James Laughead

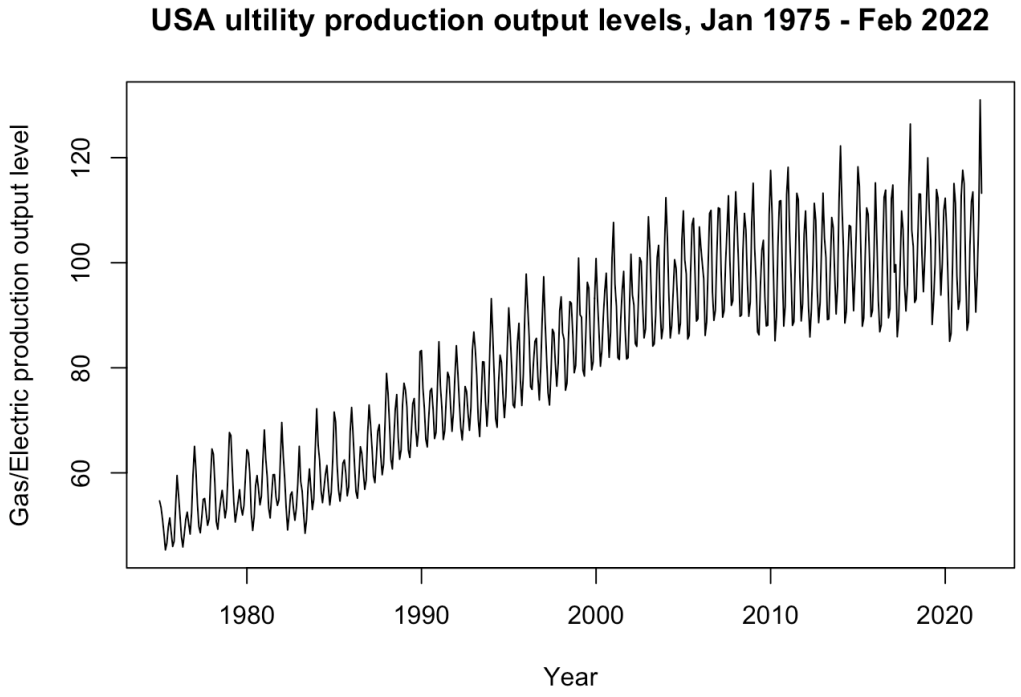
Dena Asta

STAT 5550

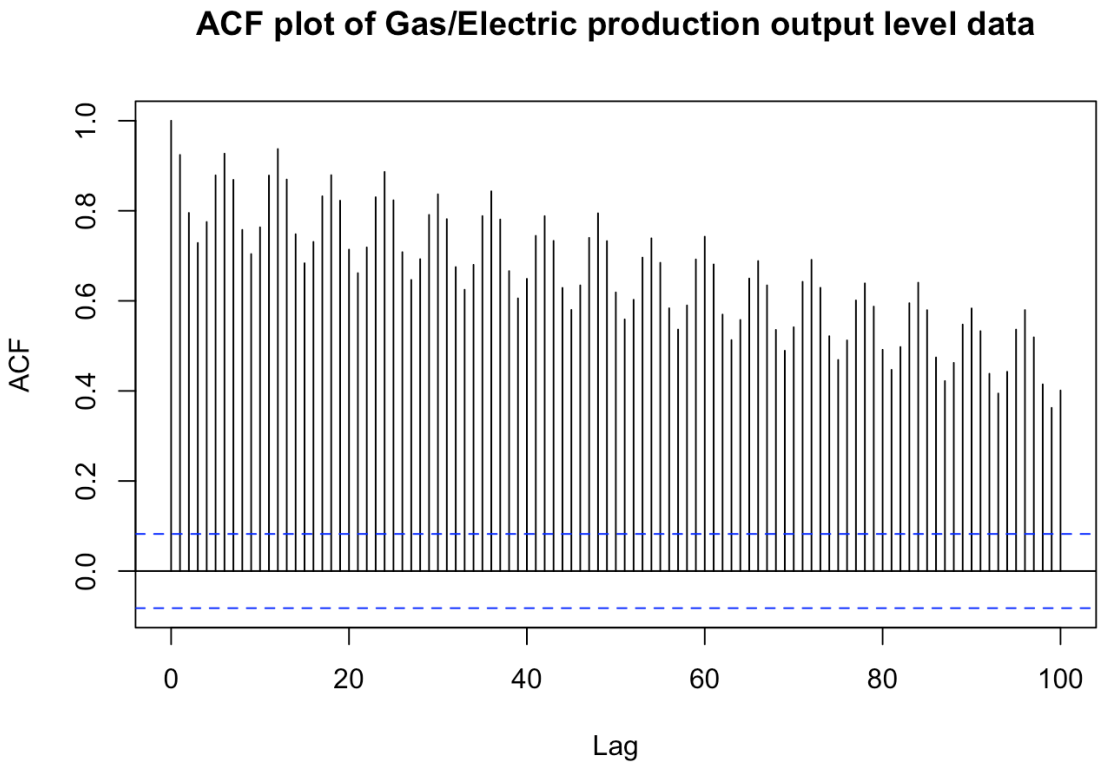
25 April 2022

A Look at USA utility output

The dataset which we are looking at in this analysis records the electricity/gas utility production output of all producers in the United States excluding territories, and has been gathered monthly from January 1975 to February 2022. The dataset was obtained from the FRED economic data website. Our goal is to use this data in order to determine a range of likely production output levels we can expect in future years. This could be useful when comparing this data to different aspects such as population growth, to see if there is an eventual max out point in the production output of utilities in the United States. A time series plot of this data can be seen in the figure below.



From a simple observation of the time series data, it is clear that that data is non stationary, as the mean is not constant and instead shows an increase in Gas/Electric production output level as the years progress. We can confirm this time series is non stationary with an ACF plot of the data, shown below.



The ACF of the data decreases slowly as the number of lags increases, which indicates that the time series is non-stationary. In this ACF plot, as the number of lags increases, there are also oscillations in the ACF values. This indicates seasonality in the data, another component which makes this dataset non-stationary, as the values are dependent on time.

Since we have identified both a trend in the data, and a seasonal component, we will have to address both of these non-stationary features. The data appears to be linear, so I decided not to apply any transformations to the raw data for analysis. Also, since the data is linear, I estimated the trend using regression methods with only t, while also modeling the seasonal component with

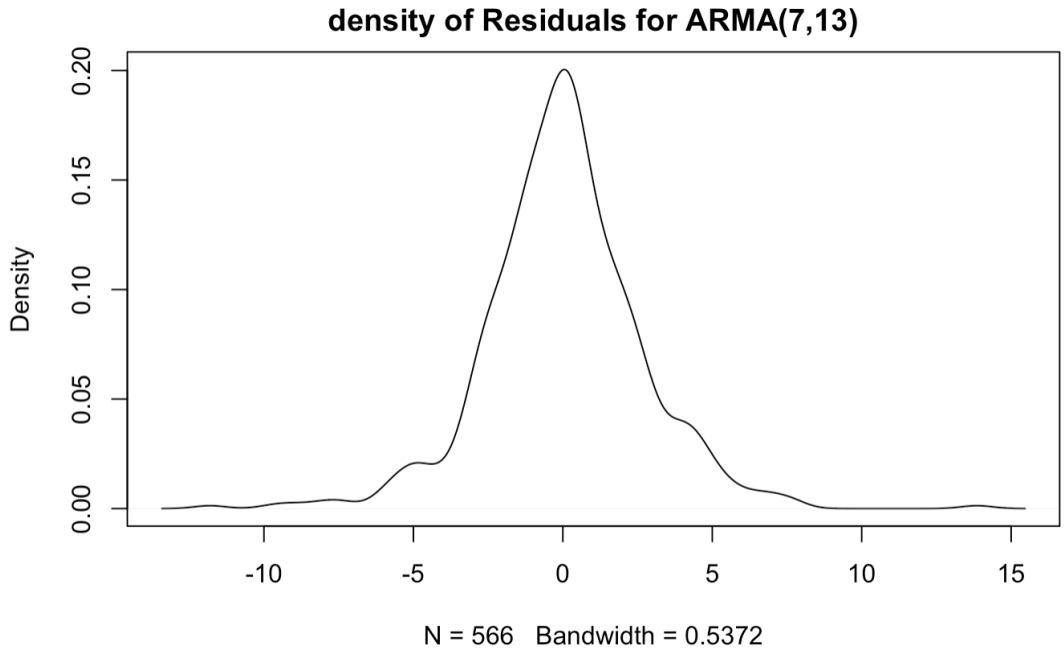
the function $c = cos(2\pi t/_{12})\ and\ s = sin(2\pi t/_{12})$, two functions which repeat after every cycle of 12, considering that our data is collected monthly and that there are twelve months in a year. After using regression modeling on t, and the functions c and s, the estimated trend was $0.107923t + 51.60552$ and the seasonal component was $(2.990434 * cos(2\pi t/_{12}) + 0.77011133 * sin(2\pi t/_{12})$.

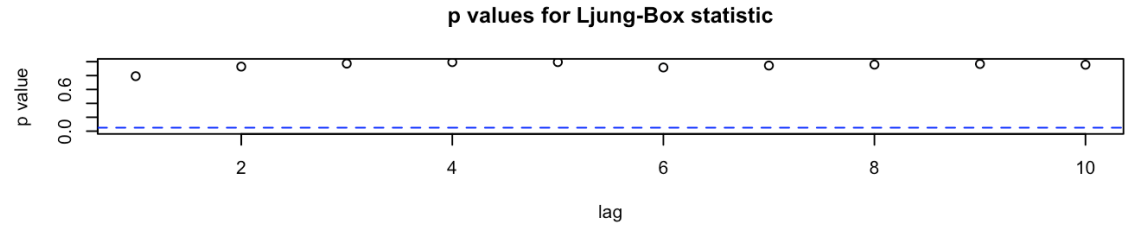
After obtaining these estimates, we can now detrend and deseasonalize the data in order to use ARMA modeling techniques on this data. Fitting several ARMA models, the first step I observed in selecting an appropriate model was the AIC of the models. A table of several ARMA models and their respective AIC is listed below. Coefficients are excluded from this graph due to the size of some of the ARMA models.

ARMA(p,q)	AIC
ARMA(1,1)	3539.58
ARMA(0,1)	3608.24
ARMA(6,11)	2747.48
ARMA(7,12)	2768.19
ARMA(7,13)	2745.69

In terms of AIC, the ARMA models with higher p and q were much better fits than those of lower p and q, and the best model in terms of AIC I tested was an ARMA(7,13) model. In order to see if this model was an appropriate fit, I performed a residual plot of the density, as

well as a Ljung box test. The plots of both of these are displayed below.





The density of the residuals for the ARMA(7,13) plot was roughly normal with a mean of 0, and the p values for the Ljung-Box test were not significant for many lags, which shows that and ARMA(7,13) model is indeed appropriate for this deseasonalized, detrended series. Combining everything, we can create an estimated model for our dataset, which is

$y_t = 0.107923t + 51.60552 + s_t + x_t$,

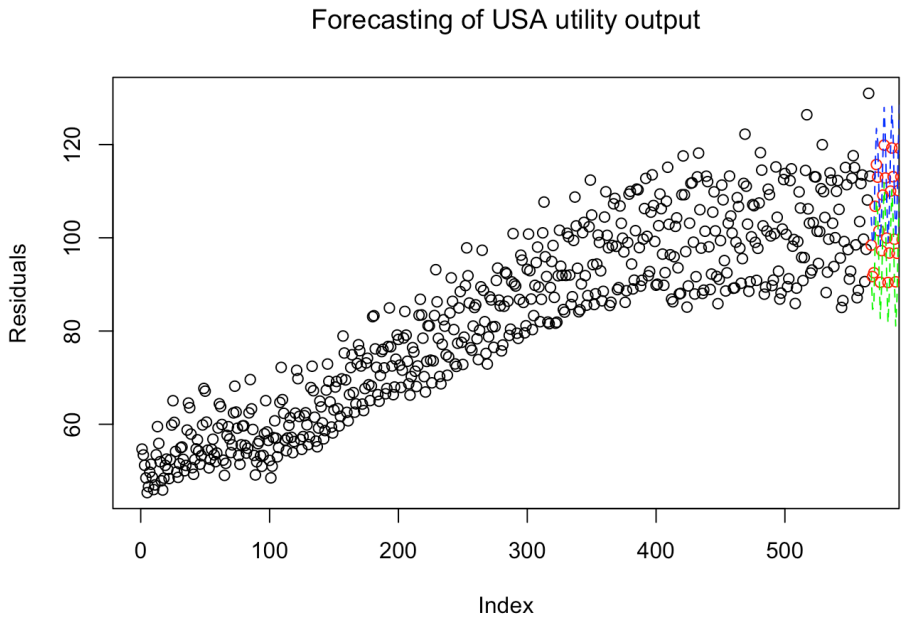
$s_t = (2.990434 * cos(2\pi t/12) + 0.77011133 * sin(2\pi t/12)$

x_t = -0.6879x_{t-1} + 0.2744x_{t-2} +-0.0565x_{t-3} + 0.2584x_{t-4} + 0.3066x_{t-5} + 0.9127x_{t-6} + 0.4488x_{t-7} +

1.1704w_{t-1} + 0.2933w_{t-2} + 0.2877w_{t-3} + 0.5759w_{t-4} + 0.1428w_{t-5} + -0.7627w_{t-6} + -0.7347w_{t-7} +

$-0.2719w_{t-8} + -0.1299w_{t-9} + -0.0845w_{t-10} + -0.1225w_{t-11} + 0.1434w_{t-12} + 0.0388w_{t-13} + w_t$, $w_t \sim$ iid N(0,6.783)

Now, we can forecast out several time periods and predict future data. In the graph below, the predictions themselves are the red points in the graph, for months of data predicted beyond February 2022. The dashed lines in the graph represent a 95 percent prediction interval for the data.

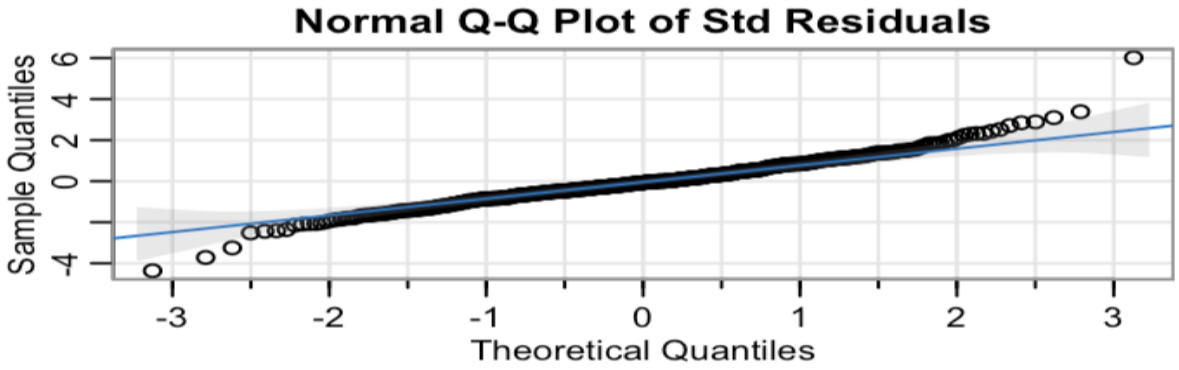


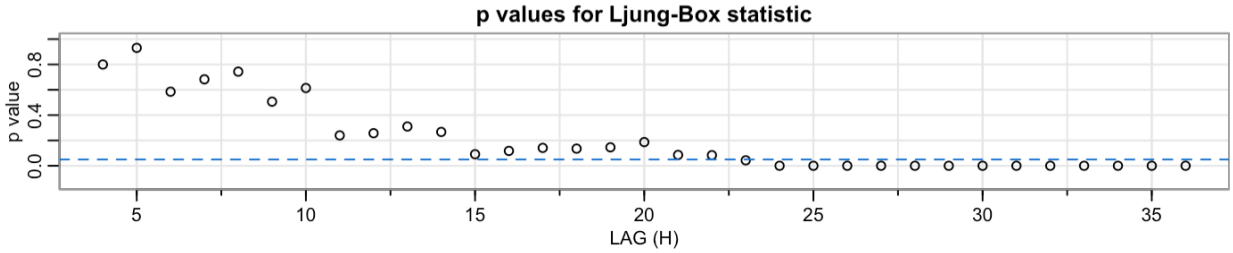
I also performed a second analysis of the data, where instead of deseasonalizing and detrending the data like I did with the ARMA model, I used differencing to achieve a similar result in the form of a SARIMA model. The table below lists several different SARIMA models I tested, with their respective AIC

SARIMA(p,d,q) x $(P,D,Q)_{12}$	AIC
SARIMA(1,1,1) x $(0,1,1)_{12}$	4.64355

SARIMA(1,2,1) x $(0,1,1)_{12}$	4.832716
SARIMA(1,1,1) x $(1,1,1)_{12}$	4.643033
SARIMA(1,2,1) x $(1,2,1)_{12}$	5.181823
SARIMA(2,1,2) x $(1,1,1)_{12}$	4.648659

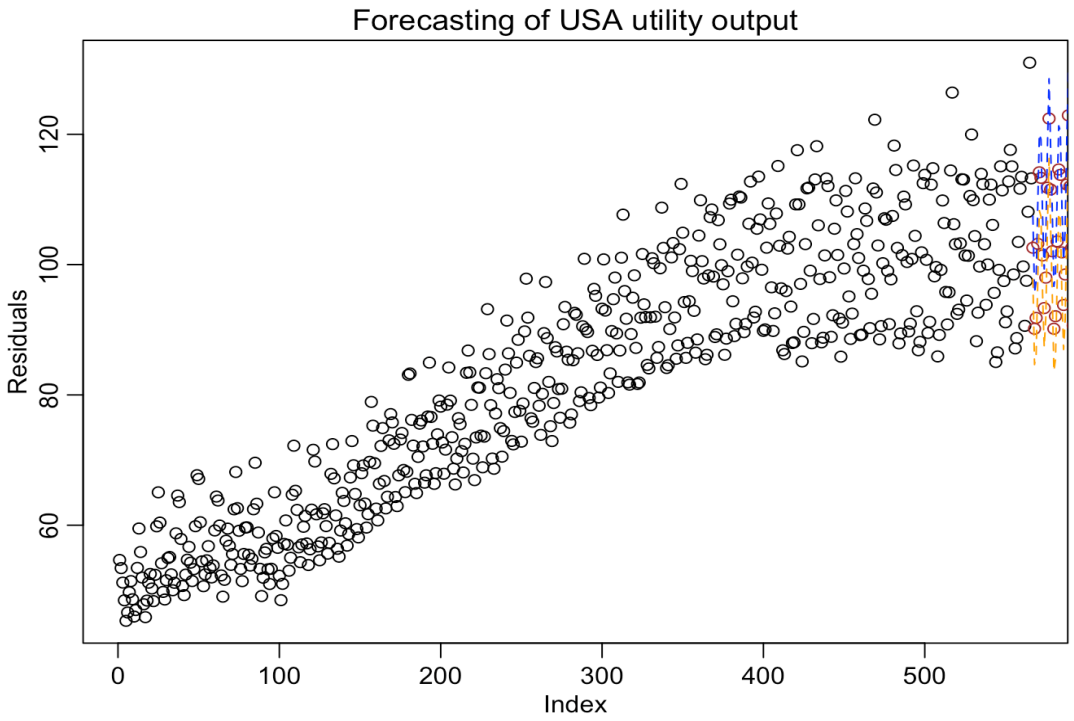
In terms of AIC, the SARIMA(1,1,1) x $(1,1,1)_{12}$ appears to be the best model fit. In order to see if this model fits the data well, I ran a qq plot of the residuals, and a Ljung-box test, both of which are shown below.





The qq plot of the residuals contains mostly all values which are close to the qq line, and the p values for the Ljung-box statistic are not significant for many lags, indicating that this model is a good fit for the data. A final model for our SARIMA(1,1,1) x $(1,1,1)_{12}$ can be written as $(1 - .4474B)(1 + .0789B^{12})\nabla_{12}x_t = (1 - .0.7888B^{12})(1 - 0.9349B)w_t$, $w_t \sim N(0,5.842)$. Using this model, we can use forecasting to predict what future values will be. A graph of the data with several

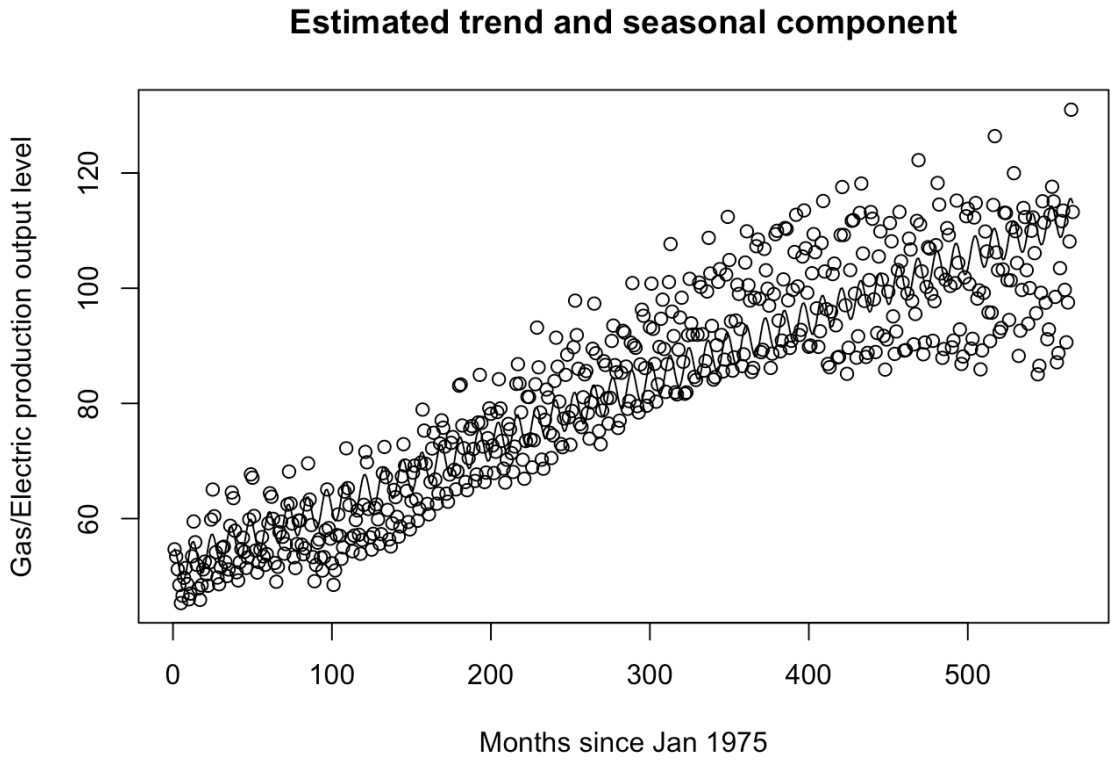
future predictions of the data, as well as upper and lower bounds creating a 95% prediction interval are displayed below.

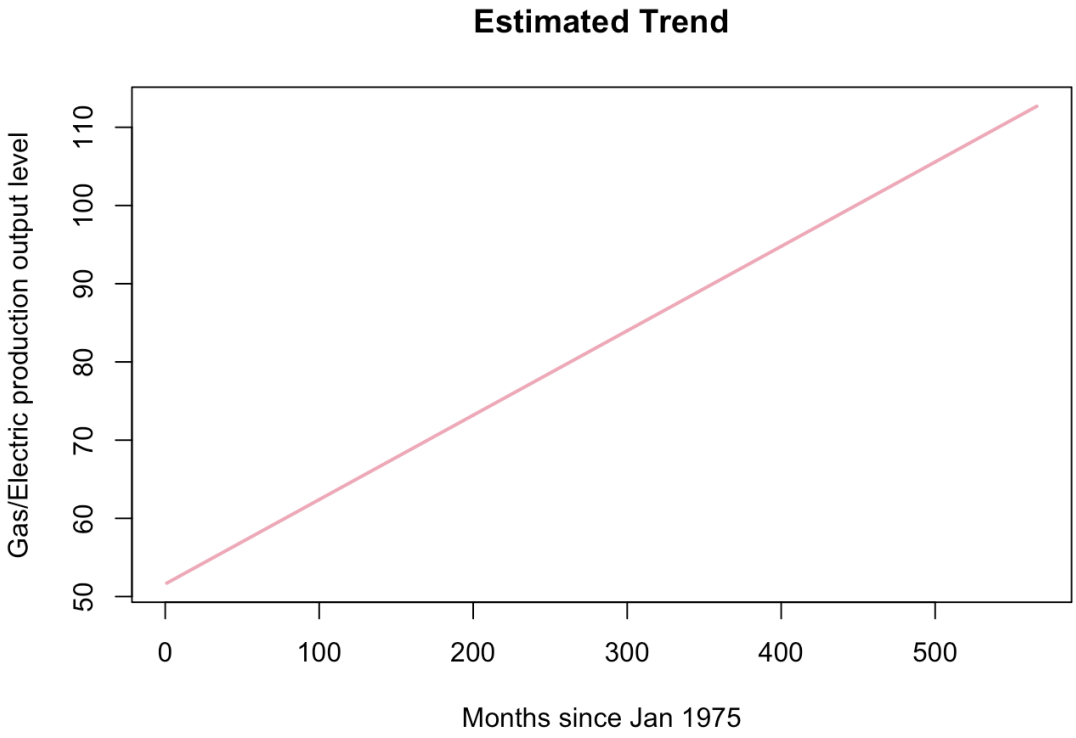


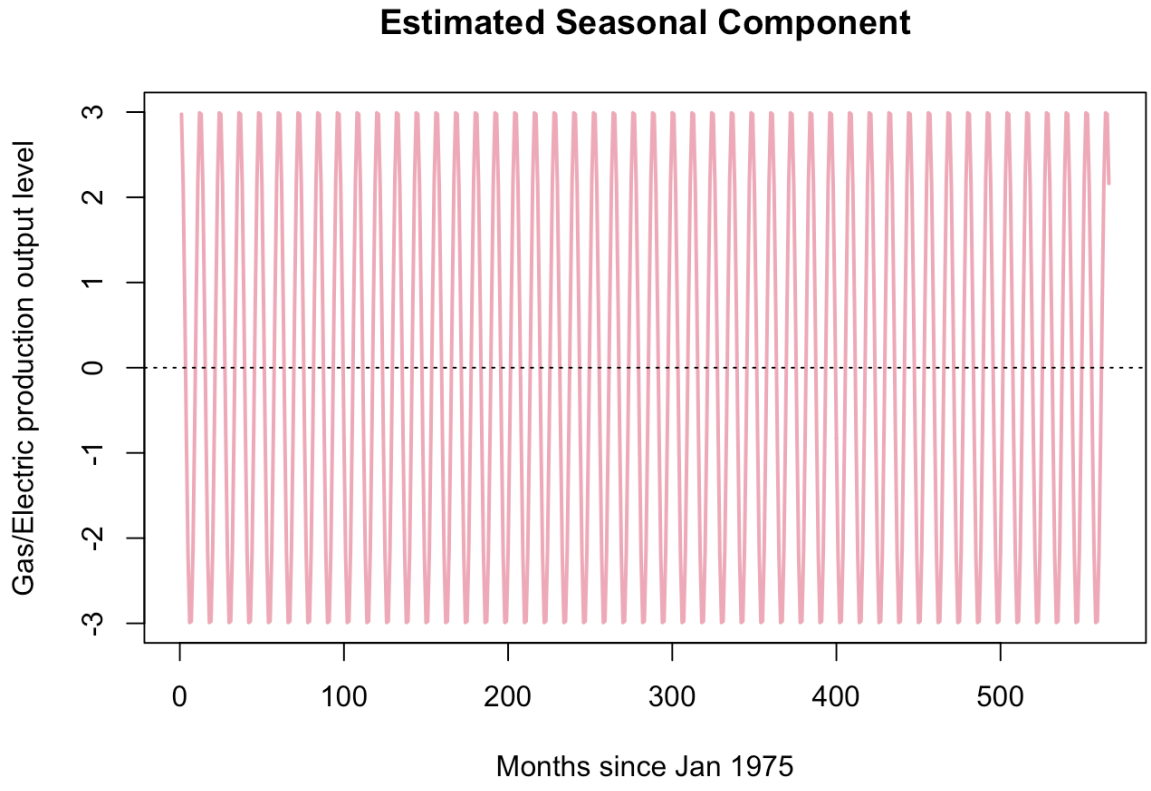
After fitting both an ARMA and an SARIMA model for this data, the future predictions achieved from forecasting were both very similar. Both forecasts predicted the seasonality to continue as well as the general trend. The prediction intervals created from the two models were also very similar. Something which differed greatly in the two models was the AIC values. The ARMA models which produced values with lower AIC included many coefficients, while the SARIMA models which produced the least AIC had few coefficients. Also, the AIC values for the SARIMA model were much lower across the board when compared to the AIC for the ARMA models. This suggests that differencing the data was very useful in creating an efficient and effective model. I prefer the SARIMA model because of the lower AIC values, as well as it was computationally quicker. Perhaps an ARMA model with even greater number of coefficients than the one I chose would have yielded lower AIC and a better fit, but with a lot of coefficients the computations take a long time for a computer, which is another reason why I would prefer using SARIMA for this dataset.

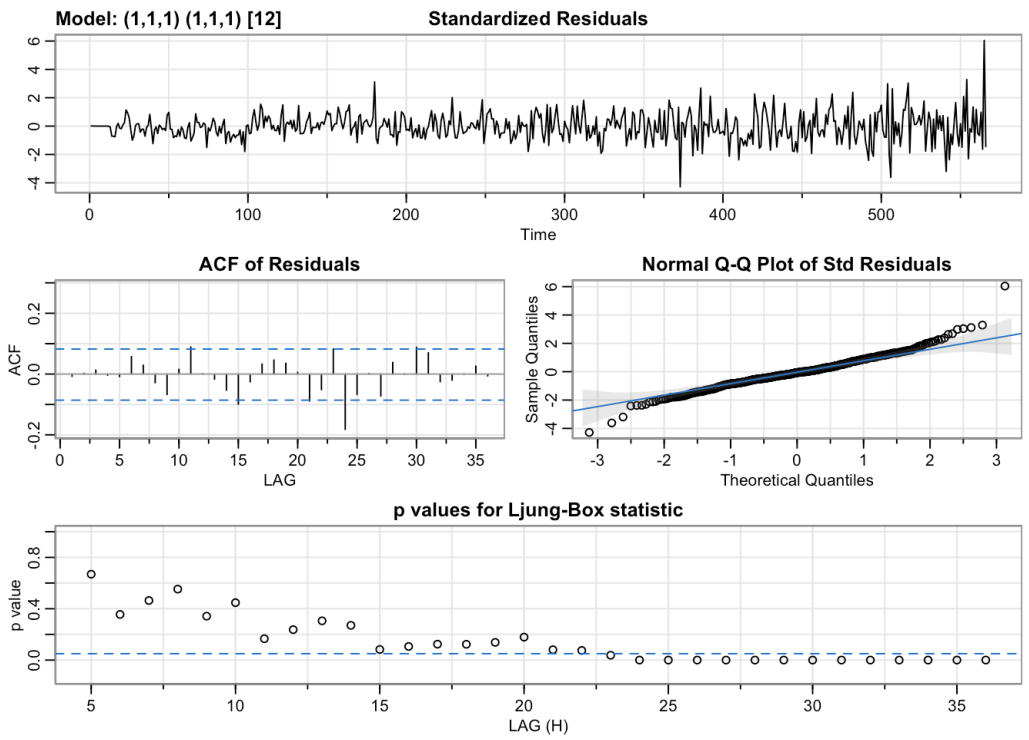
Overall, we achieved the same results from forecasting with the ARMA and SARIMA models. The utility production in the United States is predicted to maintain its seasonal pattern, and also its positive trend in the near future. It will be interesting to observe this data in the future, to see if the positive trend increases as predicted, or potentially hits a max out point and remains constant.

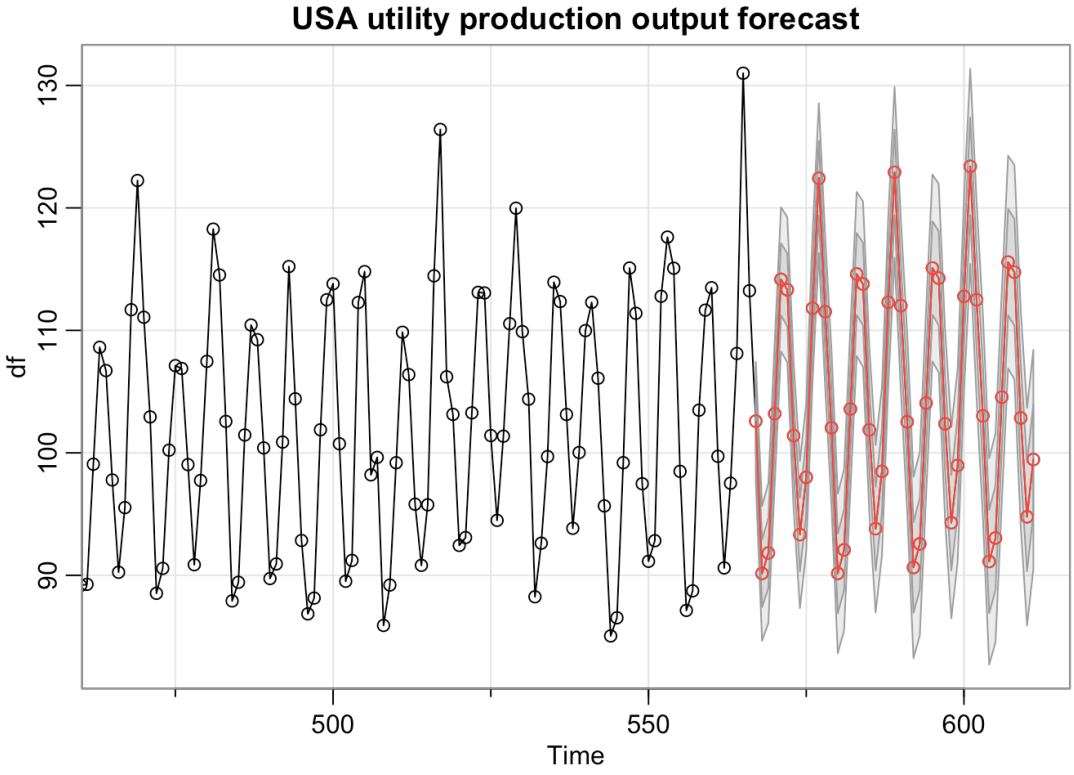
Appendix


Estimated trend and seasonal component
Gas/Electric production output level
60
80
100
120
0
100
200
300
400
500
Months since Jan 1975


Estimated Trend
Gas/Electric production output level
50
60
70
80
90
100
110
0
100
200
300
400
500
Months since Jan 1975


Estimated Seasonal Component
Gas/Electric production output level
-3
-2
-1
0
1
2
3
0
100
200
300
400
500
Months since Jan 1975


Model: (1,1,1) (1,1,1) [12]
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
LAG (H)


USA utility production output forecast
df
Time