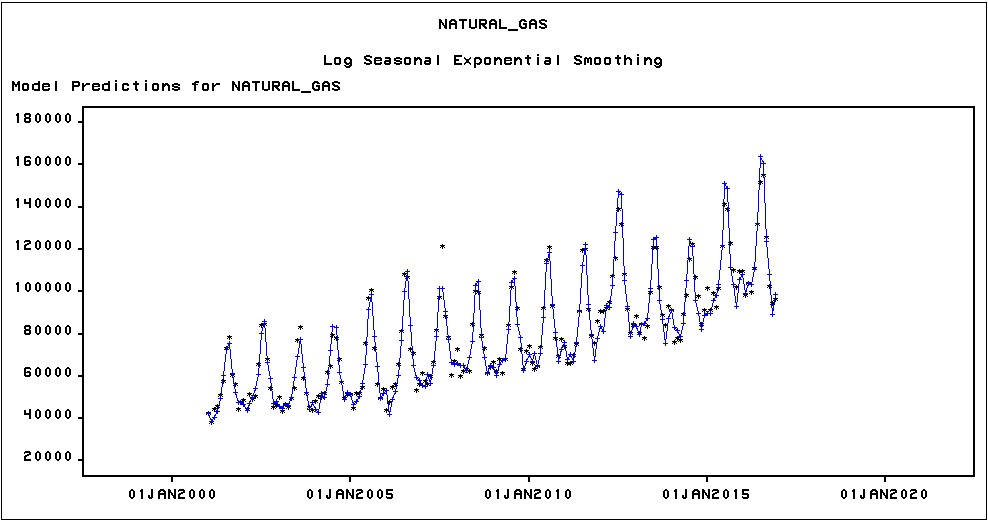


Figure 13 Actual v Predicted for Log Seasonal Exponential Smoothing

The plot of actual vs predicted values is hard to distinguish from the same plot for the best deterministic model. However, autocorrelation at lag 7 and partial autocorrelations at lags 7 and 9 are significant. Moreover, the residuals are clearly not white noise. Interestingly, the ACF, PACF and white noise tests for the Seasonal Exponential Smoothing model on the untransformed series (not shown) are all much cleaner, but it does not perform nearly as well over the Evaluation Range as the logged series does in terms of RMSE. So despite subpar ACF and White Noise Test plots, we keep Log Seasonal Exponential Smoothing as the best smoothing model.

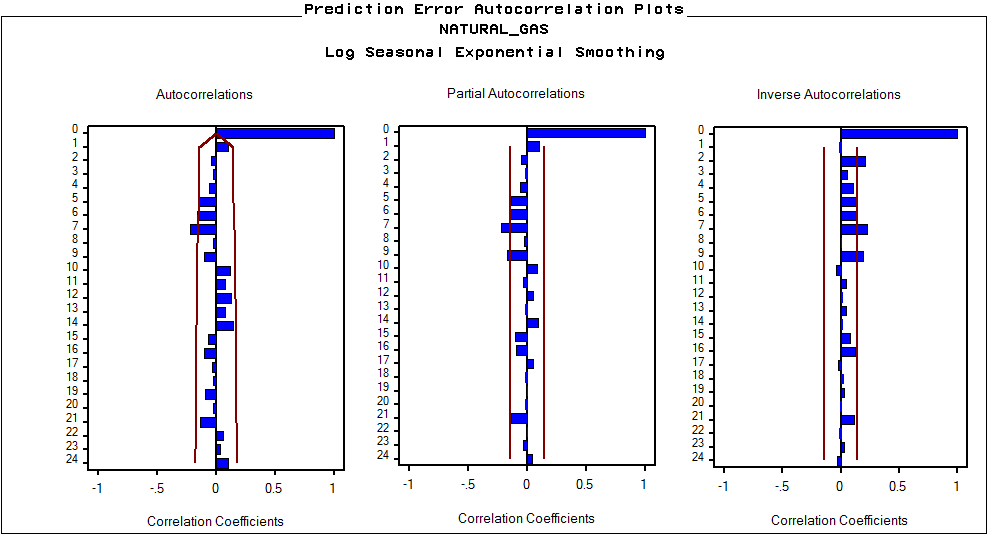


Figure 14 ACF of Log Seasonal Exponential Smoothing

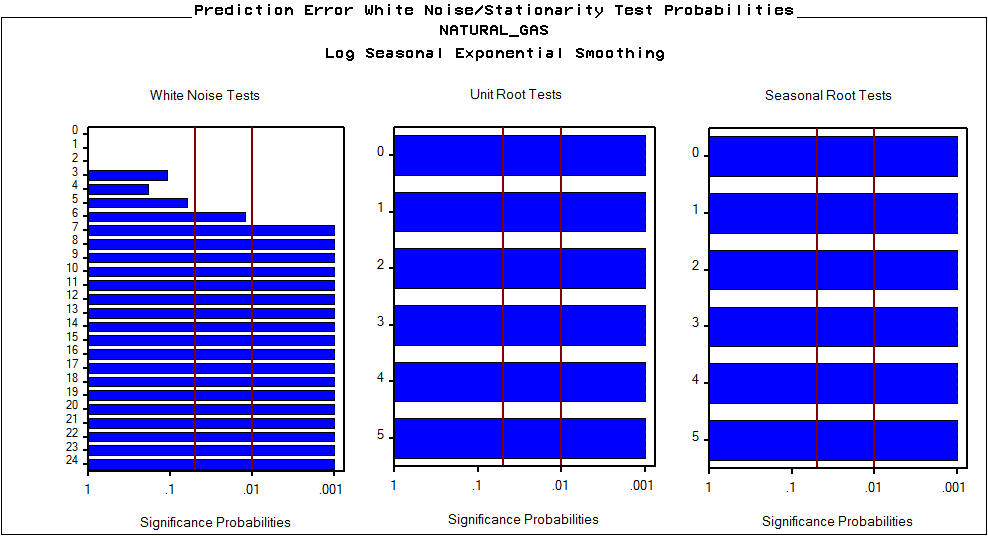


Figure 15 White Noise & Stationarity tests for Log Seasonal Exponential Smoothing

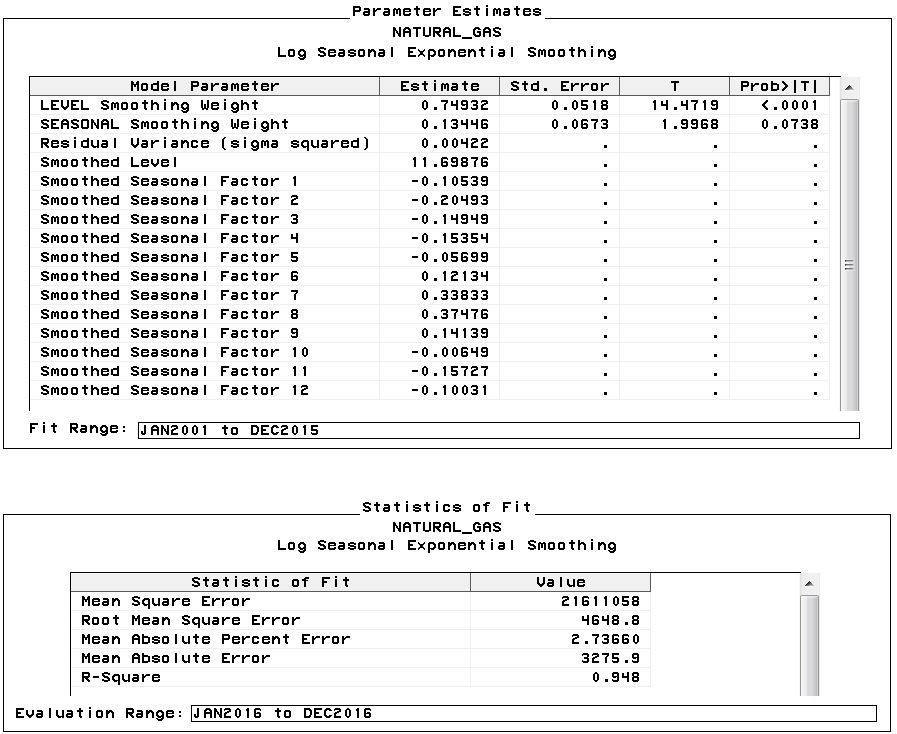


Figure 16 Parameter Estimates and Statistics of Fit for Log Seasonal Exponential Smoothing

With an R-square of 0.948 (Figure 16), this model fits our data well, explaining 94.8% of the variance over the Evaluation Range. Over the Fit Range the residual variance is a very low 0.00422. But the ACF plots are not clean and the residuals are clearly not white noise. A more advanced model may do better.

* 1. **ARIMA models**

In the autocorrelation function of the original series (Figure 4), we noted the non-stationarity of the series and indicated that differencing may make the series stationary. Autocorrelation plots for the simple differenced series is shown in Figure 17. Clearly, the simple difference alone did not induce stationarity in the series, as there are significant autocorrelations at many lags, most notably at the seasonal lags of 12 and 24. Therefore, we also took a seasonal difference. The resulting autocorrelation in the simple and seasonally differenced series is shown in Figure 18. The simple and seasonally differenced series now appears stationary.

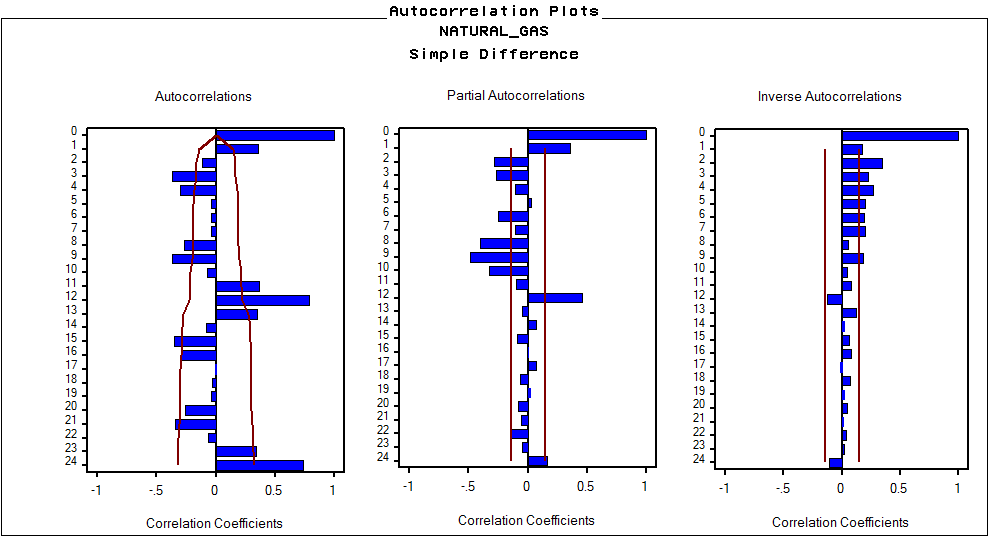


Figure 17 ACF of Simple Difference

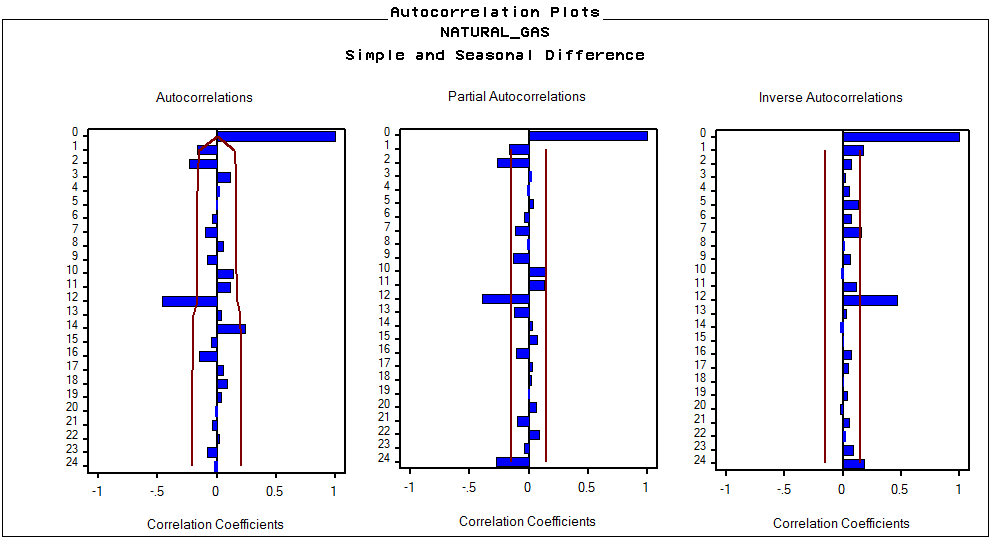
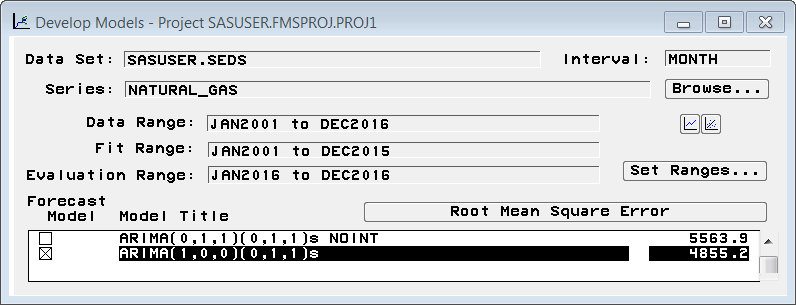


Figure 18 ACF of Simple and Seasonal Difference

From the ACF and PACF of the simple and seasonal differenced series we see that the seasonal lag 12 cuts off in the ACF, but decays in the PACF. This identifies a seasonal MA(1) term. Lags 1 and 2 are significant and may decay more quickly in the ACF than in the PACF, but some experimentation will be in order to determine whether AR or MA terms are most appropriate.

First we fit an ARIMA(0,1,1)(0,1,1) model which fits the data quite well. Then after some experimentation with MA and AR and removing simple differencing, we find that ARIMA(1,0,0)(0,1,1)+C fits the data even better. Both models reduce the residuals to white noise but as shown in Figure 19, ARIMA(1,0,0)(0,1,1)+C has a lower RMSE based on the Evaluation Range and therefore is our model of choice.

Plots of actuals vs fitted values, autocorrelation functions, and white noise and stationarity tests for the ARIMA(1,0,0)(0,1,1)+C model are shown in Figures 20-22 below.

**

*Figure 19 Seasonal ARIMA Models with RMSE*

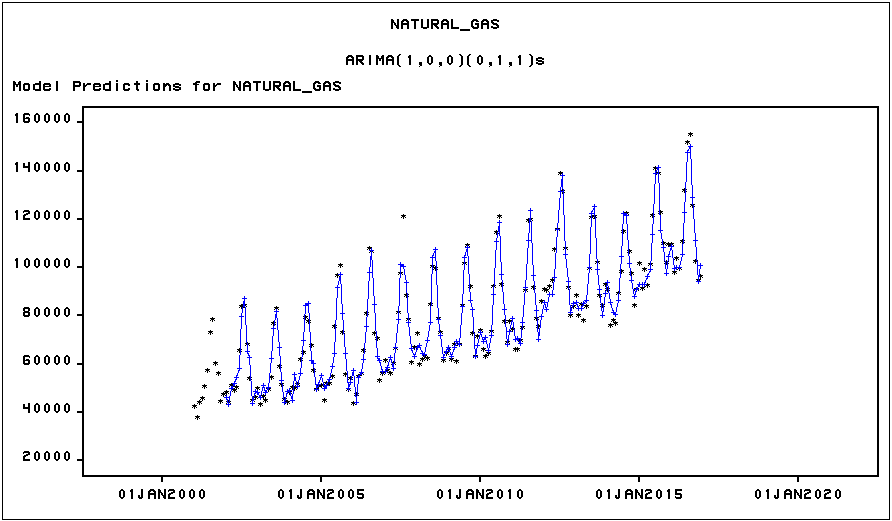


Figure 20 Actual v Predicted Seasonal ARIMA Model

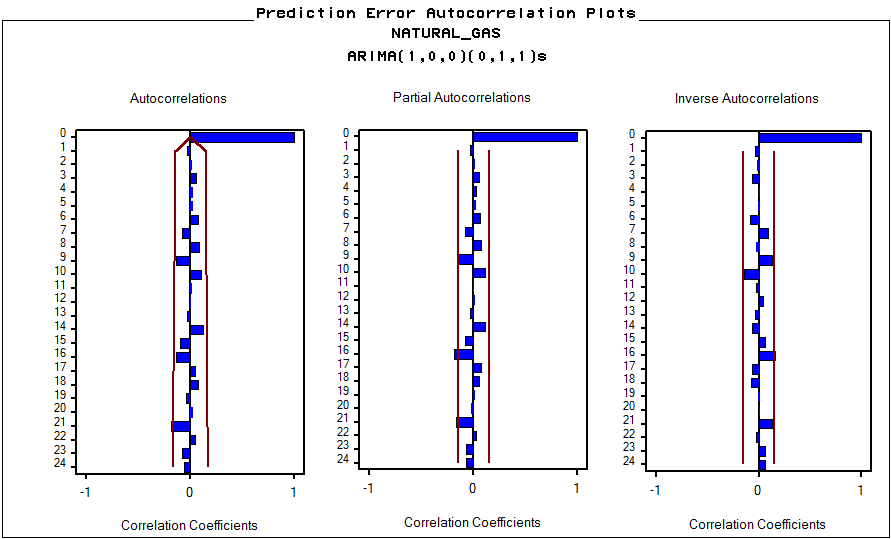


Figure 21 ACF, PACF, & IACF for ARIMA(1,0,0) (0,1,1) + C

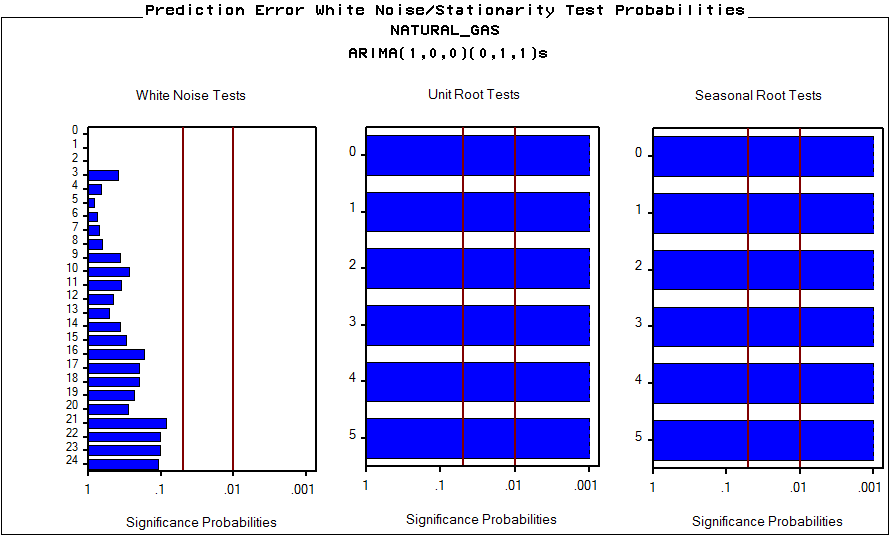


Figure 22 White Noise & Stationarity Tests for ARIMA(1,0,0) (0,1,1) + C

The plot of the actual data vs. predicted values for this model is visually indistinguishable from the prior two non-ARIMA models, except for the fact that there are no predicted values for the first 12 lags due to the seasonal differencing. However, the autocorrelation function and the white noise test show that there is no significant autocorrelation left in the residuals. In other words, the residuals have been reduced entirely to white noise. As such, this model is reliable for forecasting. This parsimonious model, with an estimated constant and only 2 parameters (Figure 23), explains 94.3% of the variance in Evaluation Range data (Figure 24). One-year forecast and error bounds are presented in Figure 25.

The model variance over the Fit Range is 23,914,031 (Figure 23). This model variance is completely inconsistent with the model variance of the Log Linear Trend + Seasonal Dummy + AR(1) model (0.00378) and with the residual variance of the Log Seasonal Exponential Smoothing Model (0.00422). This is because this model fits the untransformed data and both prior models fit the natural log of the data. As such it is unclear how one would proceed to compare the fit of models that use different transformations of the data, except by using a common metric such as RMSE.

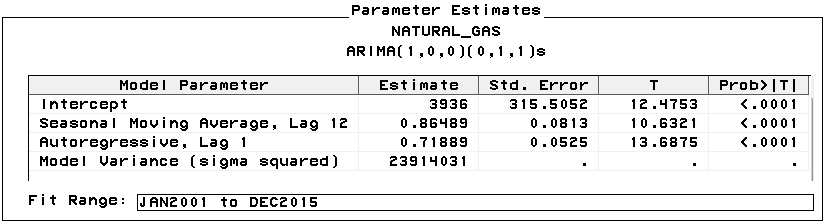


Figure 23 Parameter Estimates for ARIMA(1,0,0) (0,1,1) + C

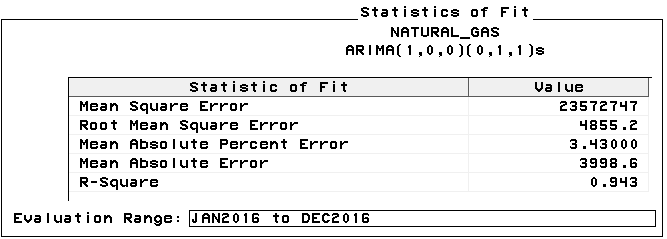


Figure 24 Statistics of Fit for ARIMA(1,0,0) (0,1,1) + C

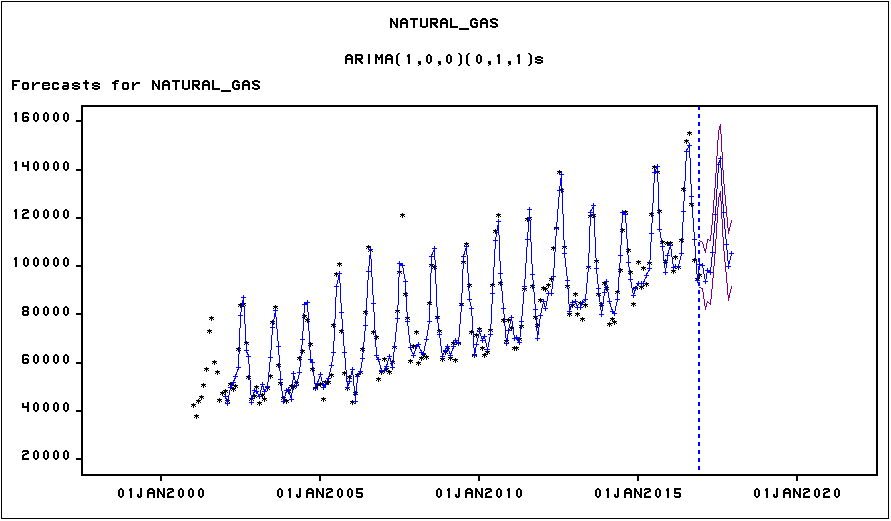


Figure 25 Forecasts by for ARIMA(1,0,0) (0,1,1) + C

1. **Bivariate Time Series Models**

Next, we conducted bivariate modeling. In these models, natural gas was treated as a function of two additional time series: coal and hydroelectric. All model parameters were estimated for a 15-year fit period from January 2001 through December 2015. Fit measures for each model were evaluated over the 12-month holdout period from January through December 2016.

Given the structural trends in power generation since the dawn of the new millennium—where the price of natural gas has dropped and supplies have increased dramatically—we expect to see that coal and natural gas are inversely related. Where one rises the other should fall. Not necessarily one-to-one, but use of each fuel for electricity generation is influenced by the same economic, financial and environmental factors. Natural gas appears to be substituting for coal since 2001. We hope to identify the strength of that substitution, if it is indeed happening, in this analysis. In contrast, we expect that power generated by hydroelectric is not related to coal or gas. This is because hydroelectric plants generate when the water flows generally peaking from March to June. This is a different seasonality than the overall electric power demand that coal or gas serve. So we expect no relationship between hydro and coal or hydro and gas.

First, we applied the same univariate time series methods employed in the previous section to identify a model for coal. The time series plot shows a non-linear diminishing series (Figure 26).

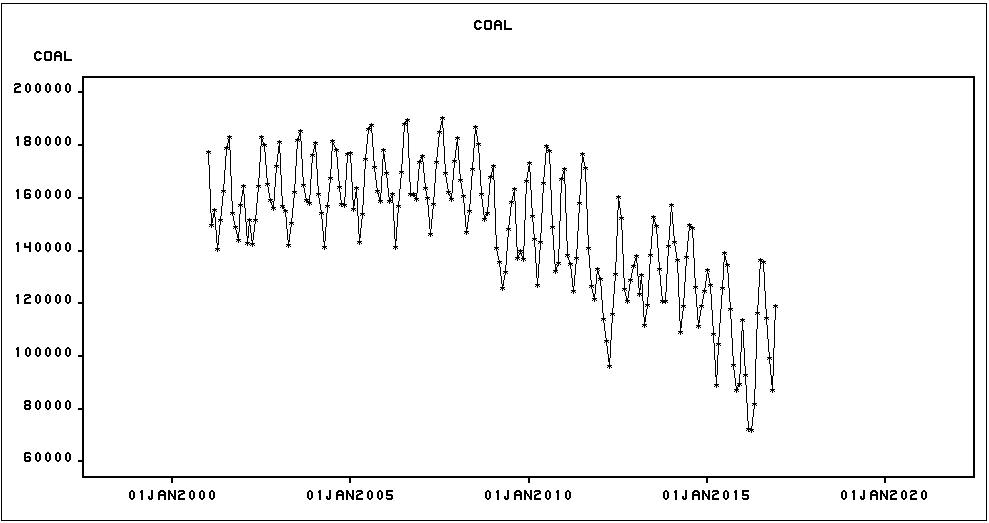


Figure 26 Coal Time Series

Contrary to our initial hypothesis, we found that the series cannot be reduced to white noise via a Quadratic Trend + Seasonal Dummies model. Instead, the best-fitting model for coal turned out to be seasonal ARIMA (1,0,0)(0,1,1)+C, the same model we identified for the variable natural gas in the previous section. The model predictions, parameters, and fit statistics are shown in Figures 27-29. The corresponding autocorrelation function (Figure 30) confirms the white noise behavior of the residuals.

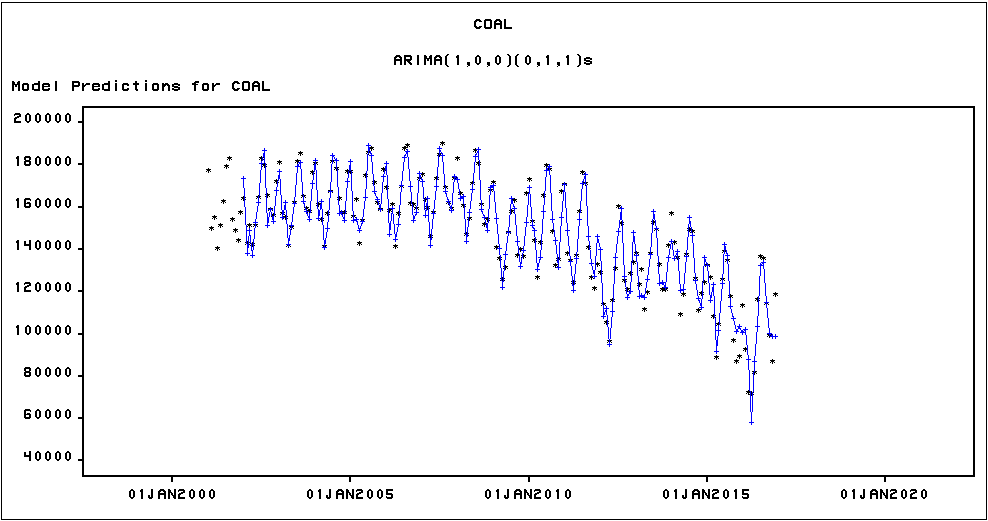


Figure 27 ARIMA (1,0,0) (0,1,1) + C Model Predictions for Coal

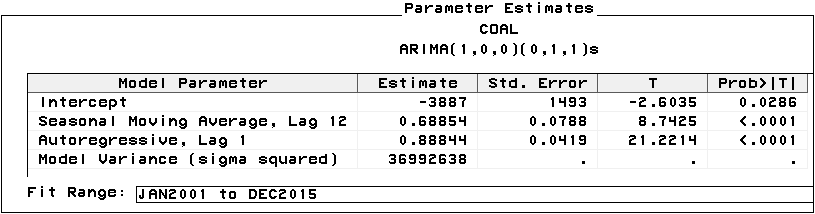


Figure 28 ARIMA (1,0,0) (0,1,1) + C Parameter Estimates for Coal

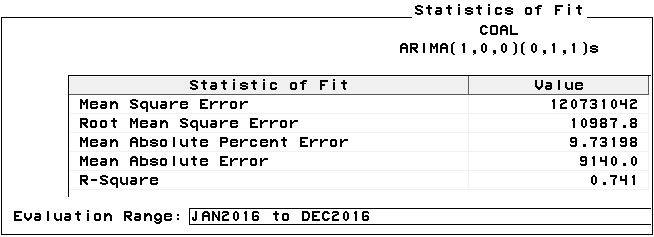


Figure 29 ARIMA (1,0,0) (0,1,1) + C Fit Statistics for Coal

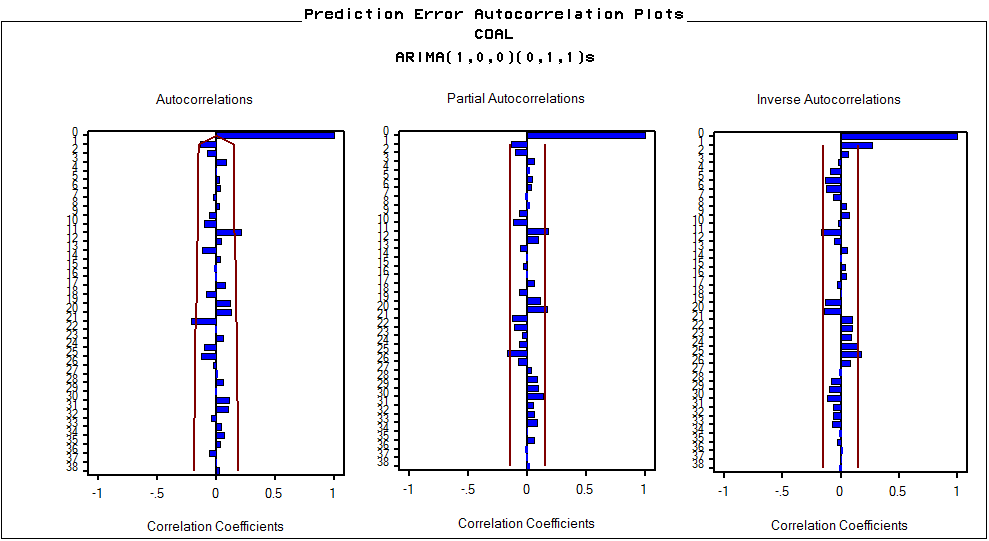


Figure 30 ARIMA (1,0,0) (0,1,1) + C ACF, PACF, & IACF for Coal

The same univariate modeling techniques were used to search for the best-fitting ARIMA model for hydroelectric (“hydro”), which is shown in Figure 31.

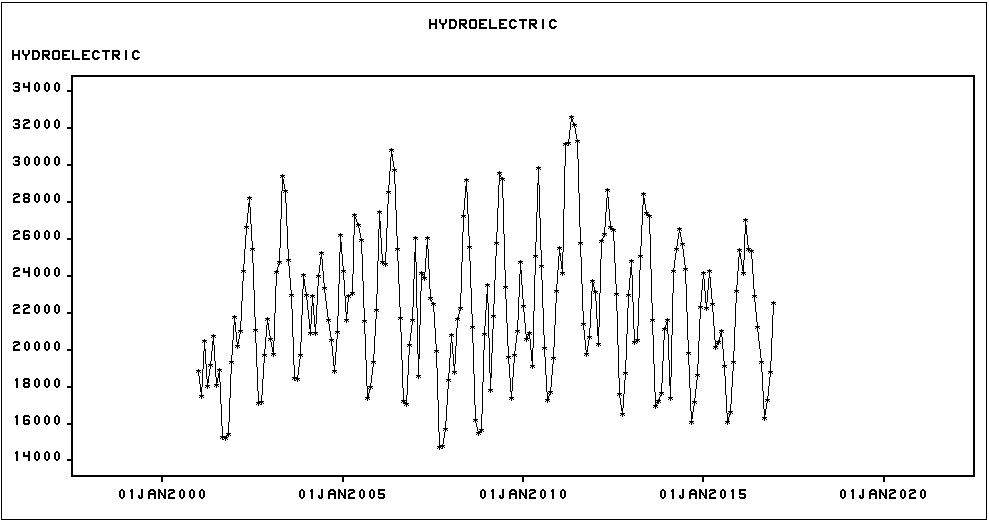


Figure 31 Hydro Time Series

We found the best-fitting model to be seasonal ARIMA (1,0,0)(1,1,0). The model predictions, parameters, and fit statistics are shown in Figures 32-34.

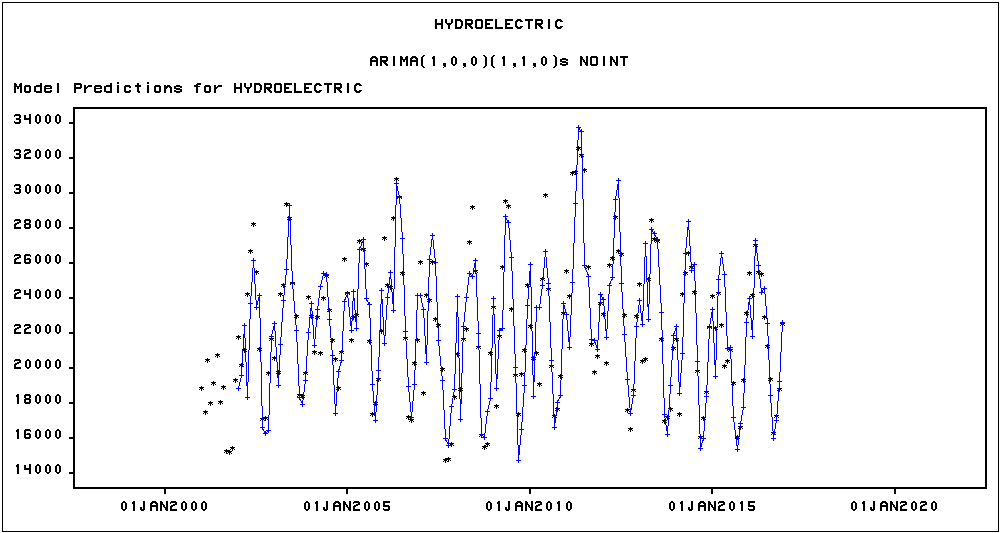


Figure 32 ARIMA (1,0,0) (1,1,0) Model Predictions for Hydro

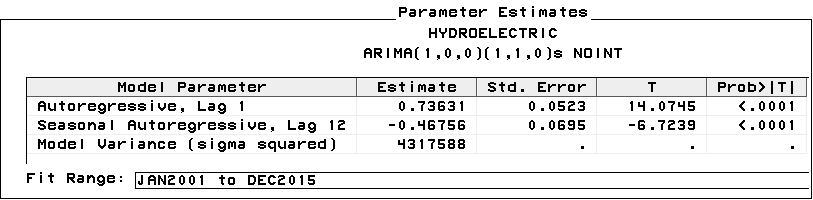


Figure 33 ARIMA (1,0,0) (1,1,0) Parameter Estimates for Hydro

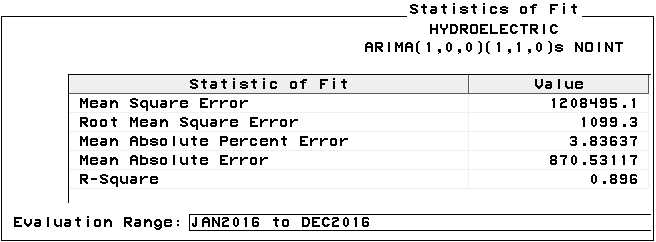


Figure 34 ARIMA (1,0,0) (1,1,0) Fit Statistics for Hydro

As displayed in Figure 35, there appears to be a slightly non-stationary behavior at the seasonal lag 24. The autocorrelation value at lag 24, though outside the 2-standard error bound, is relatively small and less than 0.25. Overall, the autocorrelation function can be considered as stationary and residuals are approximately white-noise.

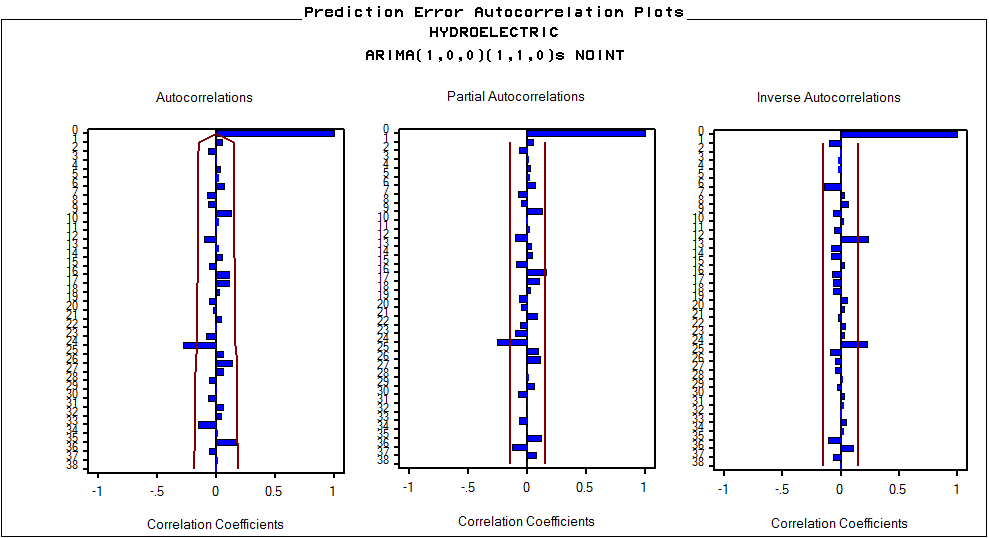


Figure 35 ARIMA (1,0,0)(1,1,0) Autocorrelation Function for Hydro

Next, with all the univariate models properly identified, we proceeded to the bivariate modeling step of pre-whitening and building the cross-correlation functions. For natural gas as a function of coal, we wrote the following SAS code:

**PROC** **ARIMA** DATA=ENERGY;

IDENTIFY VAR=natural\_gas(**12**) NOPRINT;

IDENTIFY VAR=coal(**12**) NOPRINT;

ESTIMATE P=(**1**) Q=(**12**) METHOD=ML;

IDENTIFY VAR=natural\_gas(**12**) CROSSCOR=(coal(**12**));

**RUN**;

We made sure that the same differencing, i.e. only seasonal differencing, was applied to both the input and output series. We used the pre-whitening filters of simple AR(1) and seasonal MA(1) that are consistent with the best-fitting model ARIMA (1,0,0)(0,1,1)+C we identified for coal earlier. The resulting cross-correlation function between natural gas and coal is shown in Figure 36.

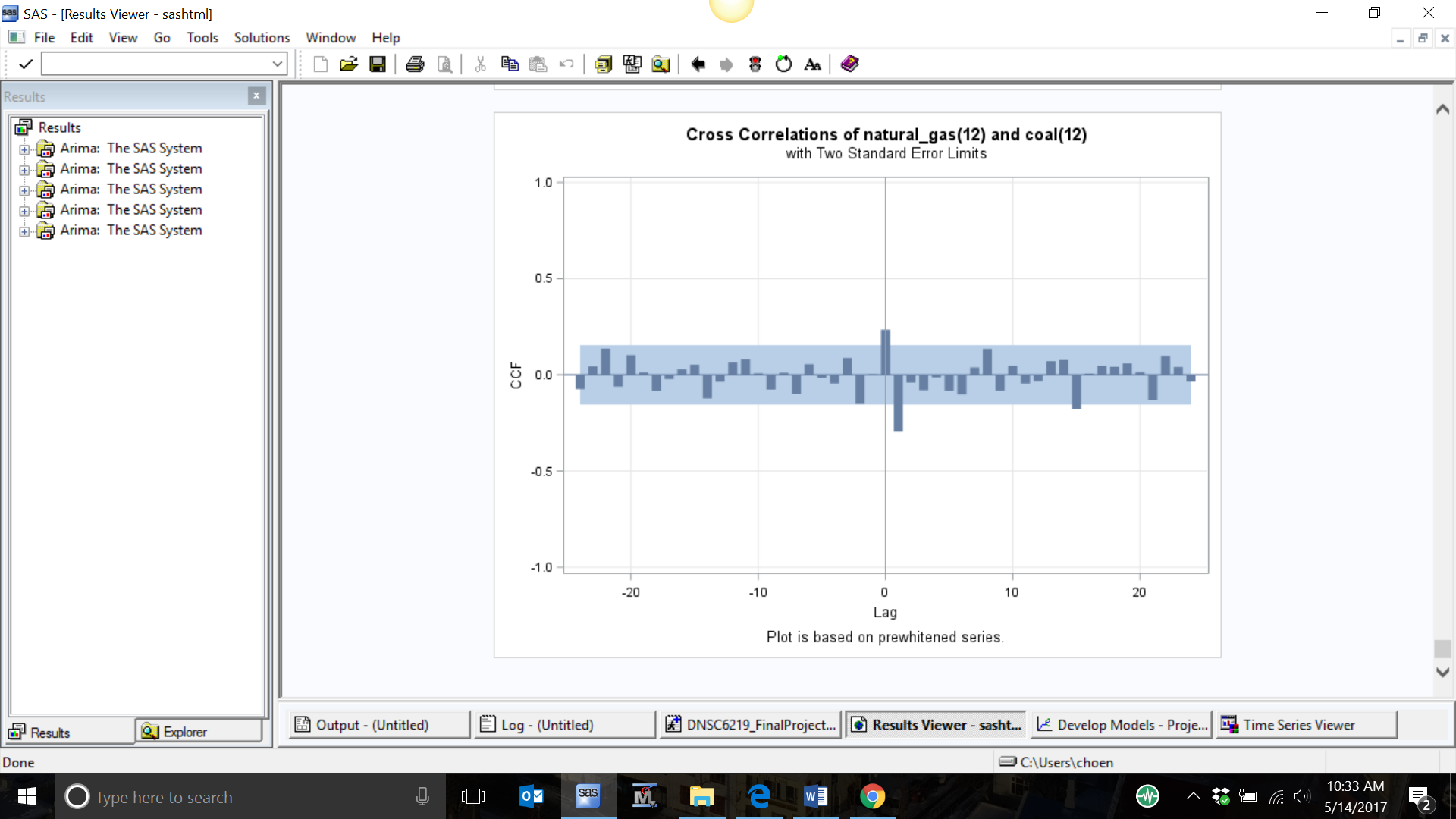


Figure 36 CCF of Natural Gas vs Coal

Responses occur at lags 0 and 1, suggesting that the change in natural gas consumption can indeed be explained by the change in coal consumption. A transfer function can be specified with b = 0, because the first response occurs at lag 0; s = 1, because of where the “decay” begins; and r = 0, due to the lack of an exponential or sinusoidal decay pattern.

**PROC** **ARIMA** DATA=ENERGY;

IDENTIFY VAR=natural\_gas(**12**) NOPRINT;

IDENTIFY VAR=coal(**12**) NOPRINT;

ESTIMATE P=(**1**) Q=(**12**) METHOD=ML;

IDENTIFY VAR=natural\_gas(**12**) CROSSCOR=(coal(**12**)) NOPRINT;

ESTIMATE INPUT=((**1**)/coal) P=(**1**) Q=(**12**) METHOD=ML PLOT;

**RUN**;

Note that again, simple AR(1) and seasonal MA(1) were applied to the residuals of the transfer function model. The resulting model predictions, parameters, and fit statistics (Figures 37-39) were examined to check for model adequacy. All the parameter estimates are statistically significant at α=0.05 and the autocorrelation function exhibits stationarity. All the p-values of the residual autocorrelations are greater than α=0.05, indicating no significant correlation exist between the residuals and the input variable coal.

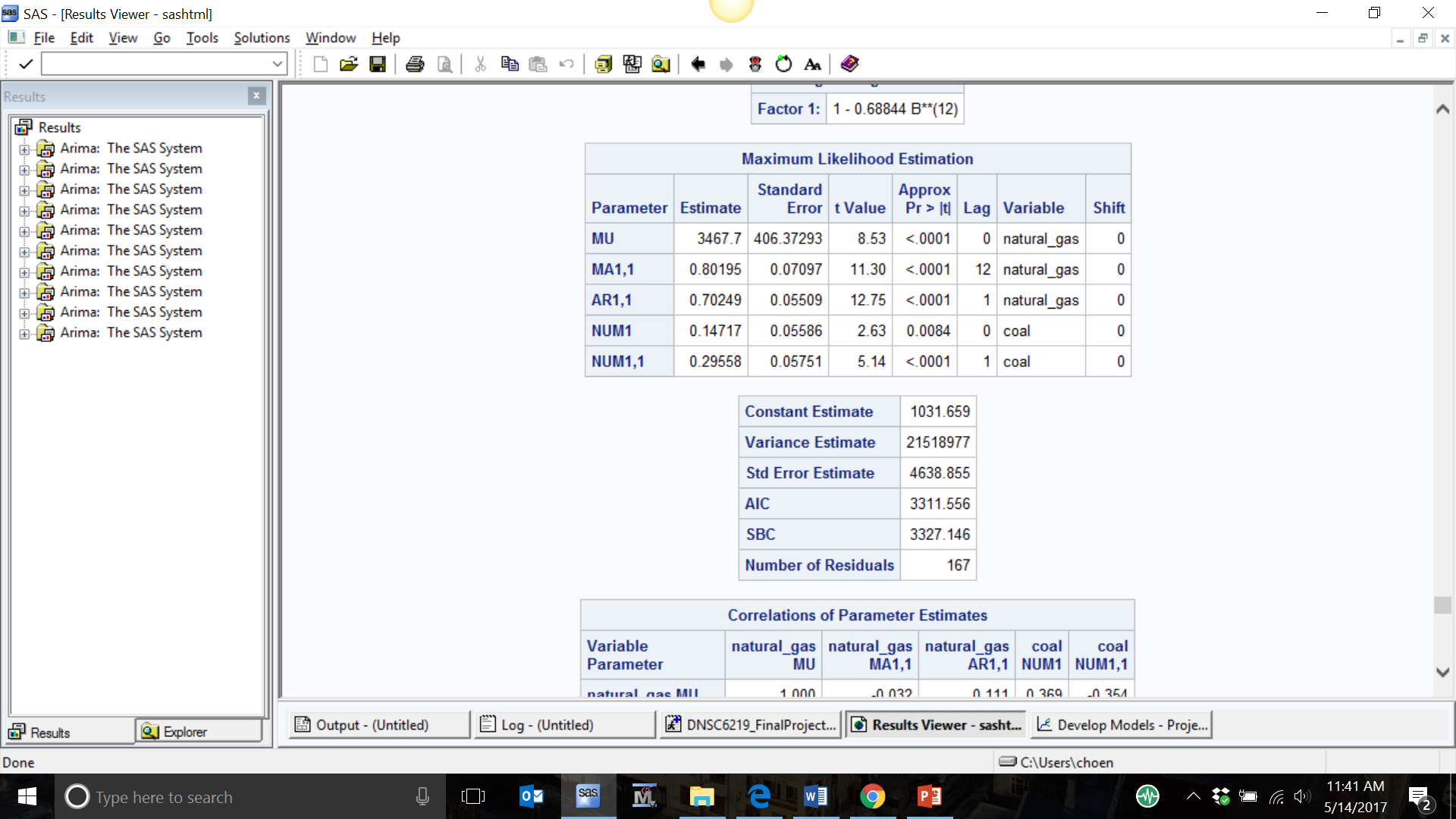


Figure 37 ARIMA Model Parameters and Fit Statistics of Natural Gas vs Coal

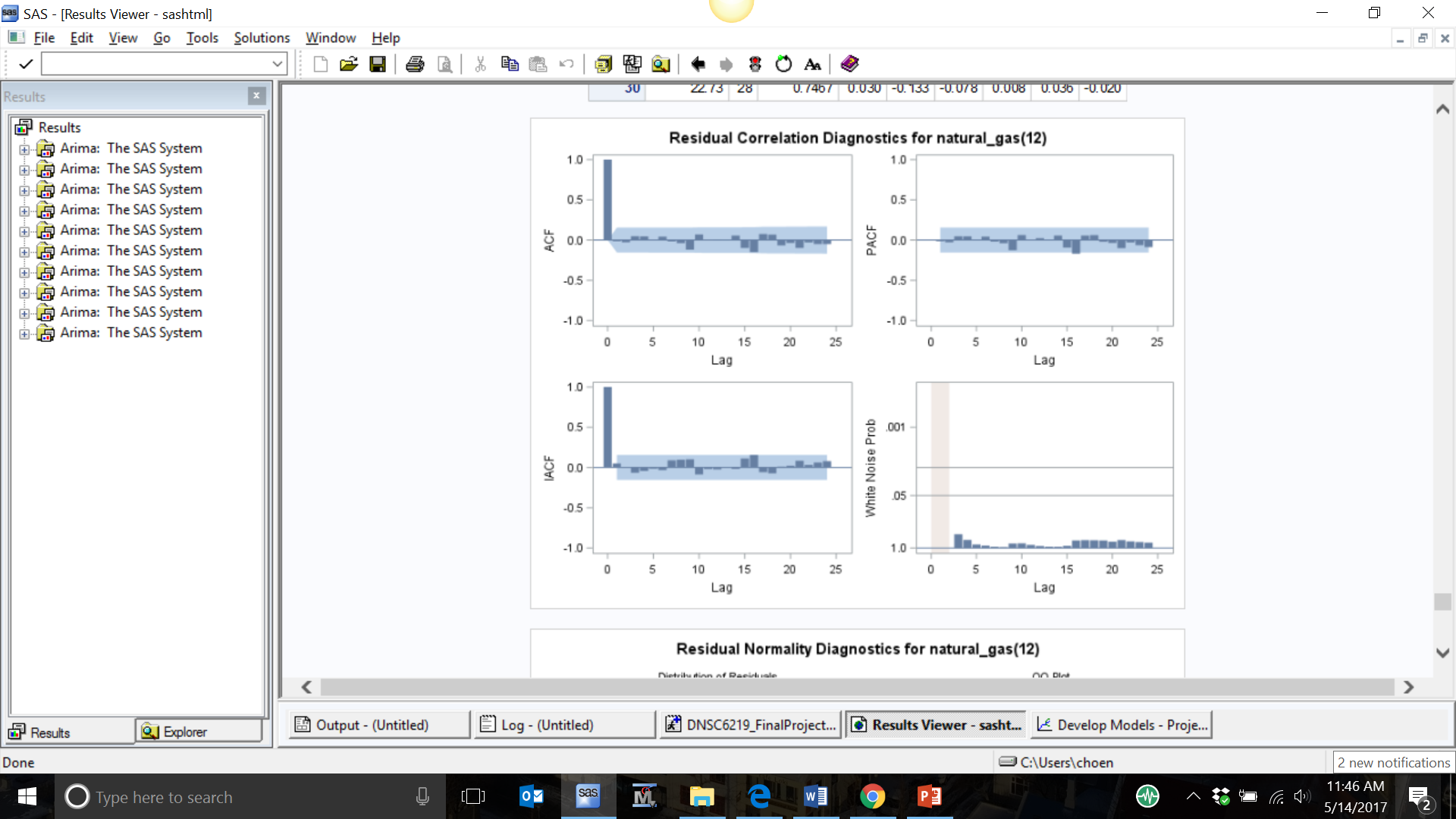


Figure 38 ACF, PACF, & IACF of Natural Gas vs Coal

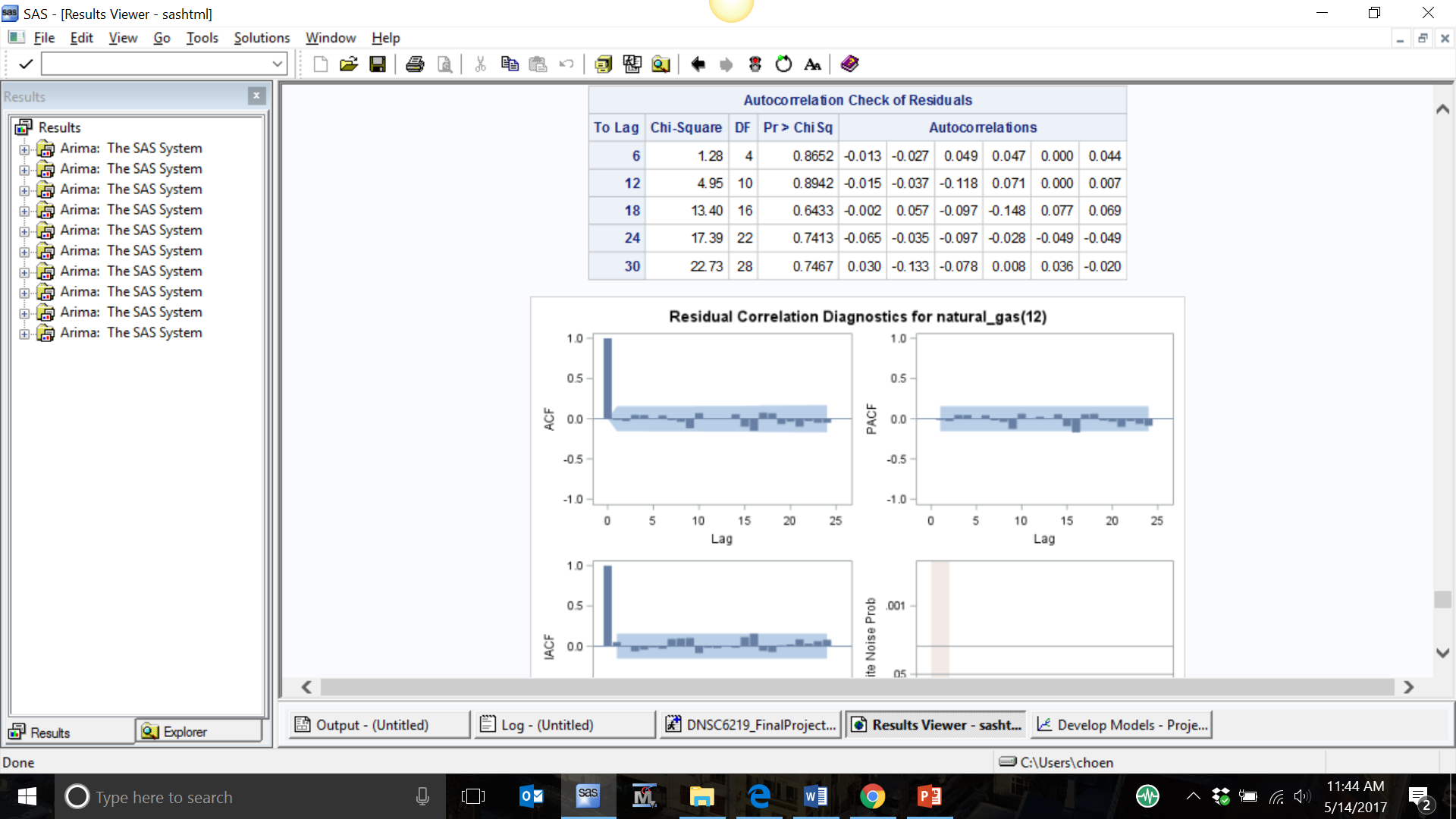


Figure 39 Autocorrelation Check of Residuals of Natural Gas vs Coal

When we attempted to model natural gas as a function of hydroelectric, we performed the same bivariate modeling steps. The following SAS code was written to pre-whiten the series and build cross-correlation function:

**PROC** **ARIMA** DATA=ENERGY;

IDENTIFY VAR=natural\_gas(**12**) NOPRINT;

IDENTIFY VAR=hydroelectric(**12**) NOPRINT;

ESTIMATE P=(**1**)(**12**) NOINT METHOD=ML;

IDENTIFY VAR=natural\_gas(**12**) CROSSCOR=(hydroelectric(**12**));

**RUN**;

Again, we made sure that the same differencing, i.e. only seasonal differencing, was applied to both the input and output series. This time, we used the pre-whitening filters of simple AR(1) plus seasonal AR(1) in accordance with the best-fitting model ARIMA (1,0,0)(1,1,0) that we obtained for hydroelectric earlier. The resulting cross-correlation function between natural gas and hydroelectric is shown in Figure 40.

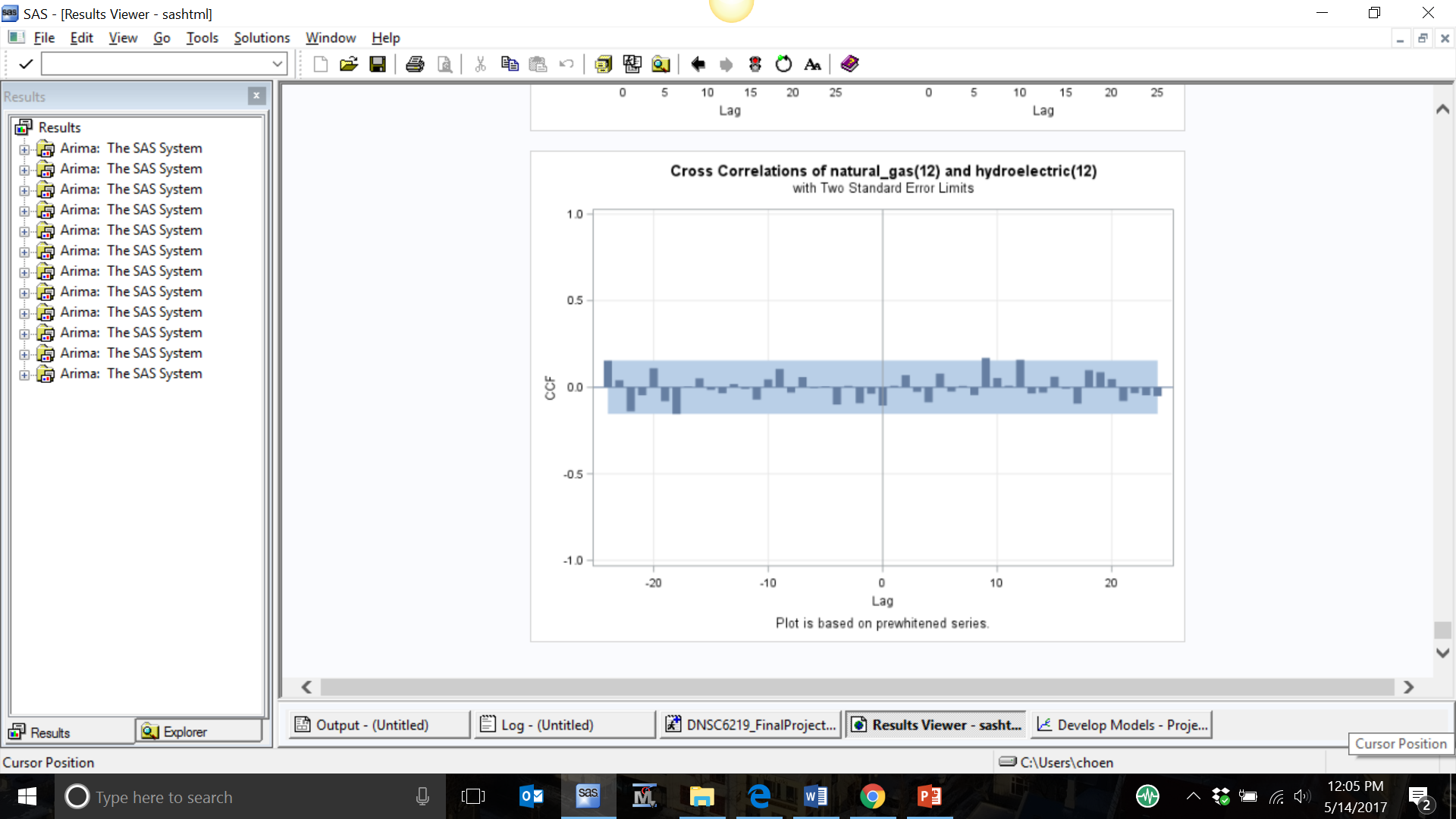


Figure 40 CCF of Natural Gas vs Hydro

The cross-correlation function displays no significant cross-correlations at any lags. Based on the cross-correlation function, we determined not to use a transfer function model. If we were to have used a transfer function model, it would need to have had parameters b = 0, s = 0, and r = 0. There is therefore no significant relationship between natural gas and hydroelectric.

1. **Comparison of Models and Conclusion**

The use of natural gas as a fuel for generating electricity has been growing sharply in the United States since the early 2000s. In this project, we used time series techniques to examine the relationships between natural gas and coal and hydroelectric as energy sources for electric power generation. We expected an inverse relationship between natural gas and coal and no relationship between natural gas and hydroelectric.

First, we modeled the natural gas time series to see if we could reduce the residuals to white noise so that the series could be used reliably for forecasting. Fit statistics over the 12-month evaluation period for all models in contention are displayed in Figure 41.

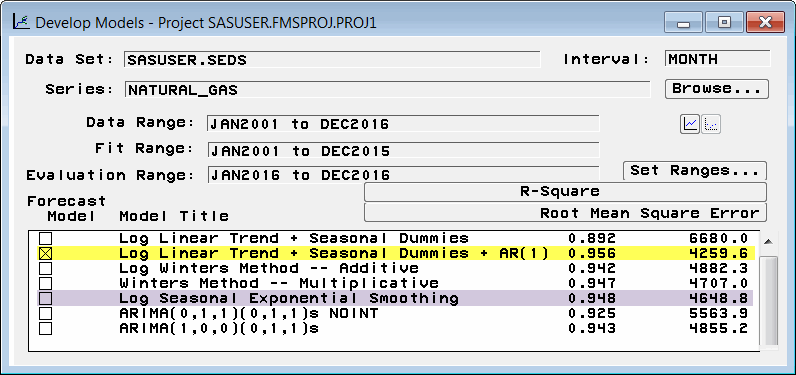


Figure 41 Model Comparison by R2 and RMSE over the 12 month Evaluation Range

From the deterministic models, we found that Log Linear Trend + Seasonal Dummies + AR(1) was the best fit over the evaluation range with an RMSE of 4259.6 and an R2 of 0.956. Of the smoothing models, the Log Seasonal Exponential Smoothing model had the second best fit with next lowest RMSE of 4648.8 and next highest R2 of 0.948. The rank order becomes cluttered after that. Among ARIMA models, the seasonal ARIMA (1,0,0)(0,1,1)+C had the best fit, with an RMSE of 4855.2 and an R2 of 0.943. Over this entire field of univariate models, only the seasonal ARIMA models reduced the residuals to pure white noise. As such, the seasonal ARIMA models can be used reliably for forecasting. However, the best-fitting univariate model for natural gas over the evaluation (holdout) range was the deterministic model Log Linear Trend + Seasonal Dummies + AR(1).

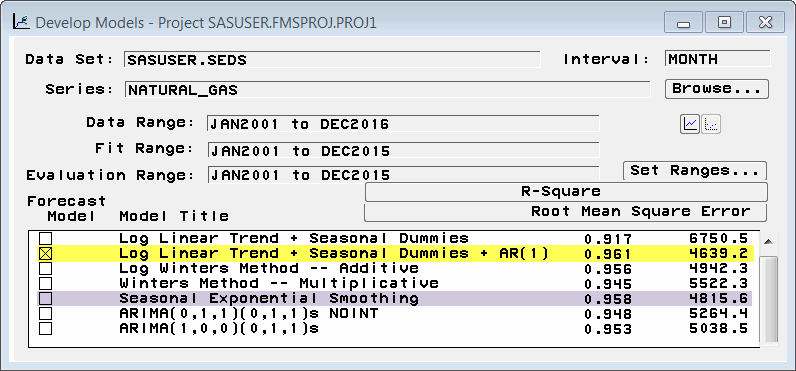


Figure 42 Model Comparison by R2 and RMSE over the 15-year Fit Range

The question arises as to whether or not this deterministic model is still the best fit when compared to other models over the fit range. Fit statistics for each model over the fit range displayed in Figure 42 answer this question. Note that the best fitting series transformation (log or not) are displayed for each model. The only shift in this regard is that over the fit range, the Seasonal Exponential Smoothing model is a better fit than the Log Seasonal Exponential Smoothing model.

Over the fit range, the Log Linear Trend + Seasonal Dummies + AR(1) is still clearly the best fit with the lowest RMSE and the highest R2. Moreover, the Seasonal Exponential Smoothing model remains as our second-best model on both measures. As before, the field becomes crowded after that. This result shows that by common fit metrics over fit and holdout periods, our leading univariate models appear to be fairly robust.

We needed bivariate time series analysis techniques to answer the key questions regarding coal and hydroelectric. Would there be an inverse relationship that might verge on some form of causality between coal and natural gas? And would there be no relationship whatsoever between hydroelectric and natural gas?

For our first question, the answer was no. The transfer function model between natural gas and coal had a holdout range standard error estimate of 4638.9 and a fit range variance of 21,518,977. As expected, this is a little better than the univariate ARIMA model’s RMSE of 4855.2 with a variance of 23,914,031. Therefore, the fit statistics make sense. However, all of the Transfer Function coefficients are positive. Even though the cross-correlation function between coal and natural gas has a significant positive correlation at simultaneous time (lag 0) and an even more significant negative correlation at lag 1, the Transfer Function coefficients did not reflect any inverse relationship at any lag between coal and natural gas. Stated another way, the cross-correlation function implies that the value for coal at time t-1 Granger causes the value for natural gas at time t, but the Transfer Function model seemed to imply otherwise. We probably should not be surprised. The substitution of gas for coal is a very long term trend. These time series models may not be the most appropriate instruments with which to measure such trends that unfold over decades.[[1]](#footnote-1) In the near term, the coefficients in the Transfer Function for coal and natural gas show that what influences coal today also influences natural gas today and even more next month. Longer term trends are opaque to it.

On the second question, the cross-correlation function between natural gas and hydroelectric revealed that there is no relationship between the two time series. Thus, our hypothesis was confirmed. We thought that even without much of a relationship there might be enough of one to merit inclusion of hydroelectric in a multivariate model with coal to predict gas, but in fact, there seems to be no relationship. Thus, we could not establish a multivariate transfer function model involving natural gas, coal, and hydroelectric.

In conclusion, we succeeded in identifying three univariate and one bivariate time series model. Some of these models may be appropriate to forecast future use of natural gas as a fuel for electric power generation one or two years into the future.

1. **Appendix**
2. /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*
3. \* PRODUCT: SAS \*
4. \* VERSION: 9.4 \*
5. \* CREATOR: Kelly Berdelle, Daniel Chen, Kevin Fizgerald, Aida Rojas, Xing Zhang \*
6. \* DATE: May 15,2017 \*
7. \* DESC: Time Series Final Project Part 2: Multivariate Time Series Analysis \*
8. \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/
9. **DATA** Energy;
10. Set SASUSER.SEDS(obs=**180**); /\*HOLDING OUT 12 OBSERVATIONS\*/
11. reformatted\_month=month(Month);
12. TIME=\_N\_;
13. /\*We identified NATURAL GAS as our output series and COAL and HYDROELECTRIC as our input series.\*/
14. /\*OVERVIEW: ACFs, PACFs, IACFs, correlations of variables, and autocorrelation check of residuals\*/
15. **PROC** **ARIMA** DATA=Energy;
16. IDENTIFY VAR=natural\_gas;
17. **RUN**;
18. **PROC** **ARIMA** DATA=Energy;
19. IDENTIFY VAR=coal;
20. **RUN**;
21. **PROC** **ARIMA** DATA=Energy;
22. IDENTIFY VAR=hydroelectric;
23. **RUN**;
24. **PROC** **ARIMA** DATA=Energy;
25. I VAR=natural\_gas CROSSCOR=(coal hydroelectric) NOPRINT;
26. E INPUT=(coal hydroelectric) PLOT;
27. **RUN**;
28. /\*OVERVIEW: Boxplots\*/
30. **PROC** **SORT** DATA=Energy;
31. BY reformatted\_month;
32. **PROC** **BOXPLOT**;
33. PLOT natural\_gas\*reformatted\_month;
34. TITLE ’Seasonal Box Plot for Natural Gas Power Output in GWh’;
35. **RUN**;
36. **PROC** **BOXPLOT**;
37. PLOT coal\*reformatted\_month;
38. TITLE ’Seasonal Box Plot for Coal Power Output in GWh’;
39. **RUN**;
40. **PROC** **BOXPLOT**;
41. PLOT hydroelectric\*reformatted\_month;
42. TITLE ’Seasonal Box Plot for Hydroelectric Power Output in GWh’;
43. **RUN**;
44. /\*STEP 1: CHECK FOR STATIONARITY: output series NATURAL GAS\*/
45. **PROC** **ARIMA** DATA=Energy;
46. IDENTIFY VAR=natural\_gas;
47. **RUN**;
48. /\*STEP 1: CHECK FOR STATIONARITY: input series COAL\*/
49. **PROC** **ARIMA** DATA=Energy;
50. IDENTIFY VAR=coal;
51. **RUN**;
52. /\*STEP 1: CHECK FOR STATIONARITY: input series HYDROELECTRIC\*/
53. **PROC** **ARIMA** DATA=Energy;
54. IDENTIFY VAR=hydroelectric;
55. **RUN**;
56. /\*Series are not stationary. Induce stationarity through differencing.\*/
57. /\*STEP 2: DIFFERENCING\*/
58. **PROC** **ARIMA** DATA=Energy;
59. IDENTIFY VAR=natural\_gas(**1**);
60. IDENTIFY VAR=coal(**1**);
61. IDENTIFY VAR=hydroelectric(**1**);
62. **RUN**;
63. /\*Differenced COAL, NATURAL GAS, and HYDROELECTRIC are not stationary and not white noise.
64. Perform pre-whitening.\*/
65. /\*STEP 3: PREWHITENING: NATURAL GAS\*/
66. **DATA** Energy;
67. Set SASUSER.SEDS(obs=**180**); /\*HOLDING OUT 12 OBSERVATIONS\*/
68. reformatted\_month=month(Month);
69. TIME=\_N\_;
70. **PROC** **ARIMA** DATA=Energy;
71. IDENTIFY VAR=natural\_gas(**12**) NOPRINT;
72. ESTIMATE P=(**1**) Q=(**12**) METHOD=ML; /\*ARIMA(1,0,0)(0,1,1)^12 + C\*/
73. **RUN**; /\*White noise! Maximum Likelihood produces better result than Unconditional Least Squares\*/
74. /\*STEP 3: PREWHITENING: COAL\*/
75. **PROC** **ARIMA** DATA=Energy;
76. IDENTIFY VAR=coal(**12**) NOPRINT;
77. ESTIMATE P=(**1**) Q=(**12**) METHOD=ML; /\*ARIMA(1,0,0)(0,1,1)^12 + C\*/
78. **RUN**; /\*White noise! Maximum Likelihood produces better result than Unconditional Least Squares\*/
79. /\*STEP 3: PREWHITENING: HYDROELECTRIC\*/
80. **PROC** **ARIMA** DATA=Energy;
81. IDENTIFY VAR=hydroelectric(**12**) NOPRINT;
82. ESTIMATE P=(**1**)(**12**) NOINT METHOD=ML; /\*ARIMA(1,0,0)(1,1,0)^12 NOINT\*/
83. **RUN**; /\*White noise!\*/
84. /\*STEP 4: 1-INPUT TRANSFER FUNCTION MODEL 1: NATURAL GAS vs COAL\*/
85. **PROC** **ARIMA** DATA=ENERGY;
86. IDENTIFY VAR=natural\_gas(**12**) NOPRINT;
87. IDENTIFY VAR=coal(**12**) NOPRINT;
88. ESTIMATE P=(**1**) Q=(**12**) METHOD=ML;
89. IDENTIFY VAR=natural\_gas(**12**) CROSSCOR=(coal(**12**));
90. **RUN**;
91. /\*The ACF is stationary except at lag 12.\*/
92. /\*The CCF shows first response at lag 0, drop at lag 1, and no patterns: b=0, r=0, s=1\*/
93. **PROC** **ARIMA** DATA=ENERGY;
94. IDENTIFY VAR=natural\_gas(**12**) NOPRINT;
95. IDENTIFY VAR=coal(**12**) NOPRINT;
96. ESTIMATE P=(**1**) Q=(**12**) METHOD=ML;
97. IDENTIFY VAR=natural\_gas(**12**) CROSSCOR=(coal(**12**)) NOPRINT;
98. ESTIMATE INPUT=((**1**)/coal) P=(**1**) Q=(**12**) METHOD=ML PLOT; /\*b=0, r=0, s=1\*/
99. FORECAST LEAD=**12** OUT=results;
100. **RUN**;
101. /\*STEP 5: ADEQUACY CHECK: NATURAL GAS vs COAL\*/
102. /\*The error model behaves white noise.\*/
103. /\*STEP 6: 1-INPUT TRANSFER FUNCTION MODEL 2: NATURAL GAS vs HYDROELECRTRIC\*/
104. **PROC** **ARIMA** DATA=ENERGY;
105. IDENTIFY VAR=natural\_gas(**12**) NOPRINT;
106. IDENTIFY VAR=hydroelectric(**12**) NOPRINT;
107. ESTIMATE P=(**1**)(**12**) NOINT METHOD=ML; /\*ARIMA(1,0,0)(1,1,0)^12 NOINT\*/
108. IDENTIFY VAR=natural\_gas(**12**) CROSSCOR=(hydroelectric(**12**));
109. **RUN**;
110. /\*The ACF is stationary except at lag 12.\*/
111. /\*The CCF shows no responses at any lags, no exponential decay, and no patterns: d=0, r=0, s=0\*/
112. /\*STEP 7: 2-INPUT TRANSFER FUNCTION MODEL: NATURAL GAS vs COAL vs HYDROELECRTRIC\*/
113. /\*PROC ARIMA DATA=ENERGY;
114. IDENTIFY VAR=natural\_gas(12) NOPRINT;
115. IDENTIFY VAR=coal(12) NOPRINT;
116. ESTIMATE P=(1) Q=(12) METHOD=ML NOPRINT;
117. IDENTIFY VAR=hydroelectric(12) NOPRINT;
118. ESTIMATE P=(1)(12) NOINT METHOD=ML NOPRINT;
119. IDENTIFY VAR=natural\_gas(12) CROSSCOR=(coal(12) hydroelectric(12)) NOPRINT;
120. ESTIMATE INPUT=((1)/coal, 1$/hydroelectric) P=(1) Q=(12) METHOD=ML PLOT;
121. RUN;\*/

1. As a follow-up to this project, if these time series are used for similar purposes again it may be worthwhile to attempt to model the relationship between Coal and Gas using a Step function in 2005. Why? Because the real modern Gas Rush began in 2005. See John Richardson’s, The History of Fracking (A Timeline), <https://energywithjr.quora.com/The-History-of-Fracking-A-Timeline>, where he states: *“Fracking boomed after the Energy Policy Act in 2005 exempted it from compliance with the Safe Drinking Water Act, the Clean Air and the Clean Water Act. Also, the CERCLA Superfund Act doesn't cover fracking sites.”* [↑](#footnote-ref-1)