

Energy Consumption Analysis (Eastern USA)

Rayyan Kazim

01/04/2024

Contents

Section 1: Introduction	2
Section 2: Modelling	2
Section 3: Results	8
Section 4: Conclusion	9
References	10

Section 1: Introduction

Data Link: <https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption>

Our data set keeps track of the hourly energy consumption (in megawatts) in the Eastern United States Region from October 2004 to August 2018 (We will convert our data to time series format where October 1, 2004 will be day 1). Out of all the observations in each day, we would like to find the sum of them. This would give us a rough estimate of the total energy consumption per day. We would like to predict the total energy consumption for October 1, 2024 to get an idea of what the energy consumption will be looking like for North Americans 20 years after the first recorded entry of this data set. It will also give us an idea of how much more/less energy we use now (from now, October 2024 is in 6 months) than before.

Importance: In general, energy consumption analysis is very important as it analyses the recordings/data of how much natural resources we use. If we effectively manage our energy consumption levels, we can build a more sustainable future. This report will give us an idea of how much energy we will be consuming in October 2024. This will tell us how little or how big of a change is needed.

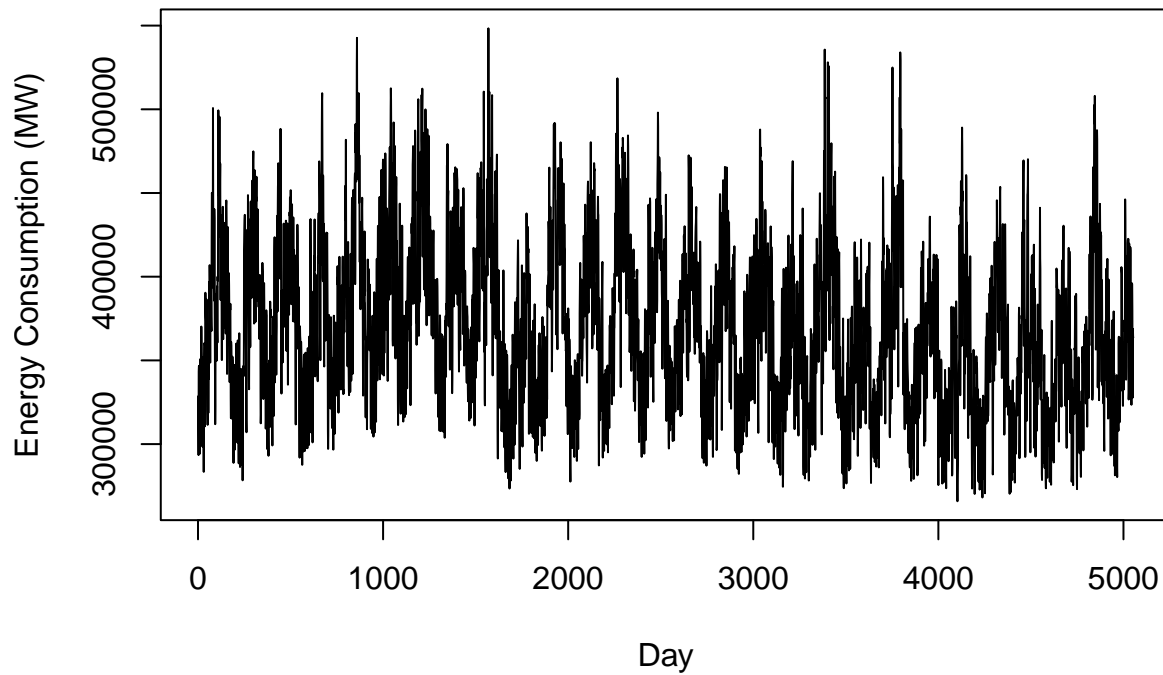
Section 2: Modelling

```
library(forecast)
library(fpp2)
library(tseries)
library(astsa)
library(knitr)
library(tidyverse)
library(magrittr)
library(ggplot2)
```

We will first clean our data set so we are left with 2 columns (Date, Total Energy Consumption (MW)). The date section will incorporate only the year, month and day (starting October 1, 2004) and the other column will represent total hourly energy consumption in megawatts. We will also convert our data to time series format where the “Date” column will be changed to “Day” (Day 1, Day 2, Day 3...).

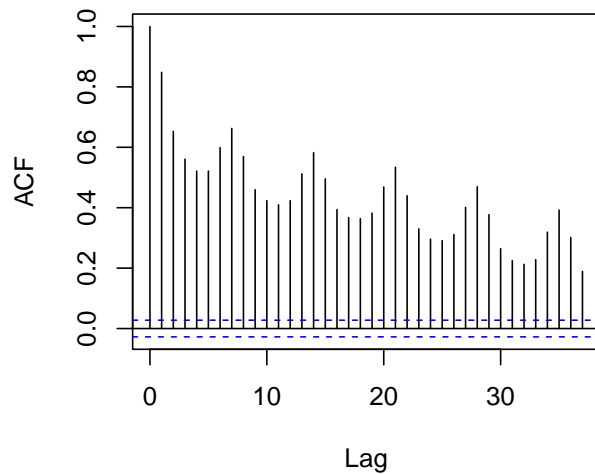
Next, we will plot to see what our time series plot looks like.

Eastern United States Energy Consumption (MW)

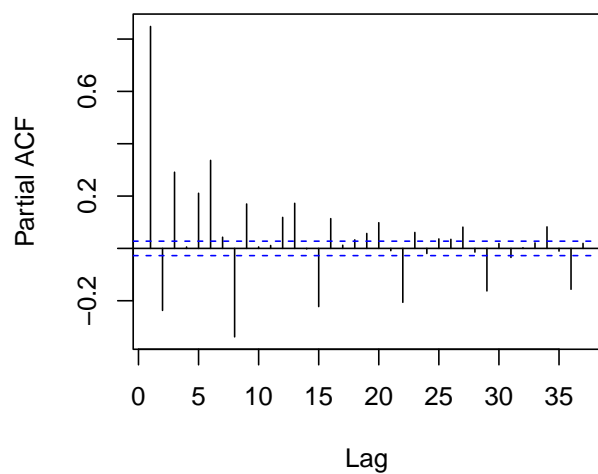


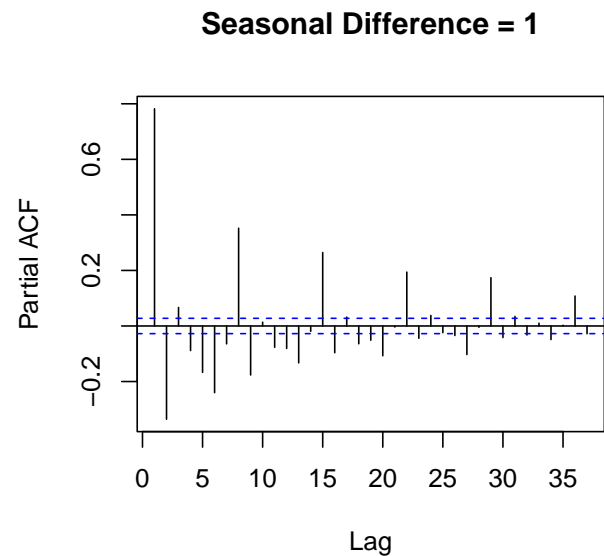
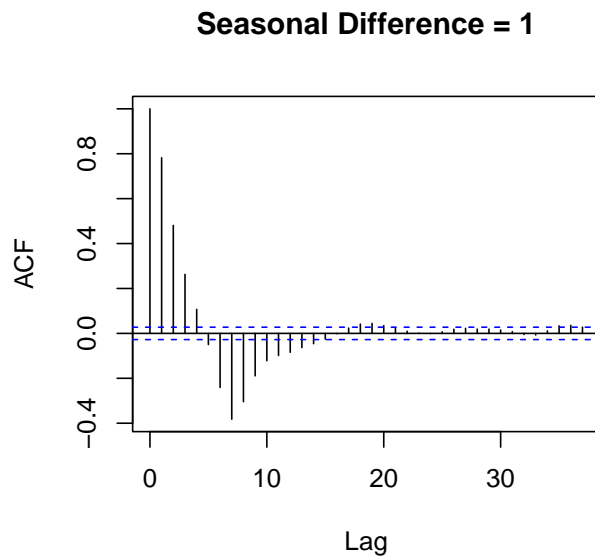
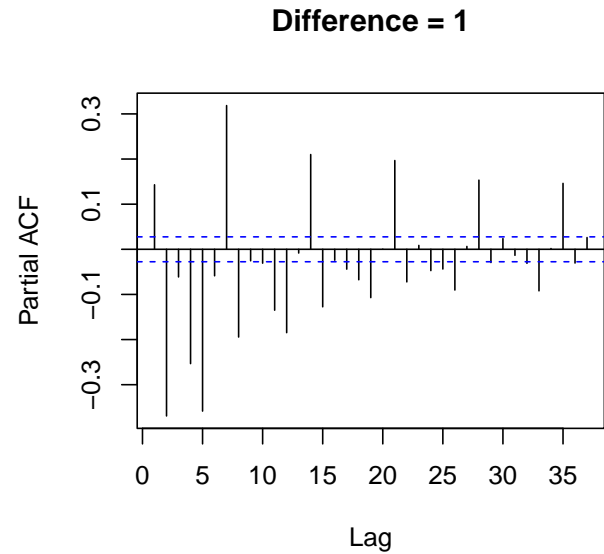
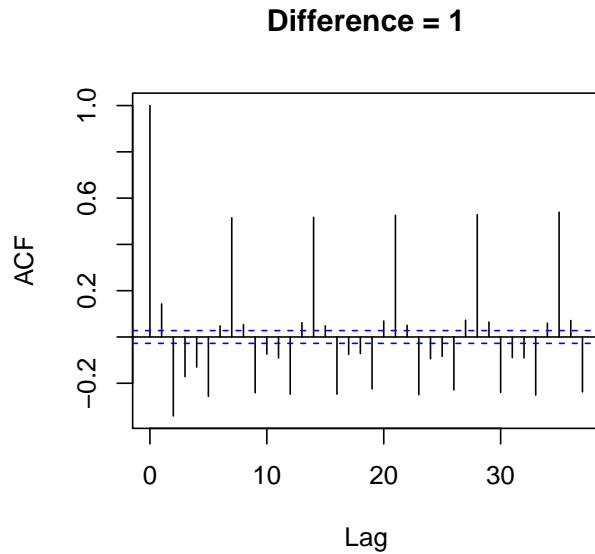
Now, we are to plot are acf's and pacf's with and without seasonality to specify our model and get an idea of what our model will look like.

ACF Plot



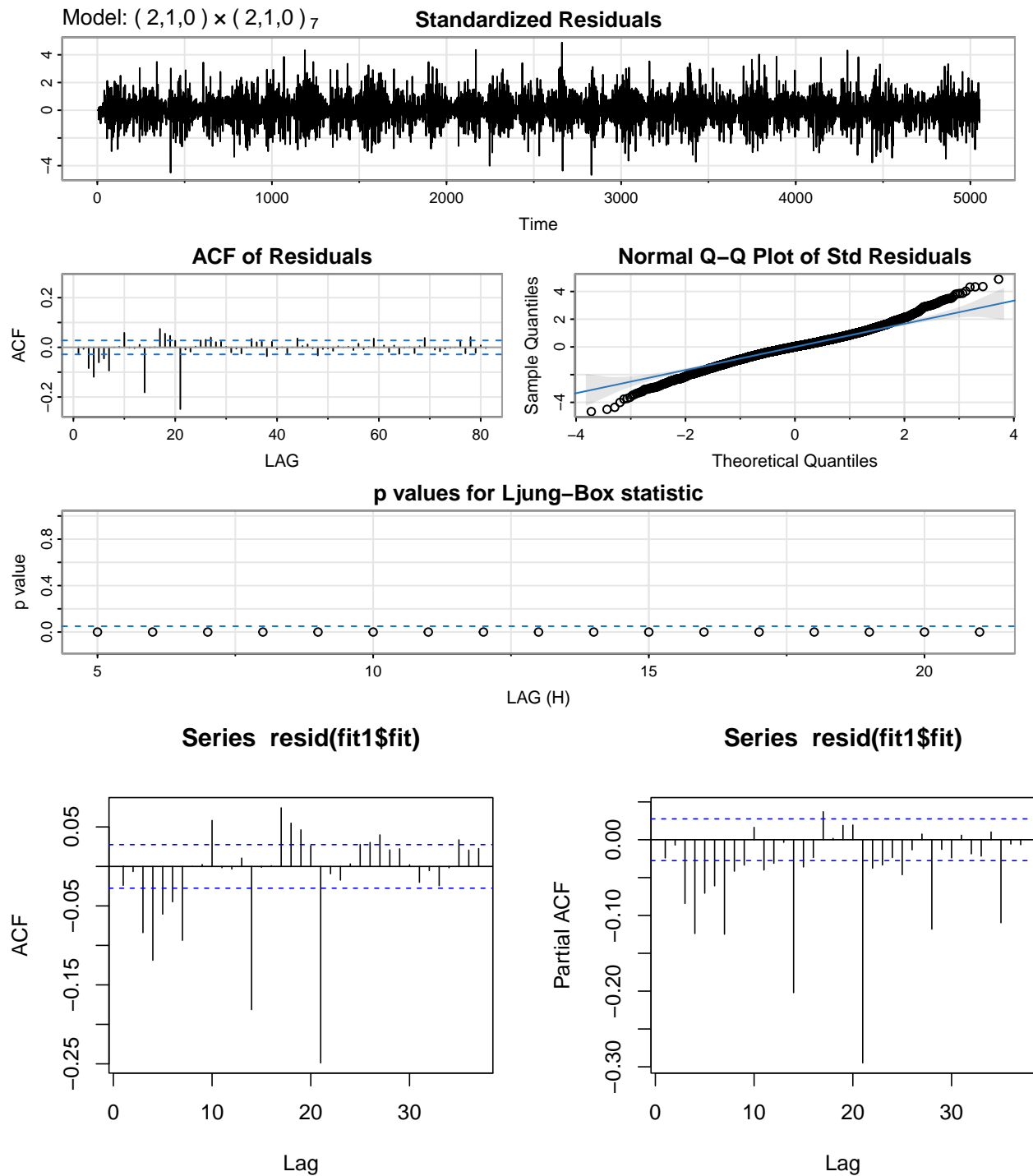
Partial ACF Plot



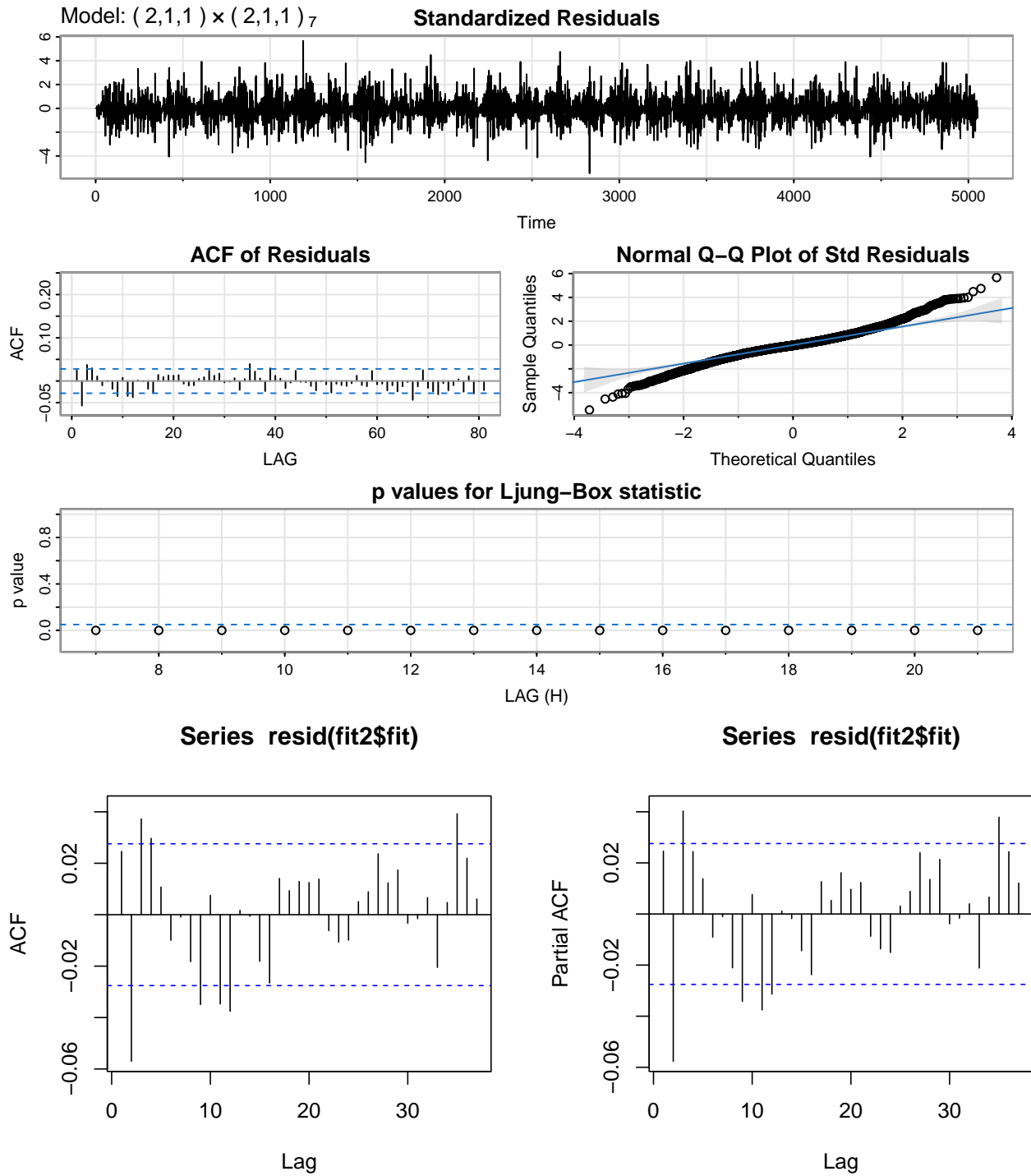


These plots appear to be odd as our points keep moving farther away from zero with more differences added. With the difference being 1, our ACF model shows that the points do not tend towards zero. With our seasonal difference being 1, our ACF model takes a large number of lags to tend towards zero. This tells us that our model may lack MA components. However, for both seasonal and nonseasonal differences being 1, our PACF plots show that our points tend towards zero for lag > 2 . This tells us about our AR components.

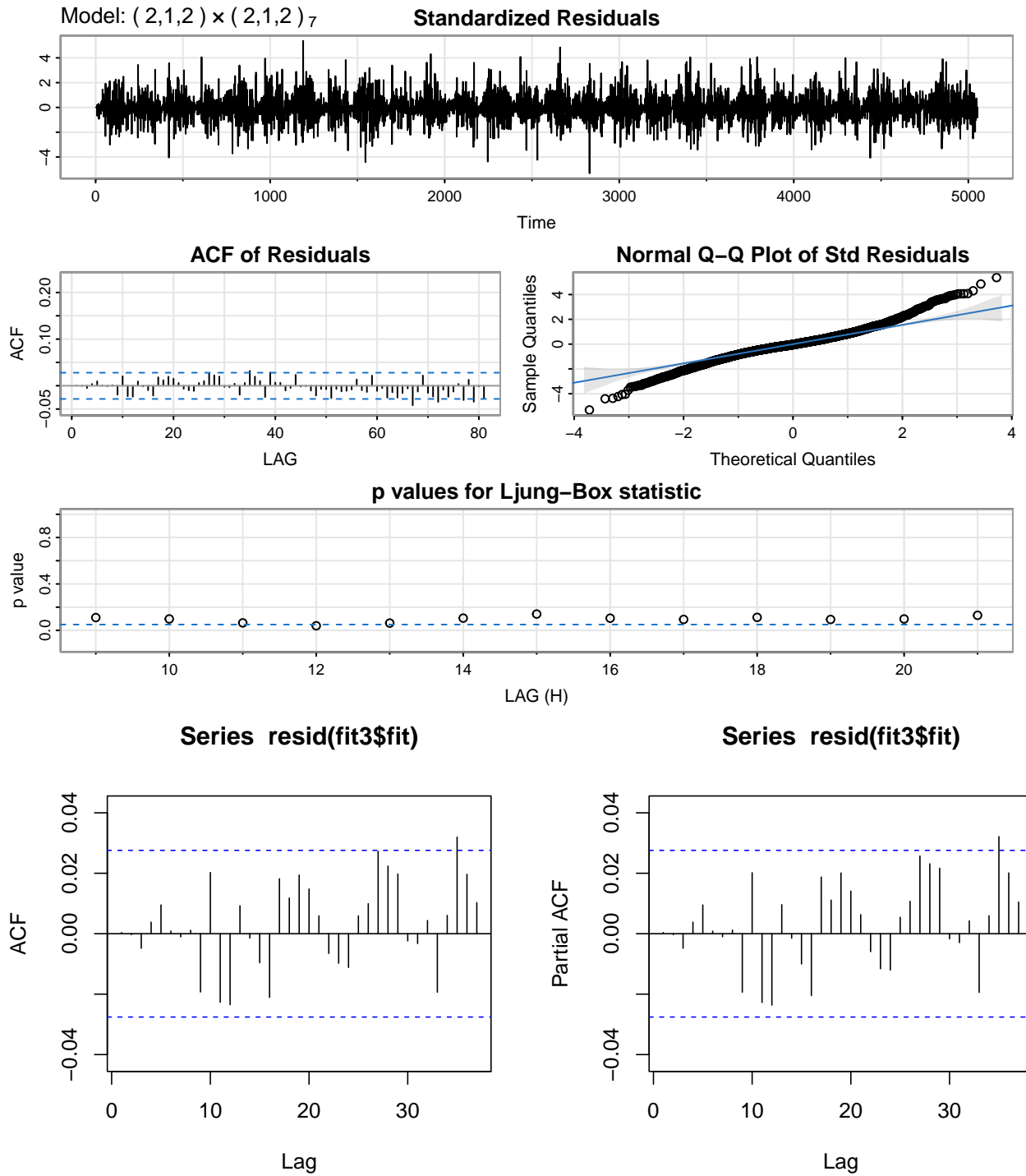
Hence, these plots imply that the parameters for our SARIMA model can be $p=2$, $d=1$, $q=0$, $P=2$, $D=1$, $Q=0$.



Looking at the ACF and PACF of Residuals plots, this might not be the best model. We can try to change our seasonal and non-seasonal, q and Q , values for our MA components, to visualize whether or not they would make a difference. As mentioned prior to the diagnostics, we will not change our difference terms as they are appearing to have odd effects.



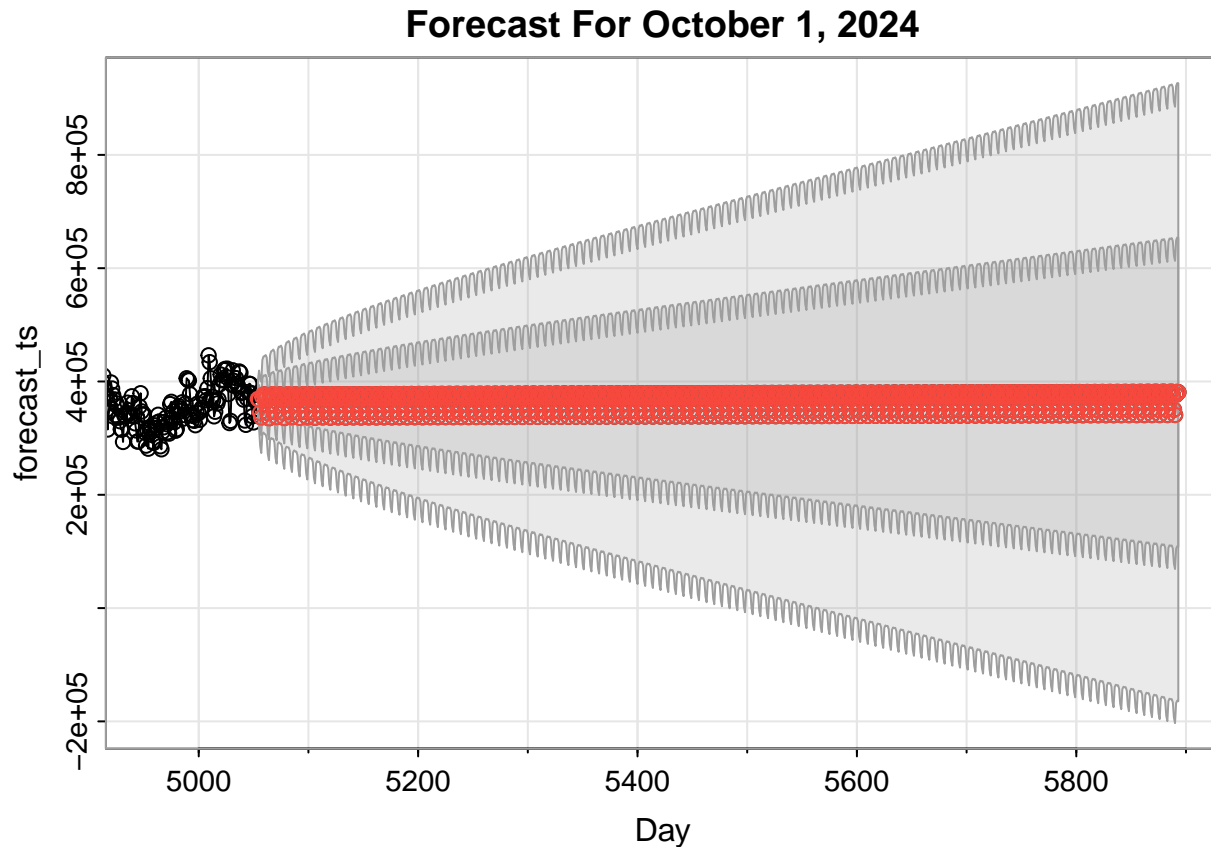
Based on the residual plots for ACF and PACF, this SARIMA model is definitely a much better fit (although, it can still be improved) than the previous one. We will continue to increase our values for our MA components



This is our best model as we can see how well it fits for ACF and PACF plots for the residuals. Our Normal Q-Q plot also indicates the points falling on the diagonal.

Next, we will forecast our model to derive an insight for October 2024. We will obtain the total energy consumption for October 1, 2024.

Based on model diagnostics, we will use SARIMA(2,1,2)x(2,1,2)₇.



Section 3: Results

Our forecast tells us that our total energy consumption on October 1, 2024 is expected to be slightly less than, but nearly 400,000 MW.

In terms of effectiveness, our forecast definitely does give an estimate of what we can expect for the future. Looking very closely at the red line in the forecast plot, the points very slowly increase over time (They get closer and closer to 400,000 MW for every few days). This plot helps us reach our goal for predicting the response value for a specific day (Oct 1, 2024) in the future, which we were successfully able to do. However, we must consider the fact that this is an estimate and not an accurate value. If our energy consumption patterns are consistent with how they were in the past, then this forecast can be close to accurate, and we can use this model to forecast even further than October 1, 2024.

Overall, our forecast is effective (through the forecast plot) in reaching our goals for predicting energy consumption levels on October 1, 2024. It is able to utilize our model to create a trend to successfully find values for the future. In the introduction, we also mentioned that we would like to see how much energy we consume now and in the future compared to past. Our forecast helps us answer this question. As mentioned before, the points on the red line (slowly increasing every few days) in the forecast plot indicated that we are using more energy now, and will be using more in the future, than we did in the past.

Section 4: Conclusion

A major limitation is that this data set only includes values up till August 2018, which is definitely a very long time before now. Hence, this data set might be more biased towards energy consumption patterns in the past as opposed to the present (unless these patterns have remained the same or similar). In future, we can either use a more modern data set, or we can continue to use this data set, combining it with more closely up-to-date values.

Another limitation is that this data consists of only the Eastern region of the United States. This could lead to potential bias as it does not represent the entire United States or Canada. We used Eastern states as sampling units for the North American population (as mentioned in the introduction). In future, to derive even more accurate conclusions for the energy consumption rates/patterns in North America, it can be beneficial to sample more states as well as provinces in Canada.

References

- Aggregate: Compute Summary Statistics of Data Subsets*. N.D. RDocumentation. <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate>.
- Cryer, J. D., and K. S. Chan. 2008. *Time Series Analysis: With Applications in r*. Springer Texts in Statistics. Springer New York. <https://books.google.ca/books?id=bHke2k-QYP4C>.
- Handbook-Team. July 18, 2023. *The Epidemiologist r Handbook*. Bookdown R Package. <https://books.google.ca/books?id=bHke2k-QYP4C>.
- Mulla, Rob. 2018. *Hourly Energy Consumption*. Kaggle. <https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption>.