Forecasting United States Energy Consumption Trends: A SARIMA Approach

Analyzing Patterns, Predicting Future Demands

Author(s): Artemiy Yalovenko, Raghav Chhabra

Date: 03/22/2024

Institution: DePaul University

Abstract

The paper's goal is analysis and forecast of United States energy consumption from January 2000 to December 2023, using a Seasonal Autoregressive Integrated Moving Average (SARIMA) model. This paper uses a compiled monthly series of total US energy consumption, obtained from the Energy Information Administration (EIA), to fit a prediction model that captures seasonal variations and time-series trends. The SARIMA model was identified by following the Box-Jenkins methodology. After ensuring data is stationary, we analyzed autocorrelation and partial autocorrelation plots of a transformed version of the original data to identify initial model orders. In the project the data was transformed with log transformation and was differenced to account for seasonality and to achieve stationarity. To estimate parameters, we used Maximum Likelihood Estimates (MLE) to determine parameter significance, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These values led us to explore more models and pick the final one. Each model's residuals were examined to determine if residuals resemble white noise. The selected model was applied to predict and visualize multiple time points in the future alongside a prediction and confidence interval calculation.

Introduction

Forecasting of energy consumption is critical in energy management, economic and national security policy formulation. Energy consumption has historically reflected economic activity and growth, technological advancements, population growth and lifestyle patterns, and used to analyze environmental impact. All in all, energy consumption in the US represents a critical factor of the nations' economic vitality.

This paper focuses on the analysis of total energy monthly consumption from 2000 to 2023, provided by the Energy Information Administration (EIA). The dataset contains information encompassing seasonal variations, economic cycles, and policy changes over the last two decades. Understanding these underlying trends is important to anticipate future energy needs of the nation, but also to aid in policy ideation or decision-making for individuals trying to leverage this knowledge.

To forecast future energy consumption, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model was employed. The SARIMA model is an extension of ARIMA model framework, incorporating complex seasonal elements into model structure. SARIMA models are often employed in fields of finance, economics, biology, and even geography. Since energy consumption (in the USA) is clearly dependent on time of year, SARIMA model is a perfect tool to use.

SARIMA model is specified by notation SARIMA(p,d,q)(P,D,Q)(S). The 'p', 'd', and 'q' are the non-seasonal components. They specify the order of autoregression, the degree of differencing, and the order of the moving average process. The 'P','D', 'Q' specifies the seasonal components of the order of autoregression, the degree of differencing, and the order of the moving average process. Finally, 'S' is the seasonality period that can be hourly, daily, quarterly or whatever seasonality frequency the data dictates.

The 'd' and 'D' are the differencing steps which allow for stabilization of the mean and seasonal patterns. The autoregressive components predict values using priorly seen values, while the moving average components use the forecasting errors to adjust the predictions.

The fitting process starts with identifying the differencing steps by estimating the best 'd' and 'D'. Then, by selecting the best seasonal and non-seasonal autoregression, the degree of differencing, and the order of the moving average process parameters by assessing a myriad of statistical measures to ensure that each order contributes meaningfully to the model while not violating critical assumptions. This paper covers our assessment of differencing orders in the model specification section and the assignment of the other orders in the fitting and diagnostics while also explaining the meanings behind the numbers.

Model specification:

Original Data:

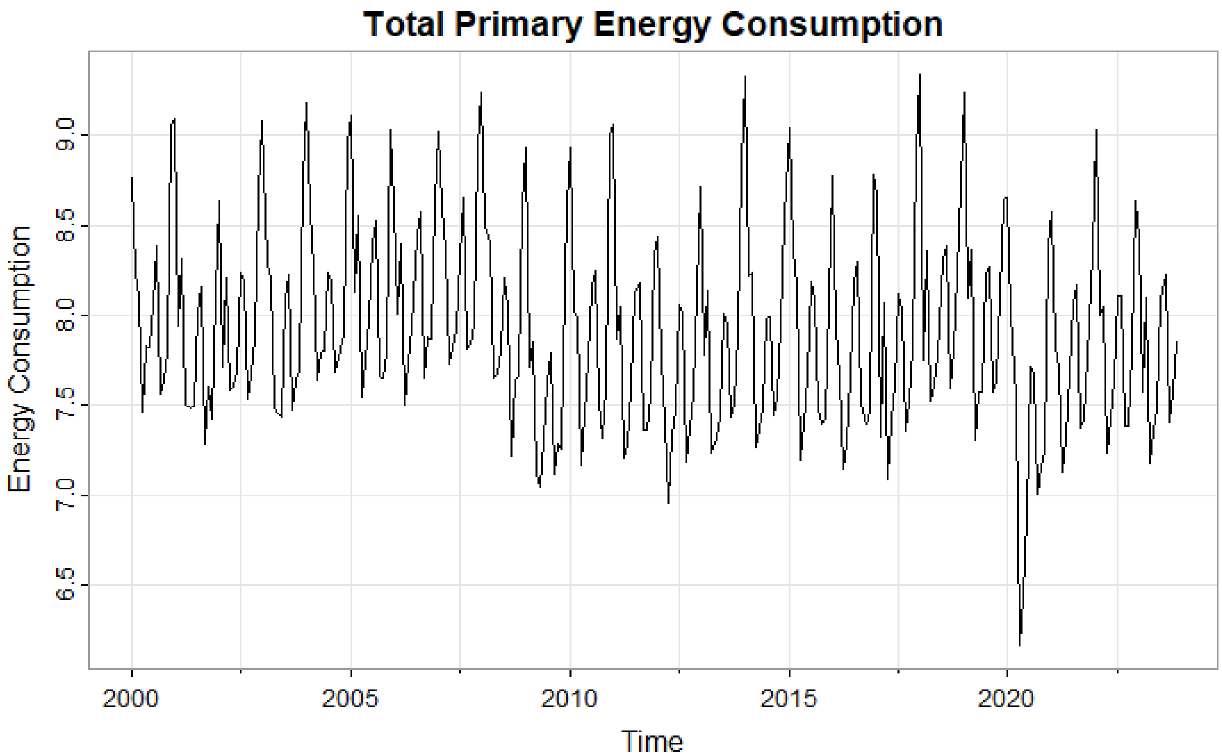


Figure 1: Energy Consumption by time

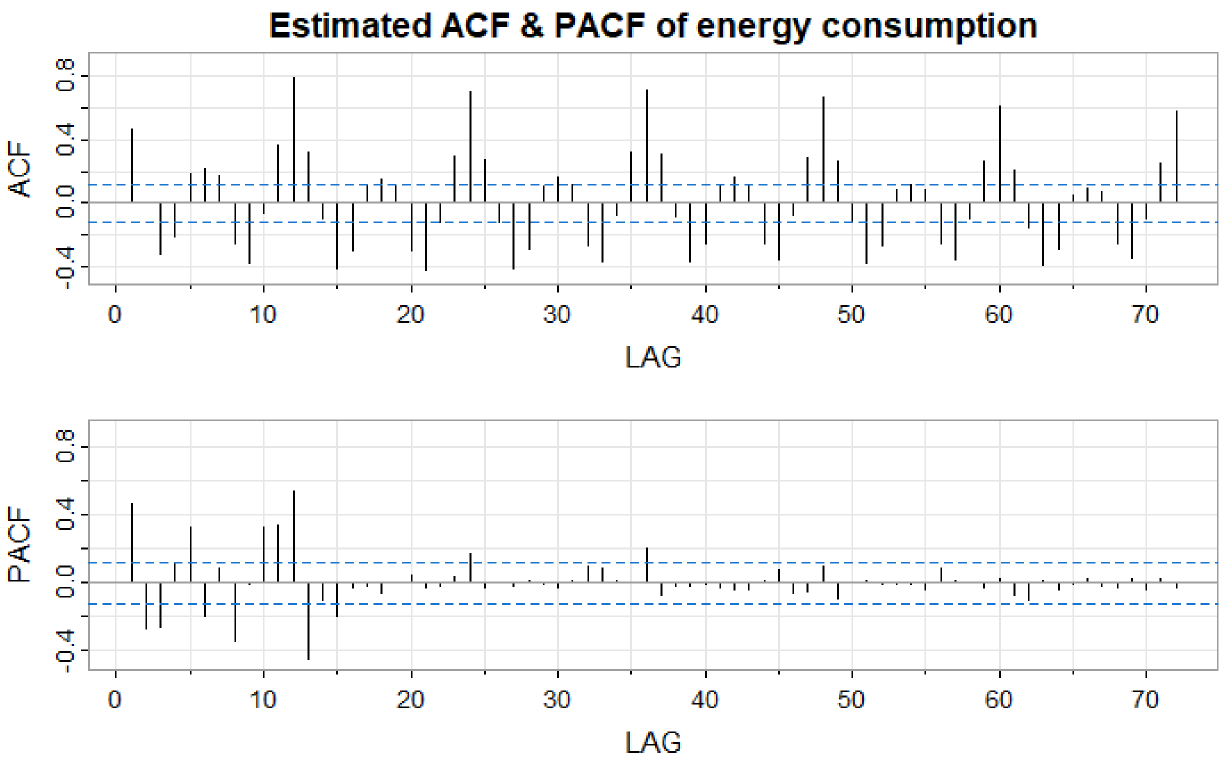


Figure 2: Estimated ACF & PACF of original Data

From Figure 1 we can see that our data is not stationary. In Figure 2 we can see the yearly seasonality on every 12 lags in ACF. To make data stationary we will differentiate the data. We will difference once on the seasonal section of the data and check the stationary assumption. If the assumption is not met, we will do double differencing. It is critical that we ensure stationarity as it is critical to SARIMA model implementation.

One Seasonality Differences Data - (1-B^12)*xt:

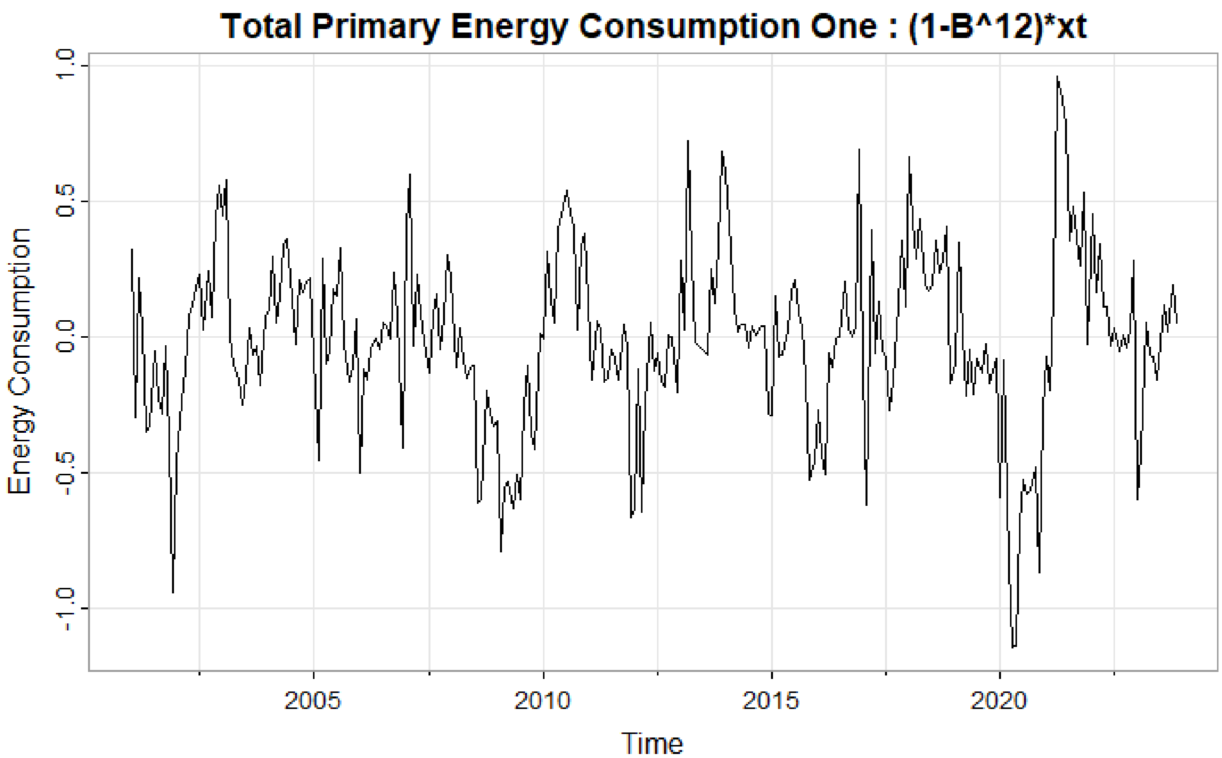


Figure 3: Energy consumption over time seasonally differenced

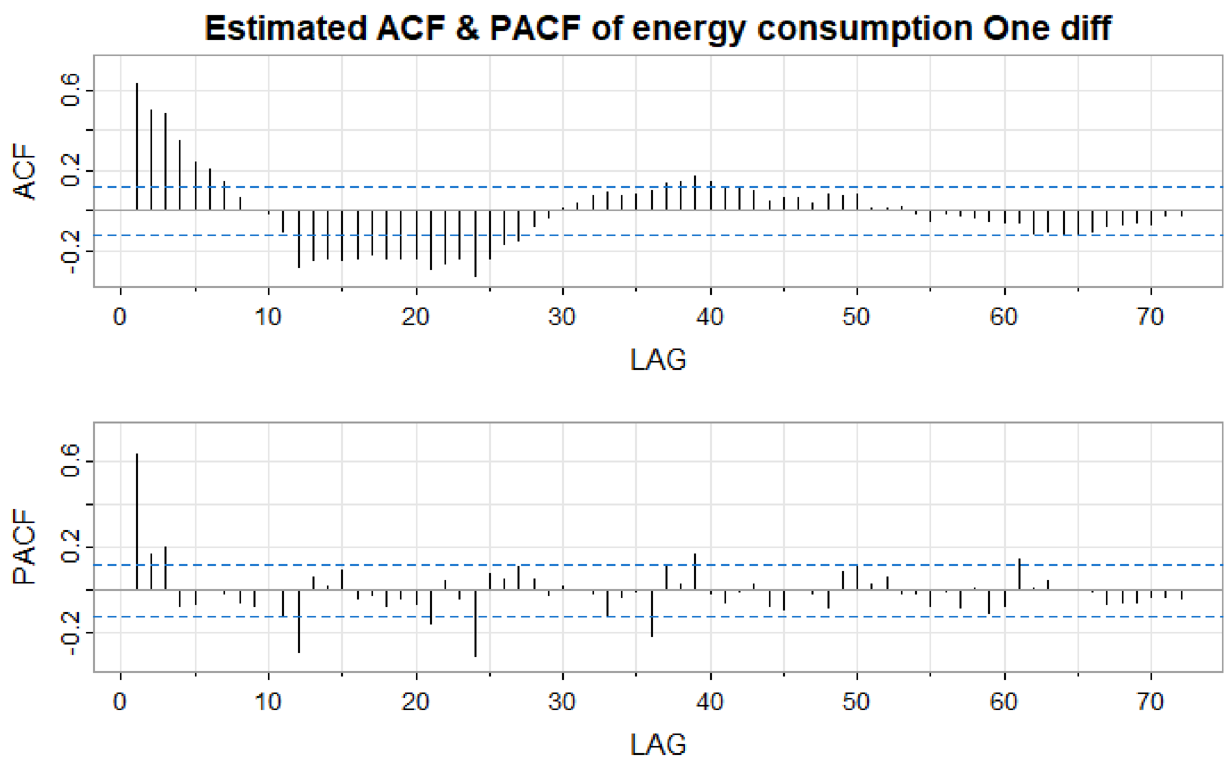


Figure 4: Estimated ACF & PACF of original Data

From Figure 3 and Figure 4 we can confirm that our data is almost stationary with a couple large spikes which could be an outlier. From Figure 4, we can see that the data doesn't seem to be stationary i.e. doesn’t have constant mean, constant variance and more vividly there is no constant autocorrelation structure as can be seen in the ACF plot. We next attempt to adjust data to achieve constant variance by first applying a log transformation.

Log data:

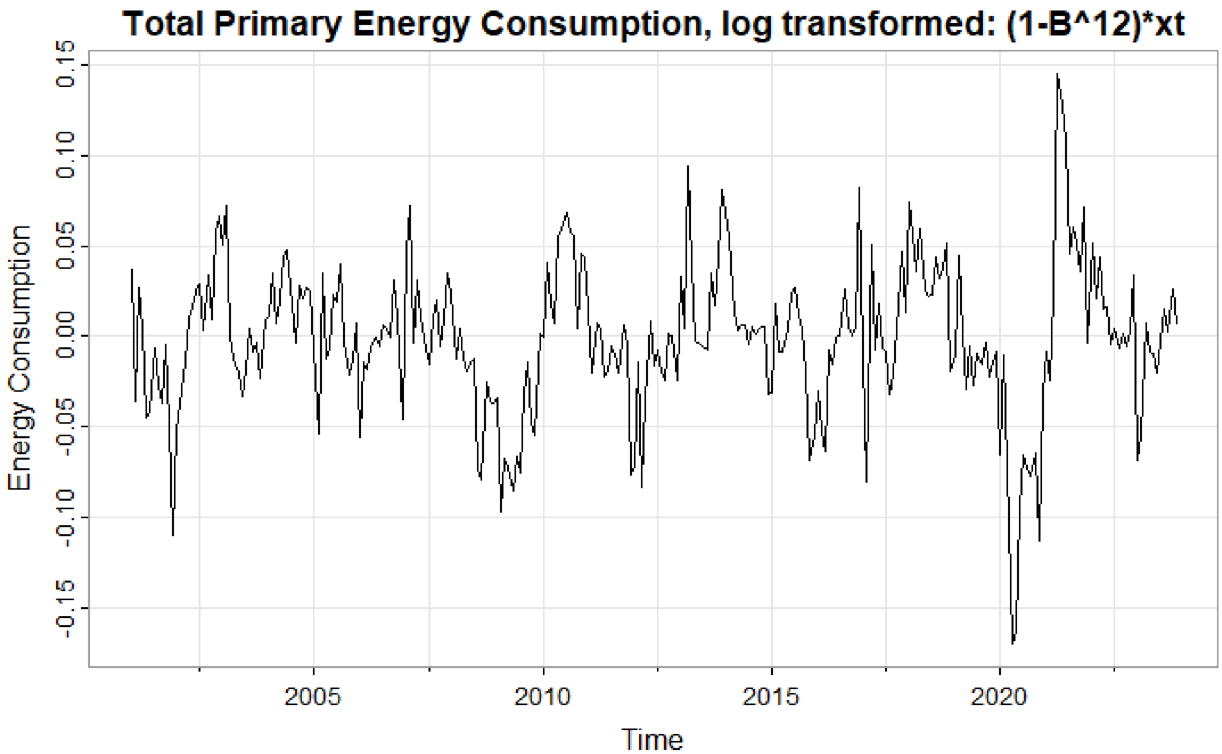


Figure 5: Energy consumption over time, log transformed and seasonally differenced

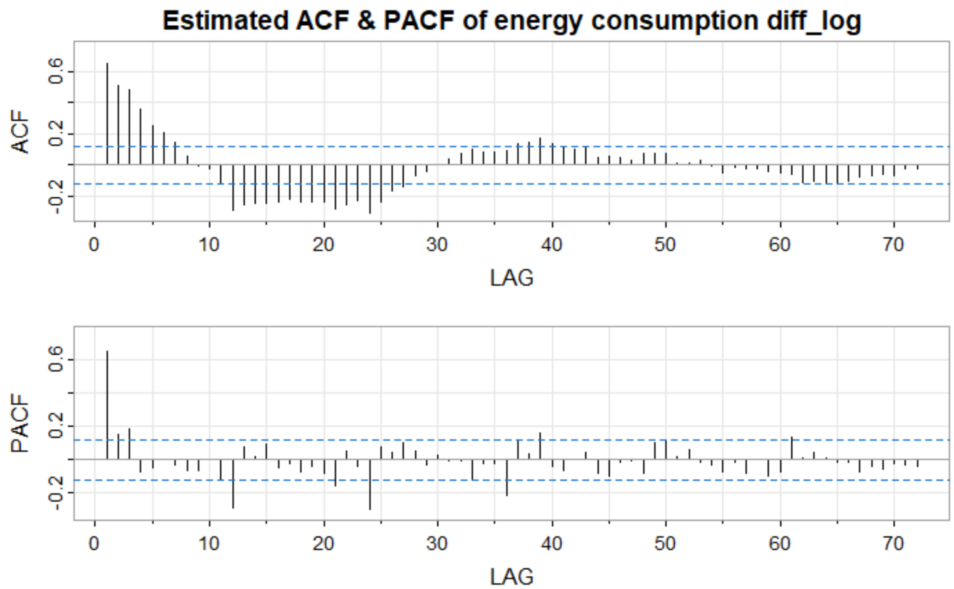


Figure 6: Estimated ACF and PACF Plots, log transformed and seasonally differenced

Log transformation had no impact on overall stationarity. Instead we attempt to difference the original data to achieve more suitable ACF and PACF values for further model building.

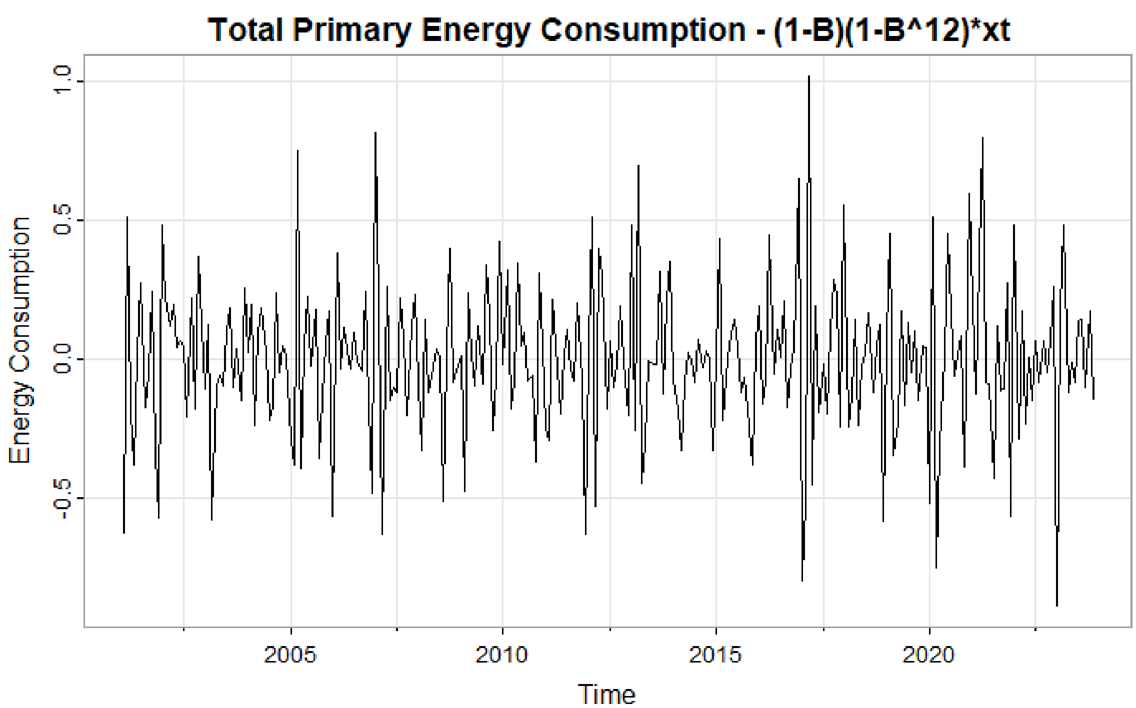


Figure 7: Energy Consumption data differenced for seasonality and stationarity

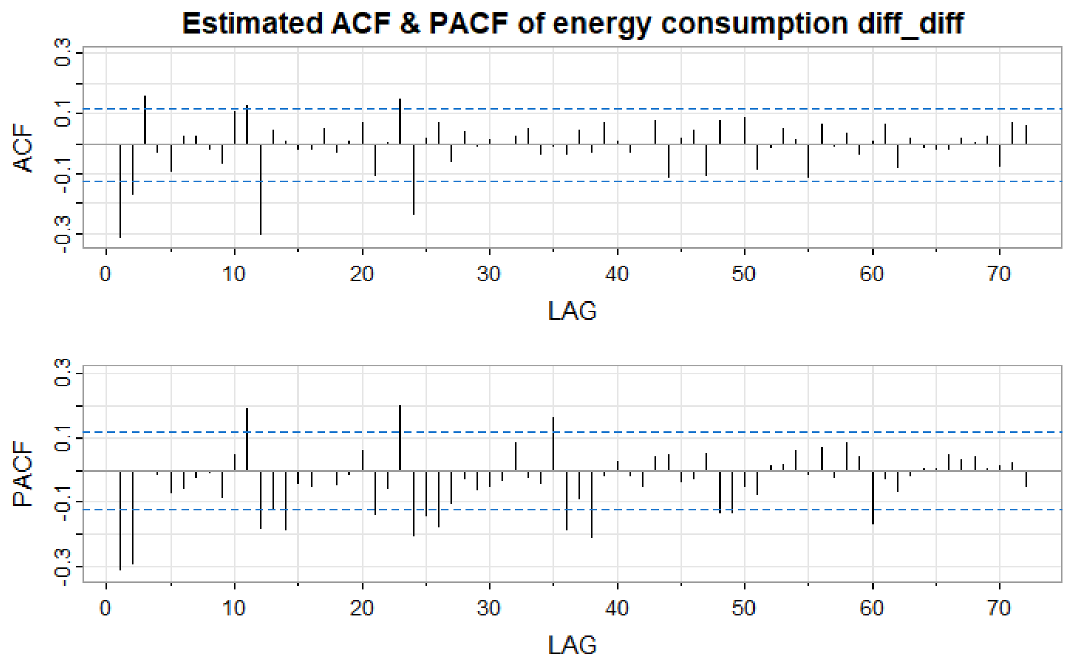


Figure 8: Estimated ACF and PACF plots on data differenced for seasonality and stationarity

After applying differencing to the seasonality stationary data, we notice a significant change in the way data looks. The data is now constant around the mean and constant in variance over time. Additionally, ACF and PACF plots confirm constant autocorrelation structure and is now also interpretable for us to inference potential first model order.

Looking for potential seasonal components which could fit well into a proposed SARIMA model, the ACF plot has high auto correlations at lags 12 and 24, indicating that seasonal moving average order (Q) may equal to 2. From the same plot we may also infer that moving average order (q) may be 2 or 3.

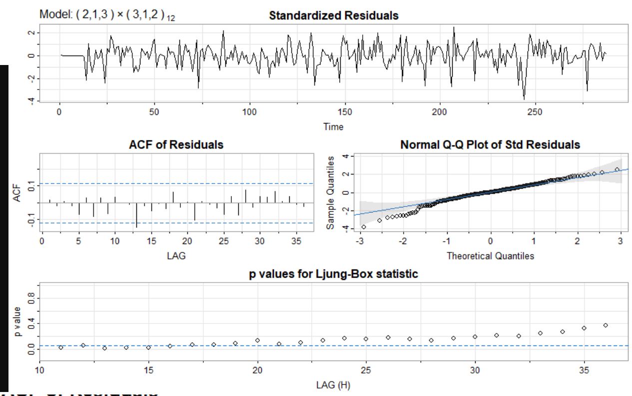
From the PACF plot we can infer that seasonal autoregressive order (P) may be equal to 3 as there are significant values at lag 12,24,36,48, and 60. From the PACF plot we may also assume that autoregressive order (p) is equal to 2 since there are 2 significant initial lags after which there is a general decay to zero.

Fitting and Diagnostics:

We carefully developed a statistical model that could provide reliable predictions at this phase of the project. Our goal in exploring the fields of diagnostics and model fitting was to identify a robust model while acknowledging the unavoidable errors in any statistical analysis. In this phase, we will be assessing if the assumption of residual independence is satisfied. There should be no autocorrelation in the residuals since it would mean that our model captured all the time series dynamics and there is nothing left over. We also want to make sure that residuals are normally distributed which would ensure that the model's errors are random and not systematic.

Model 1 <- mod.fit.213.312

```
$ttable
     Estimate     SE  t.value p.value
ar1   -0.6897 0.0597 -11.5624  0.0000
ar2   -0.9541 0.0837 -11.3949  0.0000
ma1    0.1969 0.0809   2.4339  0.0156
ma2    0.5670 0.1068   5.3113  0.0000
ma3   -0.4705 0.0789  -5.9616  0.0000
sar1   0.4308 0.2145   2.0086  0.0456
sar2  -0.3738 0.0667  -5.6019  0.0000
sar3  -0.0953 0.1065  -0.8947  0.3718
sma1  -1.2876 0.2098  -6.1362  0.0000
sma2   0.4886 0.1779   2.7469  0.0064
```



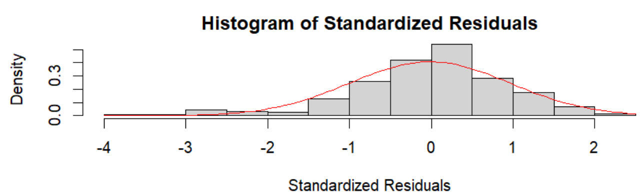
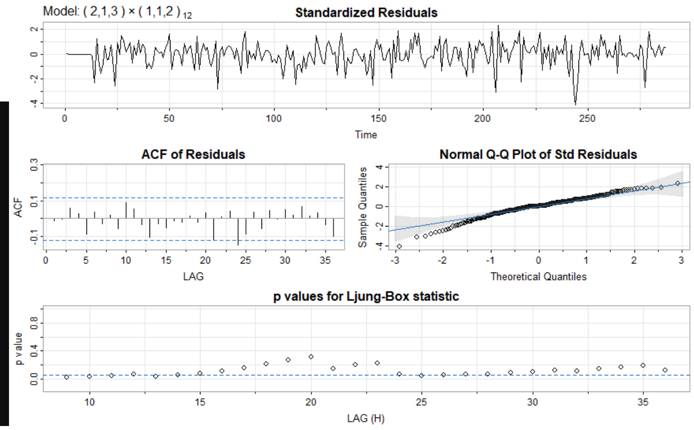


Figure 9: Estimates of model SARIMA (213.312). Figure 10&11: Residual plots for model SARIMA (213.312)

Upon reviewing the ACF and PACF plots from Figure 10 & Figure 11, it's evident that some autocorrelations persist within the residuals. Furthermore, the Ljung-Box hypothesis test indicates that autocorrelations are not zero, failing to reject the null hypothesis. Additionally, there is a violation of the normality assumption observed in the Q-Q plot. In Figure 9, the estimate for sar3 has an insignificant p-value, indicating that this model is not the most suitable choice.

Model 2 <- mod.fit.213.112

$ttable	Estimate	SE	t.value	p.value
ar1	0.4406	0.3139	1.4037	0.1616
ar2	0.3728	0.2616	1.4252	0.1553
ma1	-0.9312	0.3086	-3.0173	0.0028
ma2	-0.3100	0.3988	-0.7772	0.4377
ma3	0.2412	0.1154	2.0897	0.0376
sar1	-0.3684	0.2019	-1.8244	0.0692
sma1	-0.4047	0.1910	-2.1188	0.0350
sma2	-0.4725	0.1642	-2.8778	0.0043



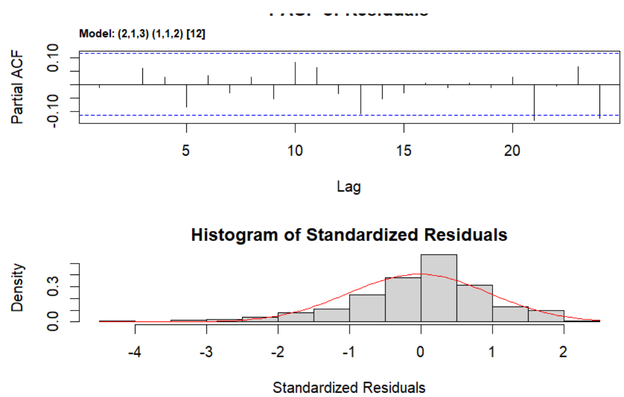
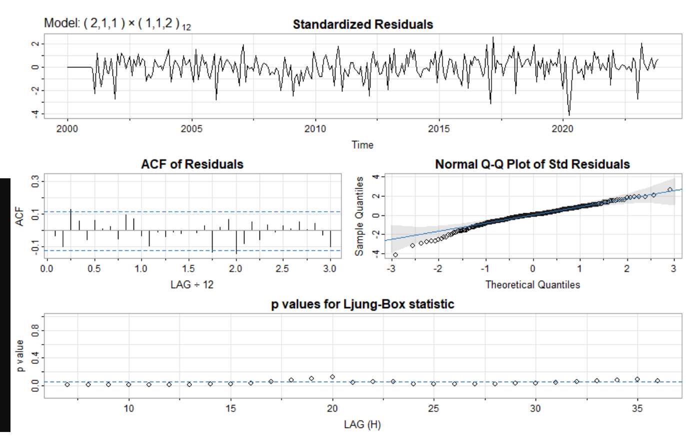


Figure 12: Estimates of model SARIMA (213.112). Figure 13&14: Residual plots for model SARIMA (213.112)

After examining the ACF and PACF plots from Figures 13&14, it's evident that some autocorrelations persist within the residuals. Additionally, the Ljung-Box hypothesis test indicates a violation of the independence of residuals, as it fails to reject the null hypothesis. Furthermore, there is a deviation from the normality assumption observed in the Q-Q plot. In Figure 12, the estimates for ar1, ar2, ma2, and sar1 are found to be insignificant, which suggests that the model is not a good fit.

Model 3 <- mod.fit.211.112

$ttable	Estimate	SE	t.value	p.value
ar1	0.5259	0.0600	8.7608	0.0000
ar2	0.1827	0.0602	3.0338	0.0027
ma1	-1.0000	0.0293	-34.1670	0.0000
sar1	-0.3728	0.1956	-1.9060	0.0577
sma1	-0.4114	0.1880	-2.1885	0.0295
sma2	-0.4948	0.1613	-3.0687	0.0024



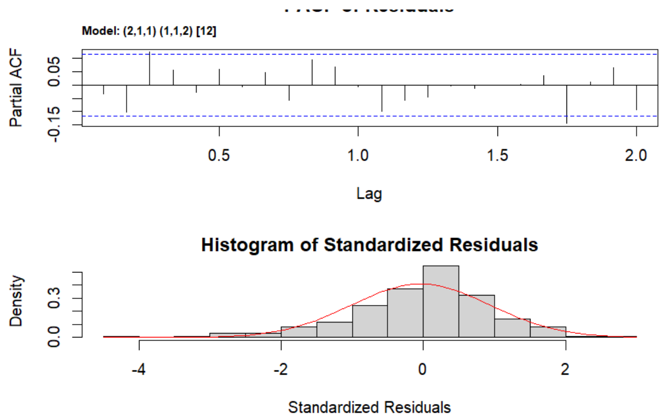


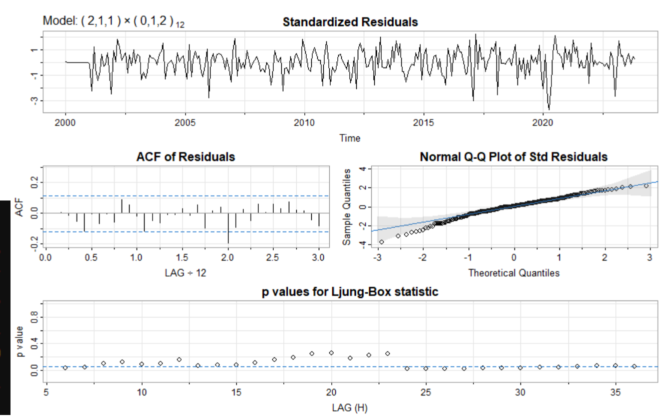
Figure 15: Estimates of model SARIMA (211.112). Figure 16&17: Residual plots for model SARIMA (211.112)

After examining the ACF and PACF plots from Figures 16 & 17, it's evident that some autocorrelations persist within the residuals. Additionally, the Ljung-Box hypothesis test suggests a failure to reject the null hypothesis, indicating a violation of the independence of residuals. Moreover, there is a slight deviation from the normality assumption observed in the Q-Q plot. In Figure 15, the estimates for sar1 are found to be insignificant. It's worth noting that the insignificant estimates for sar1 might have been overlooked if we had conducted more thorough diagnostics on the residuals plot.

Model 4 <- mod.fit.211.012

$ttable

	Estimate	SE	t.value	p.value
ar1	-0.1681	0.3611	-0.4655	0.6419
ar2	-0.1890	0.1543	-1.2251	0.2216
ma1	-0.2648	0.3779	-0.7007	0.4841
sma1	-0.7878	0.0802	-9.8234	0.0000
sma2	-0.1268	0.0818	-1.5511	0.1221



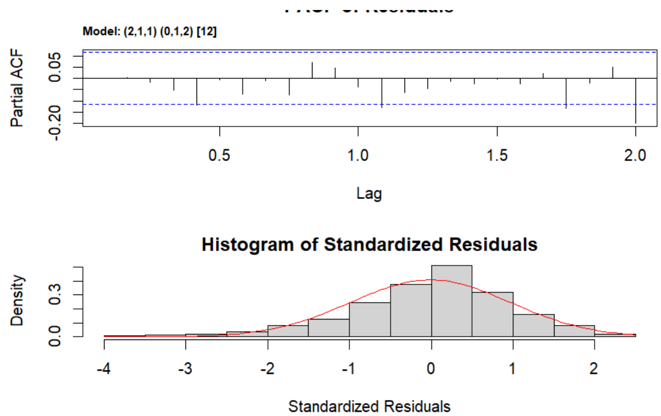
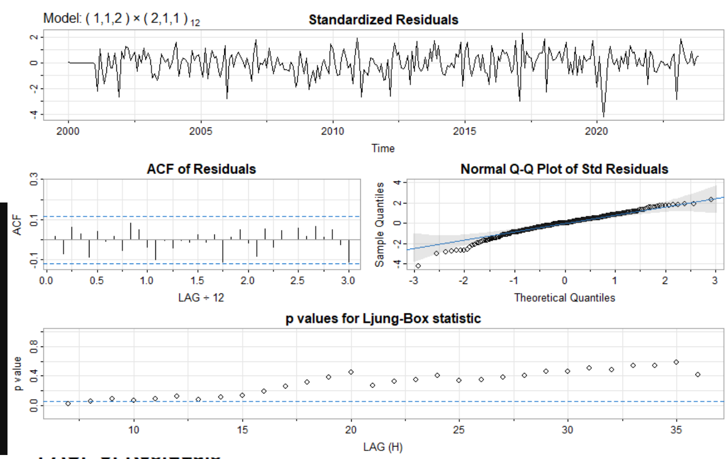


Figure 18: Estimates of model SARIMA (211.012). Figure 19&20: Residual plots for model SARIMA (211.012)

Upon inspecting the ACF and PACF plots from Figures 19 & 20, it's evident that some autocorrelations persist within the residuals. Additionally, the Ljung-Box hypothesis test indicates a failure to reject the null hypothesis, suggesting that the independence of residuals is violated. Furthermore, deviations from the normality assumption are observed in the Q-Q plot. In Figure 18, all parameter estimates except for sma1 are deemed insignificant, indicating that the model is a poor fit for the data.

Model 5 <- mod.fit.112.211

```
$ttable
     Estimate     SE  t.value p.value
ar1    0.8688 0.0466  18.6597  0.0000
ma1   -1.3968 0.0843 -16.5758  0.0000
ma2    0.3968 0.0822   4.8262  0.0000
sar1   0.0329 0.0673   0.4892  0.6251
sar2  -0.2780 0.0640  -4.3399  0.0000
sma1  -0.8388 0.0466 -17.9963  0.0000
```



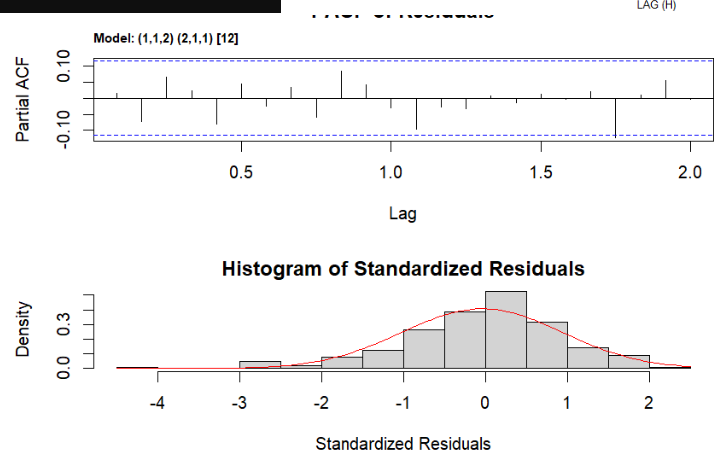


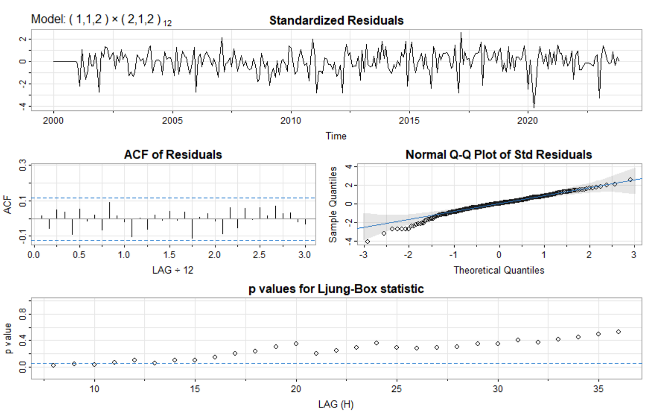
Figure 21: Estimates of model SARIMA (112.211). Figure 22&23: Residual plots for model SARIMA (112.211)

Upon examining the ACF and PACF plots in Figures 22 & 23, it's clear that the residuals are free from autocorrelation, indicating a good model fit to the temporal structure of the data. However, the Ljung-Box hypothesis test points to instances of insignificant independence among the residuals. Additionally, slight deviations from the normality assumption are observed in the Q-Q plot. In Figure 21, except for sar1, all parameter estimates are significant, suggesting that this model is a good fit.

Model 6 <- mod.fit.112.212

$ttable

	Estimate	SE	t.value	p.value
ar1	0.9090	0.0400	22.7267	0e+00
ma1	-1.4484	0.0761	-19.0226	0e+00
ma2	0.4484	0.0750	5.9779	0e+00
sar1	0.5433	0.1335	4.0703	1e-04
sar2	-0.3601	0.0679	-5.3028	0e+00
sma1	-1.4004	0.1383	-10.1228	0e+00
sma2	0.5576	0.1280	4.3560	0e+00



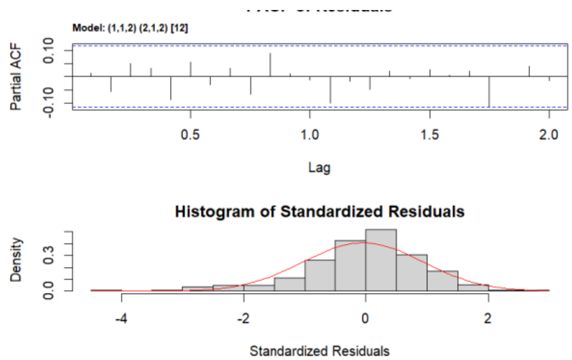


Figure 24: Estimates of model SARIMA (112.212). Figure 25&26: Residual plots for model SARIMA (112.212)

Upon reviewing the ACF and PACF plots depicted in Figures 25 & 26, we can confirm the absence of autocorrelations in the residuals, indicating that the model has effectively captured the temporal dependencies within the data. However, the Ljung-Box hypothesis test reveals some instances of insignificant independence in the residuals, while the Q-Q plot slightly deviates from the normality assumption. Despite these minor issues, all parameter estimates are significant, suggesting that this model fits well.

Next, we will evaluate the AIC, AICc, and BIC scores of our top-performing models. The model with the lowest score in any of these criteria will be selected for our forecasting.

mod.name <chr>	AIC <dbl>	AICc <dbl>	BIC <dbl>
Sarima.213.312	-0.4001888	-0.3971359	-0.2551362
Sarima.213.112	-0.3511743	-0.3491912	-0.2324950
Sarima.211.112	-0.3411512	-0.3400030	-0.2488450
Sarima.211.012	-0.3257864	-0.3249694	-0.2466669
Sarima.112.211	-0.4048825	-0.4037343	-0.3125764
Sarima.112.212	-0.4291804	-0.4276437	-0.3236876

6 rows

Figure 27: AIC, AICc and BIC values for all models

Forecasting:

After selecting the best fit model of SARIMA(112)(212), next step is to capture short-to-medium trends. Using forecast horizons of 3 months (periods), 6 months, and 14 months predicted points are plotted on the plots:

- 3 periods ahead: All data datapoints and an abbreviated plot version included.

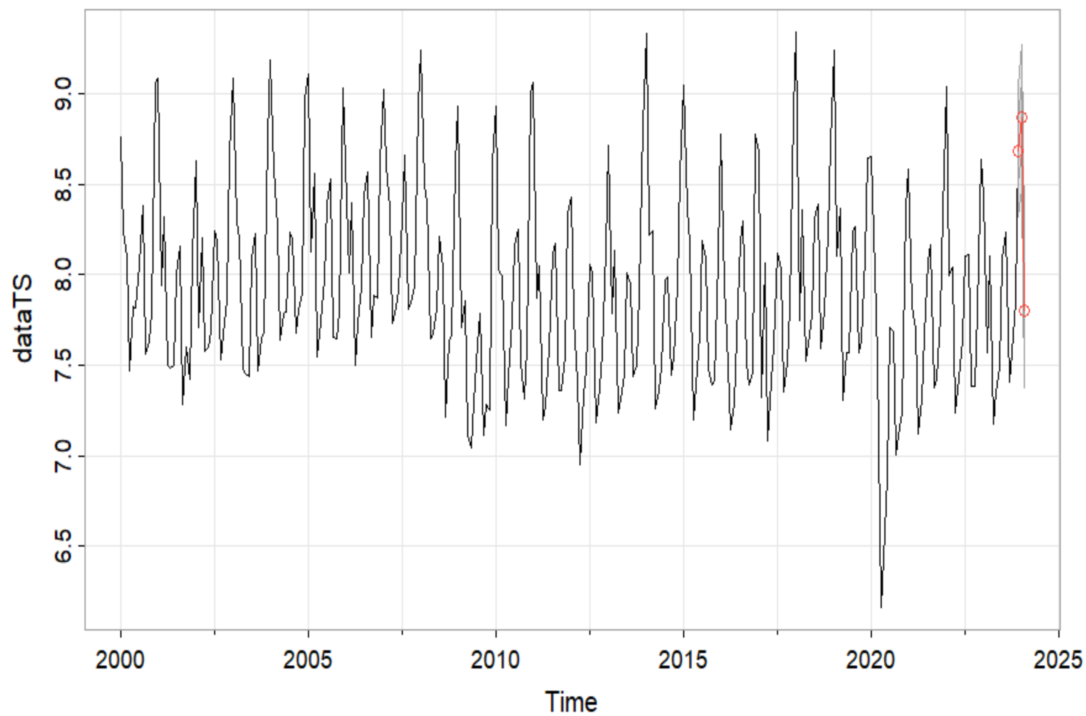


Figure 28: all data points forecast for 3 periods ahead

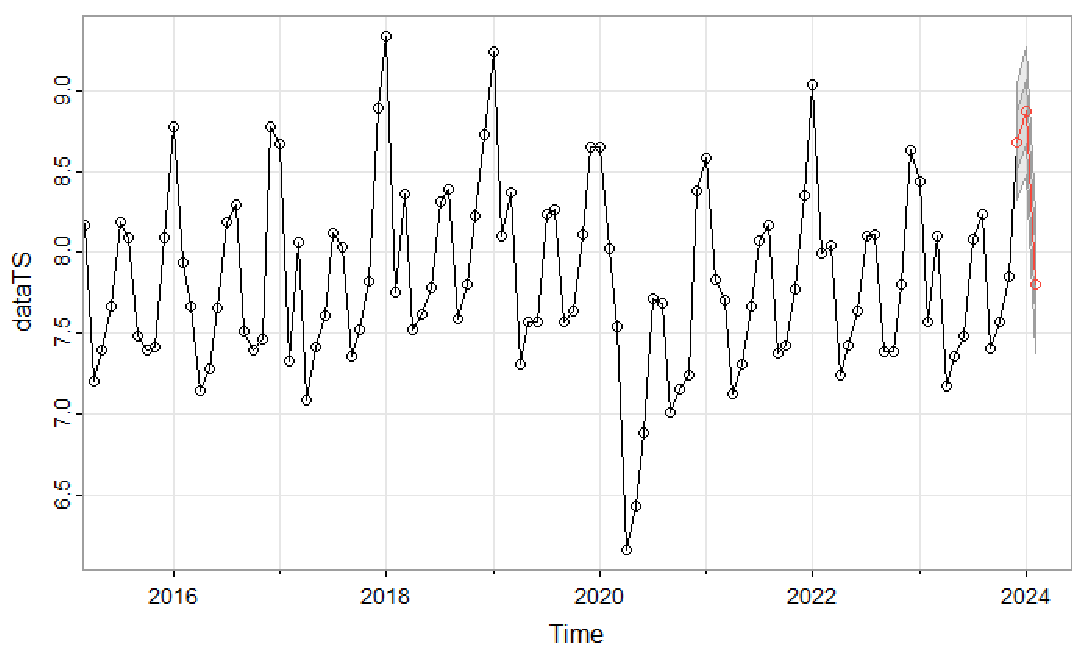


Figure 29: limited view of forecast for 3 periods ahead

- 6 periods ahead: All data datapoints and an abbreviated plot version included.

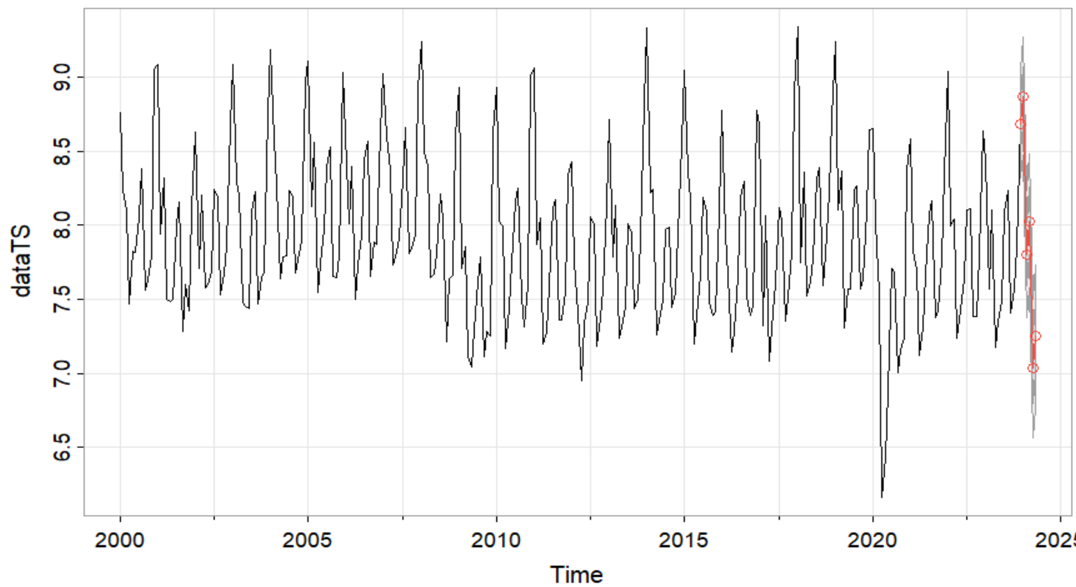


Figure 30: all data points forecast for 6 periods ahead

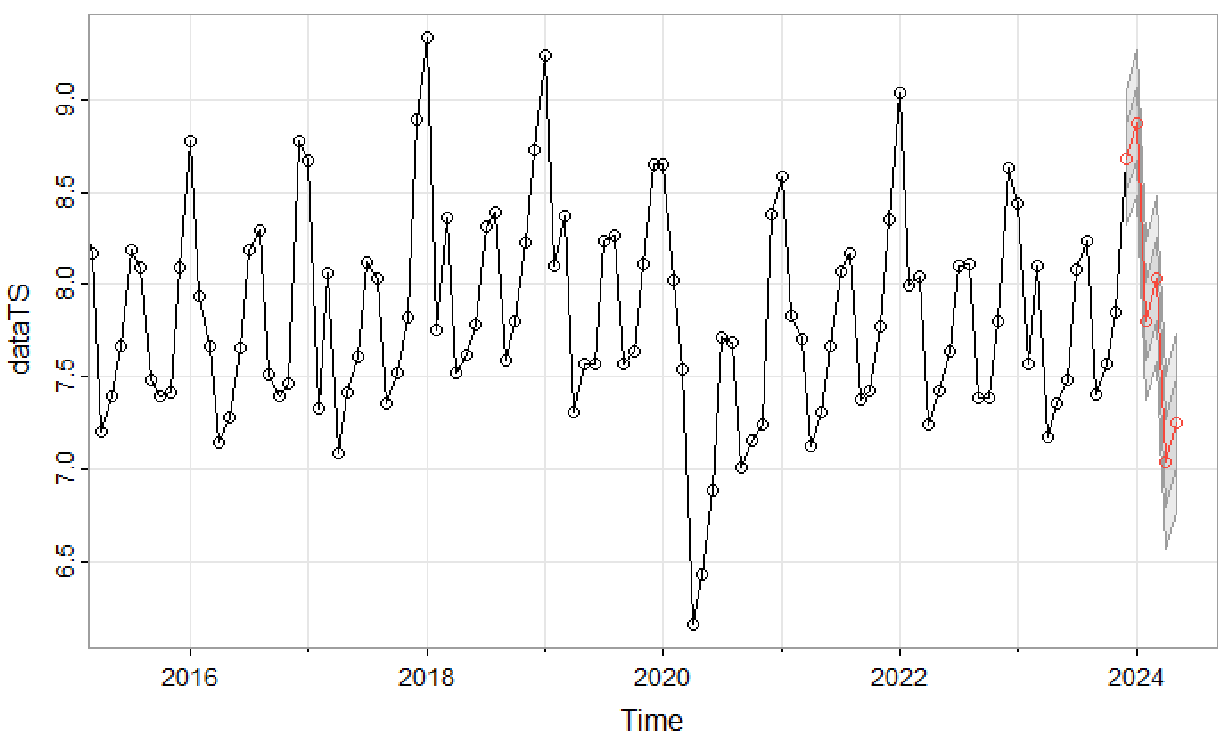


Figure 31: limited view of forecast for 6 periods ahead

- 14 Month ahead: All data datapoints and an abbreviated plot version included.

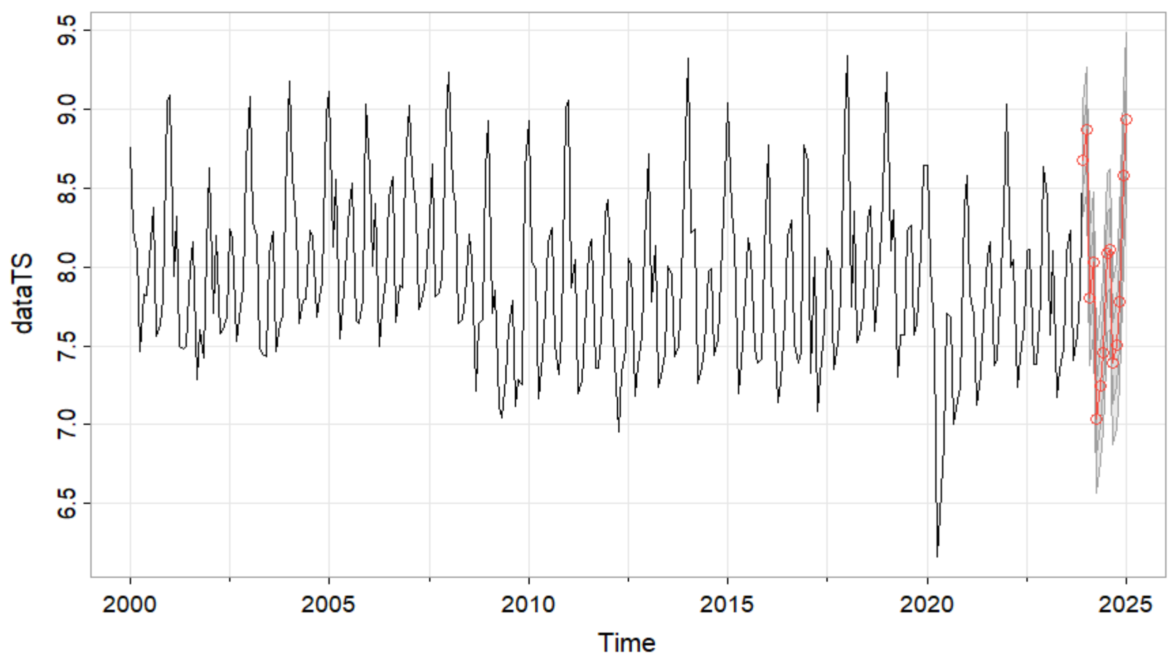


Figure 32: all data points forecast for 14 periods ahead

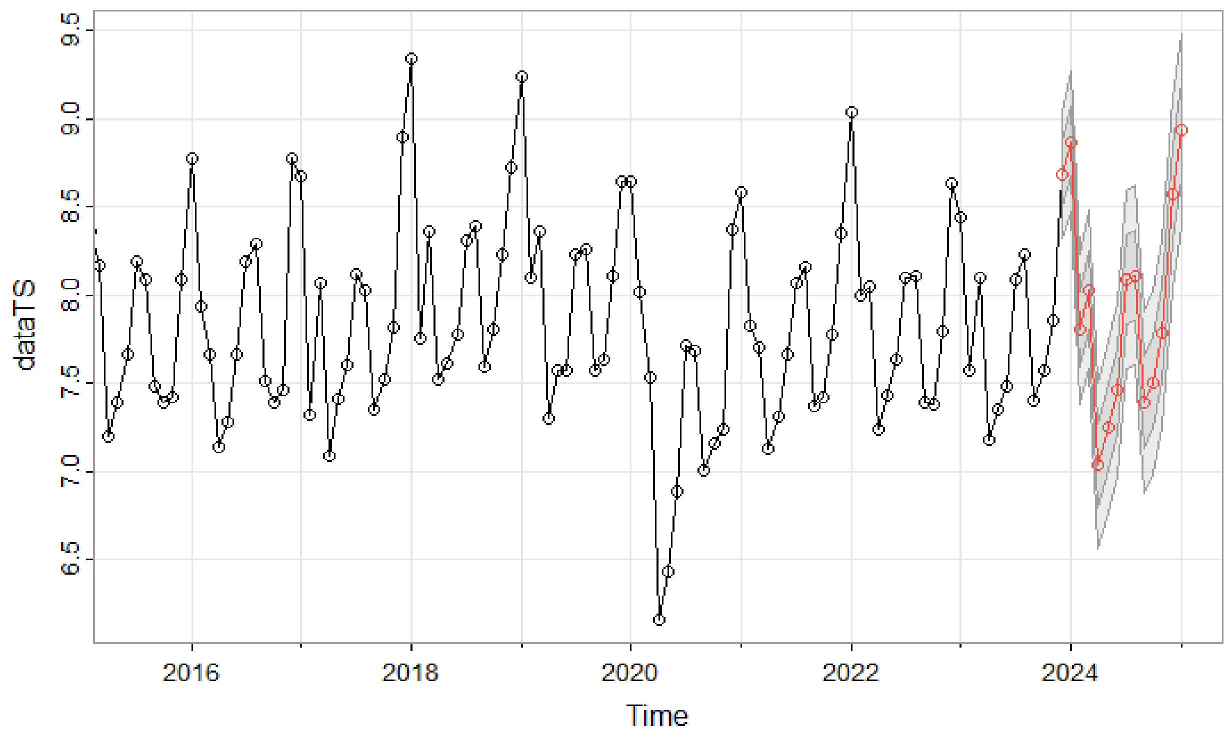


Figure 33: limited view of forecast for 14 periods ahead

Furthermore, a custom confint function was created to calculate the upper and lower bounds of confidence intervals for each forecast horizon. The function calculates the 2.5th and the 97.5th percentile values for each predicted data point.

Confidence interval calculations for 3,6, and 14 months ahead respectively:

```
# confidence interval calculation:
```{r}
confint <- function(x) {
  quantile_975 <- qnorm(0.975)
  quantile_25 <- qnorm(.025)

  calculated_values_lower <- numeric(length(x$pred))
  calculated_values_upper <- numeric(length(x$pred))

  for (i in 1:length(x$pred)) {
    calculated_values_upper[i] <- x$pred[i] + quantile_975 * x$se[i]
    calculated_values_lower[i] <- x$pred[i] + quantile_25 * x$se[i]

  }

  list(lower = calculated_values_lower, upper = calculated_values_upper)
}

```

```{r}
confint_results <- confint(forecast)
confint_results$lower
confint_results$upper
```

 [1] 8.328004 8.480649 7.385377
 [1] 9.036784 9.262214 8.222884

```{r}
confint_results2 <- confint(forecast2)
confint_results2$lower
confint_results2$upper
```

 [1] 8.328004 8.480649 7.385377 7.588641 6.574803 6.775909
 [1] 9.036784 9.262214 8.222884 8.470184 7.491564 7.721156

```{r}
confint_results3 <- confint(forecast3)
confint_results3$lower
confint_results3$upper
```

  [1] 8.328004 8.480649 7.385377 7.588641 6.574803 6.775909 6.975908 7.592416 7.611610 6.881695 6.993511
 [12] 7.263352 8.047197 8.403406
  [1] 9.036784 9.262214 8.222884 8.470184 7.491564 7.721156 7.944395 8.579990 8.614948 7.898114 8.020828
 [12] 8.299782 9.108339 9.477047
```

Figure 34: confidence interval calculations

We visually compare the original total energy consumption data against predicted values which are calculated by subtracting residual values from the chosen SARIMA model:

```
pred_x <- dataTS - resid(mod.fit.112.211$fit, frequency=12)
pred_x
```

```
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov
2000 8.755270 8.234755 8.100231 7.465028 7.827778 7.819995 8.054620 8.381258 7.561928 7.633311 7.796158
2001 9.114624 8.348169 8.119165 7.549443 7.779017 7.617501 7.872365 8.233271 7.361870 7.464301 7.611548
2002 8.462092 7.645883 7.964237 7.374304 7.574125 7.604319 8.085925 8.307441 7.417765 7.680364 7.662097
2003 8.939871 8.252280 8.418109 7.642560 7.703349 7.600997 7.962864 8.171853 7.463173 7.591033 7.769881
2004 8.966989 8.285387 8.457619 7.626973 7.723868 7.748363 8.242472 8.401943 7.502847 7.774801 7.808789
2005 9.193441 8.332403 8.347250 7.769485 7.780539 7.802640 8.327850 8.433466 7.769761 7.850725 7.827161
2006 9.121495 7.965657 8.283840 7.566372 7.640610 7.776358 8.348218 8.533654 7.724461 7.811993 7.924590
2007 8.925439 8.273051 8.521500 7.738069 7.842633 7.865940 8.390542 8.445511 7.809591 7.990773 7.951572
2008 9.280712 8.493633 8.506553 7.745245 7.790067 7.797292 8.266411 8.390796 7.485063 7.477296 7.698102
2009 8.857943 8.091434 8.097088 7.220499 7.328080 7.334695 7.844778 7.797292 7.068232 7.357269 7.411754
2010 8.724772 8.026436 8.168884 7.337179 7.370669 7.550623 8.105033 8.324819 7.527859 7.614134 7.534198
2011 9.050618 8.353585 8.147582 7.347671 7.463027 7.522623 8.058286 8.190643 7.399680 7.498014 7.571307
2012 8.670532 7.714304 7.885571 6.926171 7.046987 7.372826 7.898842 8.084648 7.265519 7.426629 7.458118
2013 8.548735 7.908496 7.823916 7.298436 7.432155 7.505341 7.929261 8.055419 7.235974 7.483993 7.581299
2014 9.083622 8.254869 8.398898 7.445401 7.401692 7.561074 7.976205 8.063500 7.317159 7.459989 7.649180
2015 8.895423 8.078355 8.172762 7.396694 7.432158 7.581503 8.073850 8.200807 7.357290 7.521907 7.580563
2016 8.528932 7.880729 7.936116 7.046152 7.263823 7.478385 8.064510 8.147982 7.465540 7.536851 7.551624
2017 8.953174 7.870924 7.642654 7.163506 7.254564 7.528312 8.048413 8.152375 7.346574 7.432408 7.674320
2018 9.000933 8.174487 8.137576 7.341514 7.586004 7.725913 8.213525 8.222483 7.598220 7.679330 7.926669
2019 9.229905 8.271099 8.145221 7.481941 7.476811 7.716761 8.132315 8.237132 7.511326 7.583180 7.784674
2020 9.000228 7.929221 8.025718 6.935802 6.793064 6.864924 7.479764 7.667156 7.034911 7.099360 7.362073
2021 8.595424 7.658244 7.794616 6.876222 7.172029 7.514910 8.119708 8.091130 7.398614 7.470634 7.588071
2022 8.873507 7.888740 8.161615 7.371748 7.466892 7.617924 8.093382 8.148800 7.400625 7.462417 7.732478
2023 8.964141 7.672814 7.760208 6.990116 7.276632 7.501257 8.064116 8.071134 7.441085 7.489093 7.766144
          Dec
2000 9.056578
2001 8.608933
2002 8.622317
2003 8.739368
2004 8.852377
2005 8.780335
2006 8.897142
2007 8.850968
2008 8.761538
2009 8.348889
2010 8.653035
2011 8.593359
2012 8.454810
2013 8.778272
2014 8.879294
2015 8.406910
2016 8.464045
2017 8.759562
2018 8.881215
2019 8.760426
```

Figure 35: predicted values calculation

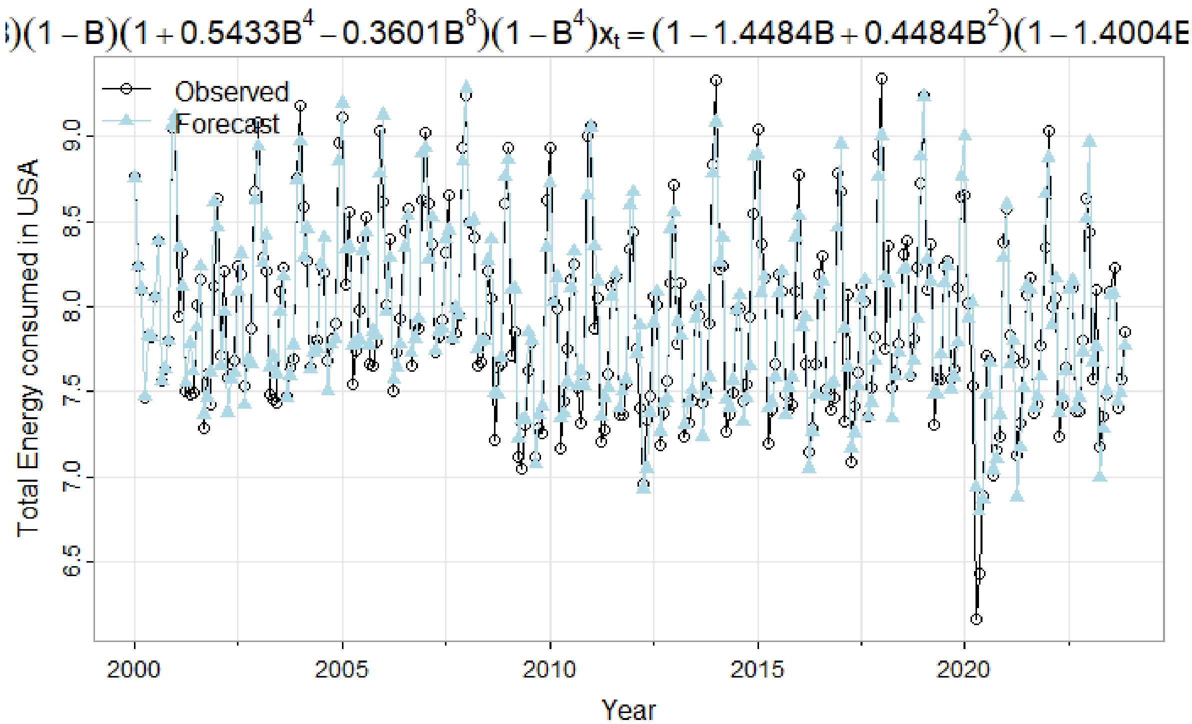


Figure 36: visual representation of the predicted values compared to original data

Finally, given the coefficients of the SARIMA(112)(212) (see figure 24) we can find the mathematical representation of the model:

- **Final model:** (1+.909*B)*(1-B)*(1+.5433*B^4-.3601*B^8)*(1-B^4)*x[t] = (1-1.4484*B +.4484*B^2)*(1-1.4004*B^4 + 0.5576*B^8) * w[t])

Discussion:

In this project, we used the SARIMA model to analyze and forecast the United States' energy usage from January 2000 to December 2023. The EIA dataset included monthly consumption that included different seasonal changes.

Finding seasonality and trends in the original data was the first step in our investigation. The ACF and PACF plots showed distinct seasonal trends, which required data processing to become stationary. We were able to resolve these problems with double differencing, which produced a more comprehensible and stable dataset for model building.

Choosing the right SARIMA model orders was a part of the model specification process. Optimal values for the autoregressive and moving average parameters, together with their seasonal equivalents, were identified by analyzing the autocorrelation and partial autocorrelation plots. The challenges obtaining an ideal model fit were found through repeated fitting and diagnostics. Some models showed deviations from normality, violated the independence criterion, or had significant autocorrelation in the residuals. Most of the fitting process was without issue, we were mostly able to follow the hints that coefficient statistics and residual analysis provided us to locate the best model. In the end, we did indeed compromise slightly as our Ljung-Box statistic's first autocorrelation p-value was hovering just under .05 threshold. Yet the SARIMA (112)x(212) model we chose was certainly the best out of the ones we tested, having all significant parameter estimates and the lowest AIC, AICc, and BIC values.

However, our analysis is not without limitations. Many variables go into energy consumption and some of them are quantitatively unpredictable such as politically related issues. Geography and politics may present an issue to generalizability of this model. Energy consumption data of countries in the Northern hemisphere which have similar climates and an energy management policy which aligns with the US' may use this model, but it is unclear how effective it would be. Population size, availability of natural resources, logistical challenges, and availability of energy related technologies may cause the model to be ineffective even in the seemingly best candidates.

In conclusion, our SARIMA model offers important insights into US energy consumption trends, it is essential to acknowledge its limitations and uncertainty in predicting/forecasting real world data. More research and refinement of model techniques is needed to improve the accuracy and generalizability of the model. In future we may add variables which directly or indirectly affect energy consumption in the US resulting in a more robust and accurate model to make decisions in energy management and policy formation. Yet still, the model presented in this paper offers valuable insights to those who seek to leverage the knowledge of the trend to their benefit.

Bibliography:

U.S. Energy Information Administration - EIA - Independent Statistics and Analysis

Class Notes

Appendix:

library(astsa)

library(readxl)

source("examine.mod.R")

data <- read_excel("TTLEnergyConsumption.xlsx")

dataTS <- ts(data$`Total Primary Energy Consumption`, frequency=12, start=c(2000, 1))

tsplot(dataTS, main="Total Primary Energy Consumption", ylab="Energy Consumption", xlab="Time")

acf2(as.numeric(dataTS), max.lag=72, main="Estimated ACF & PACF of energy consumption")

#We can see that data is not stationary. Data does not have a trend, but is seasonal. We can see the seasonal pattern every 12 months in PACF & ACF plot.

#Lets address the seasonality pattern by applying seasonal difference: (1-B^12)*xt

dataTS_diff <- diff(dataTS, lag=12, differences=1)

tsplot(dataTS_diff, main="Total Primary Energy Consumption One : (1-B^12)*xt", ylab="Energy Consumption", xlab="Time")

acf2(as.numeric(dataTS_diff), max.lag=72, main="Estimated ACF & PACF of energy consumption One diff")

```

#Data is now seasonality stationary, but it doesn't seem to be stationary i.e. has constant mean, constant variance and more vividly there is no constant autocorrelation structure as can be seen in the ACF plot. Lets attempt to adjust data to achieve constant variance by first applying a log transformation.

dataTS_log <- log(dataTS)

dataTS_diff_log <- diff(dataTS_log,lag=12, differences=1)
```

```
tsplot(dataTS_diff_log, main="Total Primary Energy Consumption, log transformed: (1-B^12)*xt", ylab="Energy Consumption", xlab="Time")

acf2(as.numeric(dataTS_diff_log), max.lag=72, main="Estimated ACF & PACF of energy consumption diff_log")


#(1-B)(1-B^12)*xt


dataTS_diff_diff <- diff(diff(dataTS,lag=12, differences=1))

tsplot(dataTS_diff_diff, main="Total Primary Energy Consumption - (1-B)(1-B^12)*xt", ylab="Energy Consumption", xlab="Time")

acf2(as.numeric(dataTS_diff_diff), max.lag=72, main="Estimated ACF & PACF of energy consumption diff_diff")



#Fitting and Diagnostic


mod.fit.213.512 <- sarima(as.numeric(dataTS), p=2, d=1, q=3, P=5, D=1, Q=2, S=12)

mod.fit.213.512


mod.fit.213.412 <- sarima(as.numeric(dataTS), p=2, d=1, q=3, P=4, D=1, Q=2, S=12)

mod.fit.213.412


mod.fit.213.312 <- sarima(as.numeric(dataTS), p=2, d=1, q=3, P=3, D=1, Q=2, S=12)

mod.fit.213.312


examine.mod(mod.fit.213.312, 2,1,3,3,1,2,12)


mod.fit.213.112 <- sarima(as.numeric(dataTS), p=2, d=1, q=3, P=1, D=1, Q=2, S=12)

mod.fit.213.112
```

```
examine.mod(mod.fit.213.112, 2, 1, 3, 1, 1, 2 ,12)

mod.fit.211.112 <- sarima(dataTS, p=2, d=1, q=1, P=1, D=1, Q=2, S=12)
mod.fit.211.112

examine.mod(mod.fit.211.112, p=2, d=1, q=1, P=1, D=1, Q=2, S=12)

mod.fit.211.012 <- sarima(dataTS, p=2, d=1, q=1, P=0, D=1, Q=2, S=12)
mod.fit.211.012

examine.mod(mod.fit.102.012, p=2, d=1, q=1, P=0, D=1, Q=2, S=12)

mod.fit.112.211 <- sarima(dataTS, p=1, d=1, q=2 , P=2, D=1, Q=1, S=12)
mod.fit.112.211

examine.mod(mod.fit.112.211, p=1, d=1, q=2 , P=2, D=1, Q=1, S=12)

mod.fit.112.212 <- sarima(dataTS, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12)
mod.fit.112.212

#THE ONE!!!

examine.mod(mod.fit.112.212, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12)

#FORCASING

forecast <- sarima.for(dataTS, n.ahead = 3, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12, plot.all = TRUE)
```

```
forecast2 <- sarima.for(dataTS, n.ahead = 6, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12, plot.all = TRUE)

forecast3 <- sarima.for(dataTS, n.ahead = 14, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12, plot.all = TRUE)

forecast <- sarima.for(dataTS, n.ahead = 3, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12, plot.all = FALSE)

forecast2 <- sarima.for(dataTS, n.ahead = 6, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12, plot.all = FALSE)

forecast3 <- sarima.for(dataTS, n.ahead = 14, p=1, d=1, q=2 , P=2, D=1, Q=2, S=12, plot.all = FALSE)

print(forecast)
print(forecast2)
print(forecast3)


# confidence interval calculation:

confint <- function(x) {
  quantile_975 <- qnorm(0.975)
  quantile_25 <- qnorm(.025)

  calculated_values_lower <- numeric(length(x$pred))
  calculated_values_upper <- numeric(length(x$pred))

  for (i in 1:length(x$pred)) {
    calculated_values_upper[i] <- x$pred[i] + quantile_975 * x$se[i]
    calculated_values_lower[i] <- x$pred[i] + quantile_25 * x$se[i]
```

```
  }

  list(lower = calculated_values_lower, upper = calculated_values_upper)
}

confint_results <- confint(forecast)
confint_results$lower
confint_results$upper

confint_results2 <- confint(forecast2)
confint_results2$lower
confint_results2$upper

confint_results3 <- confint(forecast3)
confint_results3$lower
confint_results3$upper



pred_x <- dataTS - resid(mod.fit.112.211$fit, frequency=12)
pred_x



tsplot(dataTS, ylab = 'Total Energy consumed in USA', xlab= "Year", type = "o",

      main = expression(paste((1+.909*B)*(1-B)*(1+.5433*B^4-.3601*B^8)*(1-B^4)*x[t] == (1-1.4484*B
+.4484*B^2)*(1-1.4004*B^4 + 0.5576*B^8) * w[t])))

lines(pred_x, col = "lightblue", type = "o", pch = 17)
```

```
legend("topleft", legend = c("Observed", "Forecast"),

       lty = c("solid", "solid"), col = c("black","lightblue"), pch = c(1, 17), bty = "n")


print(expression(paste((1+.909*B)*(1-B)*(1+.5433*B^4-.3601*B^8)*(1-B^4)*x[t] == (1-1.4484*B
+.4484*B^2)*(1-1.4004*B^4 + 0.5576*B^8) * w[t]) ))
```