PHINEAS PHAM

DEC 15th, 2022

MATH 422 - Dr. White

Bike Sharing Analysis

I. Introduction:

Climate change has recently become one of the most trending topics. Thus, cities are reimagining their transportation infrastructure. Shared mobility concepts, such as car sharing, bike sharing, or scooter sharing have been utilized and publicly favored in urban cities like Washington D.C. The bike sharing system was first launched in the District of Columbia in August 2008, attracting 1600 users during its first two years of operation. In this project, I will try to forecast the future demand for bike sharing daily to help address balance the supply with demand and potentially reduce operating costs for bike sharing suppliers. The goal of the project is to figure out the patterns of bike usage and predictor variables for daily bike sharing customers.

After trying time series analysis with multiple methodologies, I find out that the number of users are highly correlated with the number of users the day before, and useful predictor variables are maximum temperature, precipitation level, wind speed, and holiday. Furthermore, using periods of one year and 11 months are useful in capturing the wavy trend of the total customers.

II. Methods:

a) Data source:

I use the bike sharing dataset from Kaggle made available by Julia Nikulski. The author published the data set on Kaggle, mentioning that the data was collected from Capital Bikeshare for the bike sharing demand, NOAA's National Climatic Data Center for weather data, and the DC Department of Human Resources for data on public holidays. We have the data from 2011 to 2018, which is a subset of the actual one. As the data set is cleaned, I will import the dataset into R like normal. Then, I will use the ‘total_cust’ and ‘date’ columns, which have information on the total number of customers that use bike sharing daily, and the date of the record. With those columns, I will transform them into time series, identify the trends from the plot, and find the best model. The author of the dataset has collected data from Capital Bikeshare for the bike

sharing demand, NOAA's National Climatic Data Center for weather data, and DC Department of Human Resources for data on public holidays. Thus, this can potentially produce mistakes or bias to the dataset during the data gathering process. And I acknowledge that my dataset is a subset from the total bikesharing usage in the D.C. area.

The time series of daily use of bike sharing in the D.C. area I am modeling was generated by some true model that I don't know, but any individual observation is affected by random factors (weather, economic reasons, etc.) so I am observing signal + noise, and the noise is random. Thus, I think my data is a random sample from a true generating process that I am trying to model.

b) Data wrangling:

The data set has many NA values in variables that I will not use in this project. However, the data on total number of users have 4 consecutive missing values at indices 1849, 1850, 1851, and 1852. Figure 1 shows us the missing values and its neighbors trend.

I fill those missing values by filling the middle numbers from indices 1848 and 1853, forming a straight line from those values. The end results would include values filling the gap from figure 1, and figure 2 shows how the gap in the time series filled with new values.

c) Overall trend:

I assume that the total number of customers is not constant and is increasing over time. The increase should be in a cyclical trend, and the cycles are bigger overtime. I reason that because the demand for bikesharing will increase overtime as it is becoming more popular to the citizens and travelers, while the usage will always fall for seasons that not many use bikes like winter or bad weather.

In this project, I will try several approaches to fit the ‘total_cust’ time series: the non-parametric trend, the function of time model, the SARIMA model, spectral analysis, and regression with weather factors such as temperature, precipitation, and wind speed. Figure 3 shows the overall ‘total_cust’ time series after we handle the missing values.

d) Approaches:

First of all, I perform a log transform to the time series. Doing analysis on log transformed time series will become beneficial for a time series with increasing volatility like this one. Figure 4 shows the log transformed version of ‘total_cust’ time series.

I perform a non-parametric trend with Ksmooth. The best smoothing series signifies the wavy trend shown in figure 5. Following that, I perform decomposition of the log version of the time series (figure 6) to draw out the overall trend and seasonality in the time series. A clear seasonality matches with what we see in the smoothed version, and the overall trend is not a linear but rather bended to a curve, this suggests I may use a time squared as a variable for the linear model (the models then imply that time squared is not significant).

My best ARIMA on the log transformed series is one with first order differencing, AR(1), and MA(1) and no seasonality. As a result, I use this model for forecasting.

In order to perform xreg with the log transformed series, using log on independent variables is necessary. Thus, wrangling temp_max, precip, and wind variables is a must to do a log transformation on these values. I transform temp_max from Celsius to Fahrenheit scale, and increase precip values by 1 so that there are no negative values for log transformation. These changes will make no impact on the analysis as we are analyzing the relationship between variables. However, my xreg model does not provide useful information. It fails to pass several tests like Ljung-Box and normal Q-Q plot. Thus, I move on to conduct spectral analysis.

Figure 7 and 8 shows the results of our periodogram. Frequencies of 8 and 9 (periods of around one year and 11 months) are the most important frequencies in our time series. Thus, I will include these periods to fit cosine frequencies in our analysis by using linear regression. Fitting a log of time and 2 cosine trends above, I get a pretty good model and useful findings. When fitting the model to the actual data (figure 9), we can see it captured the most important features in the time series. Looking at the figure 10, the adjusted R-squared of 60% argued that the model is helpful, about 60% of the observations can be explained by this model, which would perform well for forecasting and other uses. Only the p-value for Xsin2 is insignificant, but I decide to leave it in the model as the model performance shows no significant change if I removed it.

III. Results and Conclusions:

After fitting various models, ARIMA(1,1,1) is the best model. Figure 11 shows the residual plots of this ARIMA model. It looks stationary with 2 concerning outliers, but it passes 4 stationary tests. The ACF and PACF tests only have 3 bad lags at 11, 27 and 29 among the first 35 lags. These bad lags seem random, as it is unreasonable to see that the number of total

customers has a relationship with the number of 11, 27 and 29 days ago. Thus, I call it a pass for ACF and PACF tests. Next, I check the model utility tests. From figure 12, we can see that the Q-Q plot of the residuals does not look normal, and the p-values for the Ljung-Box statistic show signs of concern. Figure 13 shows the density plot of the residuals, and they look good with no significant skew. Thus, we are likely to trust the point forecast instead of forecast intervals. I build a forecasting model in figure 14. The forecast looks reasonable as it captures my opinion on the trend of increasing the number of customers in the near future. The auto.arima() function suggests an ARIMA(3,1,2) model. Although this model has a bit better AICc score, some of the terms, as shown in figure 15, have p-values higher than 0.05, implying that those terms are not significant. Thus, I decide to keep the ARIMA(1,1,1) model.

The regression model with weather factors does not pass the model utility tests. As shown in figure 16, it fails the Ljung-Box tests and normal Q-Q plot. Thus, we will try another model.

Our spectrum model on log transformed 'total_cust' discussed above shows a very good performance. Figure 17 shows a good relationship between residuals and fitted. The residuals from figure 18 looks like there is still a trend in the wavy trend in the middle, but as it passes all 4 stationary tests, we will go with this model as it performs well, shown in figure 9.

Last but not least, I gather the findings from the previous model to attempt to perform a regression with spectrums and other variables including temp_max, precip, wind, and holiday. Based on figure 19, the model can capture 72.3% if the observations, and p-values show evidence that all the terms in this model are significant. This model has mostly the same residuals and 'residuals vs fitted' graphs with the spectrum model, which makes it a not so great model. Based on the coefficients we have, we can conclude that there is a positive correlation between maximum temperature and number of customers, and there are negative correlations between precipitation, wind speed, holiday and total number of customers. These imply that customers enjoy bike sharing more when it is hotter, and less if there is more rain and windy and during holidays. Figure 20 shows how well this model can capture versus the real data. But this also leads to a potential problem of overfitting.

IV. Conclusion:

By applying various techniques, the project was interesting as each approach opens a new set of findings that can potentially add to other models. With the cyclical nature of the dataset, spectrum analysis has become the cornerstone of our findings. Variables like temperature, wind speed, and holiday are helpful in attempts for a better fit model. I learned that we can always find better models by trying more approaches. Based on the model I presented and figure 18, the number of customers are following a cyclical pattern of around 1 year and 11 months, and impacted factors of temperature, precipitation level, wind speed, holiday. For future work, I will try to see if there are more interesting factors affecting the demand for bike sharing in the area.

V. References:

1. Nikulski, J. (2020, June 9). Bike Sharing Washington DC. Kaggle. Retrieved December 15, 2022, from https://www.kaggle.com/datasets/juliajemals/bike-sharing-washington-dc?select=bike_sharing_dataset.csv
2. Motivate International, I. (n.d.). System data. Capital Bikeshare. Retrieved December 15, 2022, from https://ride.capitalbikeshare.com/system-data
3. National Centers for Environmental Information (NCEI). (n.d.). Climate Data Online Search. Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Retrieved December 2022, from https://www.ncdc.noaa.gov/cdo-web/search
4. Holiday schedules. DCHR. (n.d.). Retrieved December 15, 2022, from https://dchr.dc.gov/node/1630311
5. Motivate International, I. (n.d.). About company & history. Capital Bikeshare. Retrieved December 15, 2022, from https://ride.capitalbikeshare.com/about#:~:text=In%20August%202008%2C%20the%20District,its%20two%20years%20of%20 operation.

VI. Appendix:

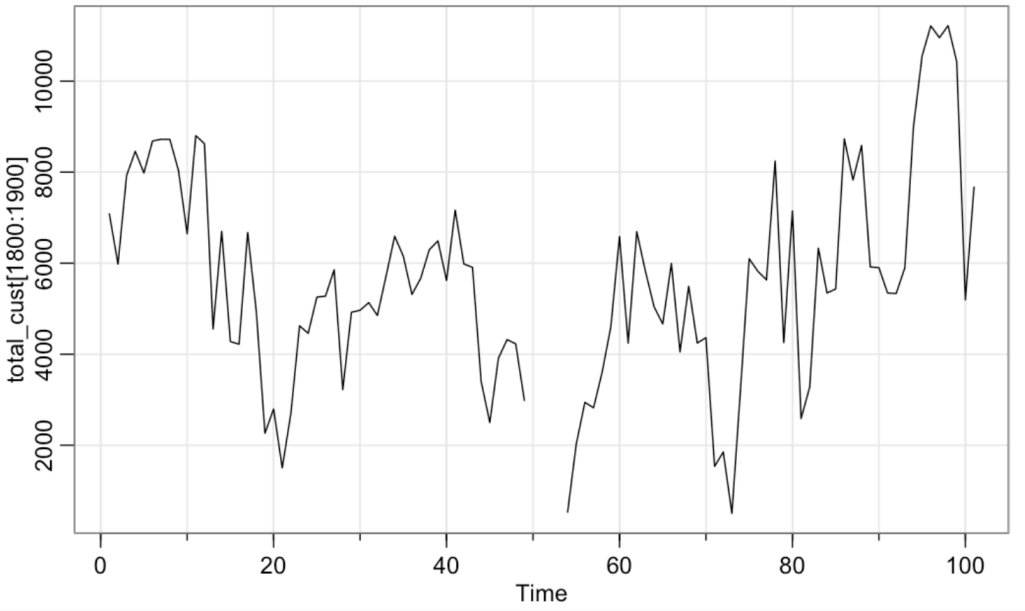

Figure 1: A subset of ‘total_cust’ where missing values are

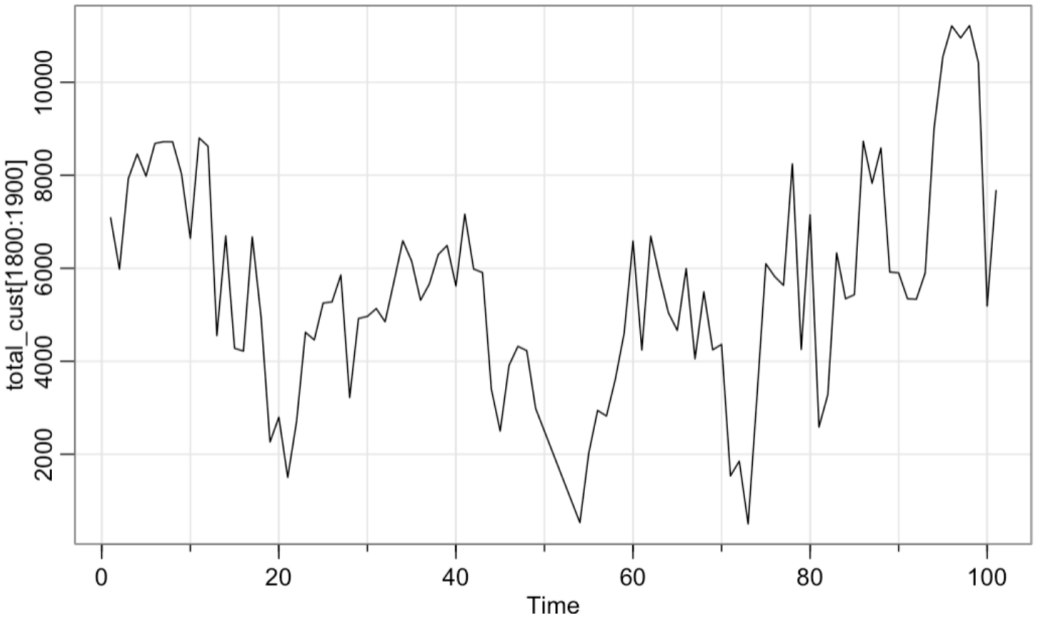

Figure 2: Time series after handling missing values

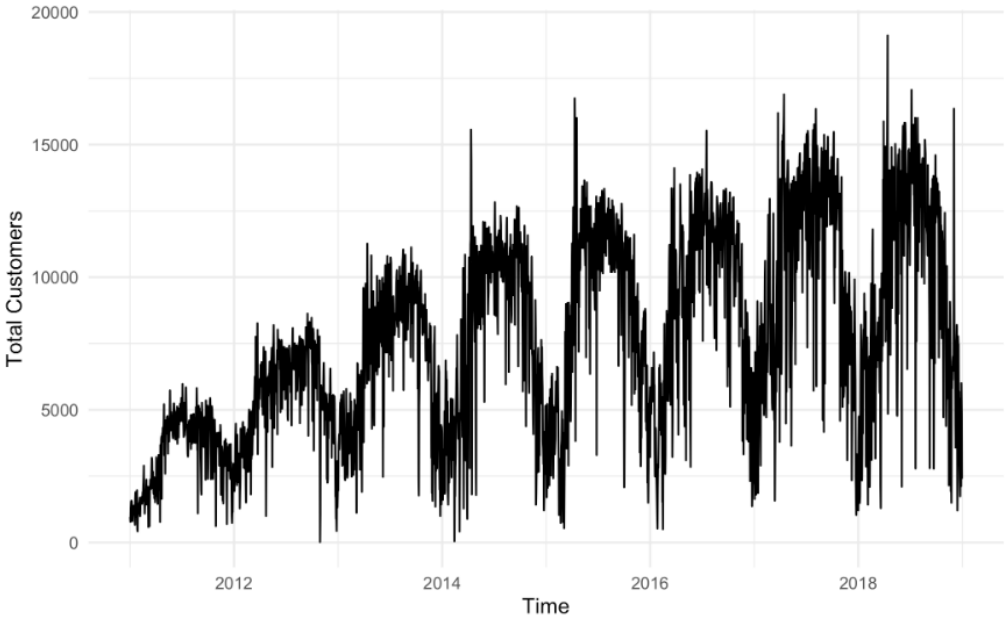

Figure 3: ‘total_cust’ time series

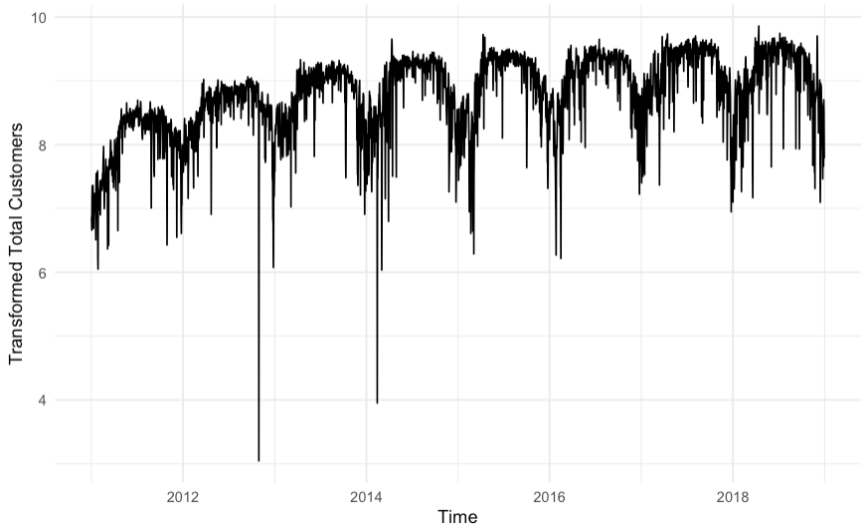

Figure 4: Log transformation of ‘total_cust’ time series

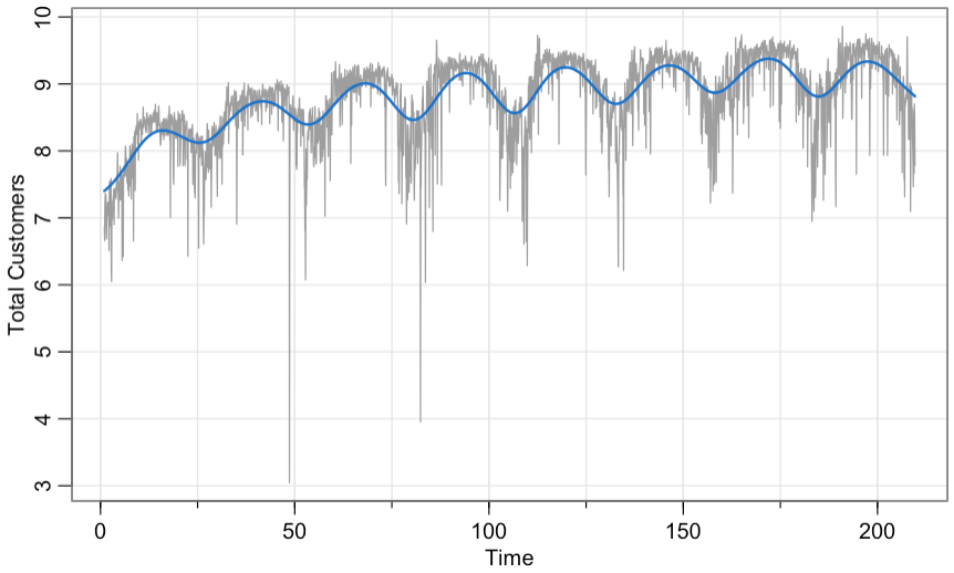

Figure 5: Ksmooth on log time series

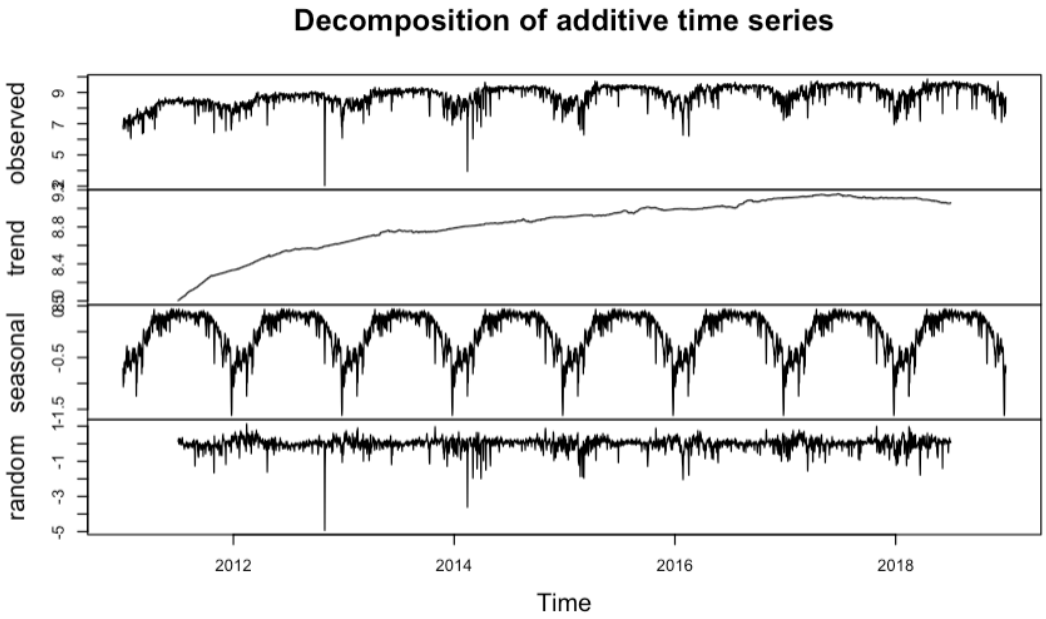

Figure 6: Decomposition of log time series

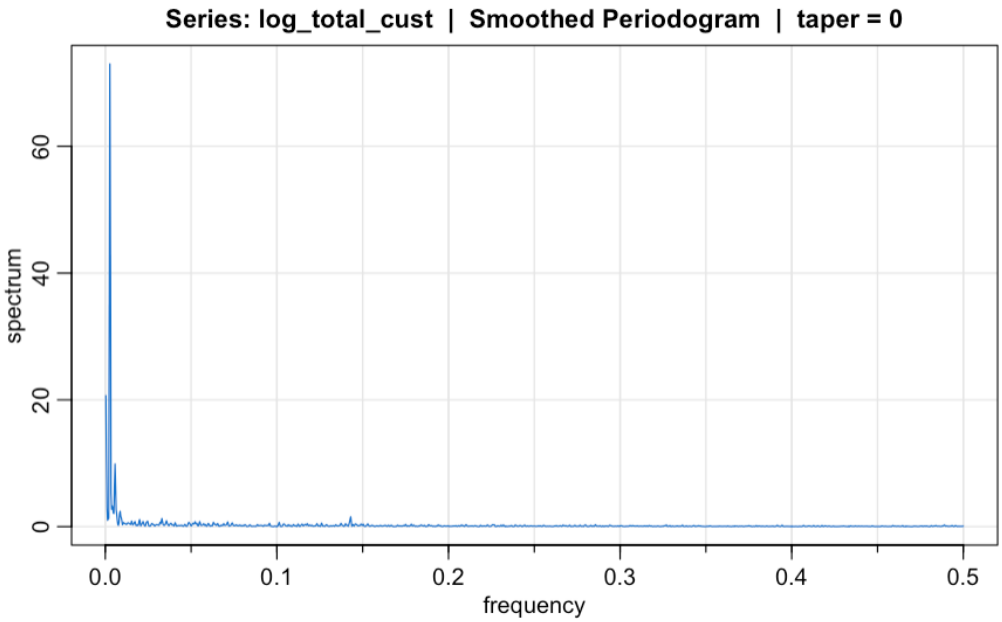

Figure 7: Smoothed Periodogram

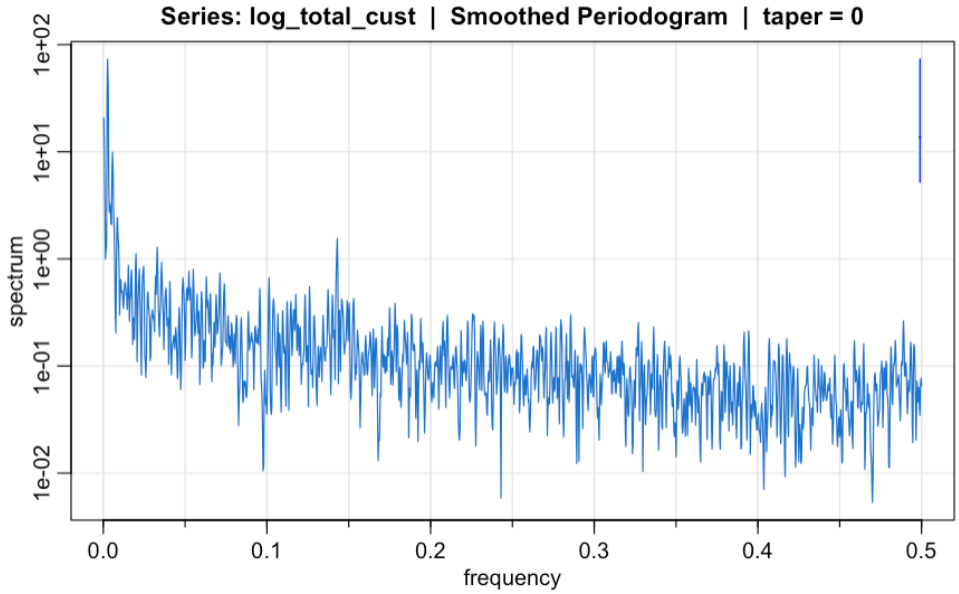

Figure 8: Smoothed Periodogram with confidence interval

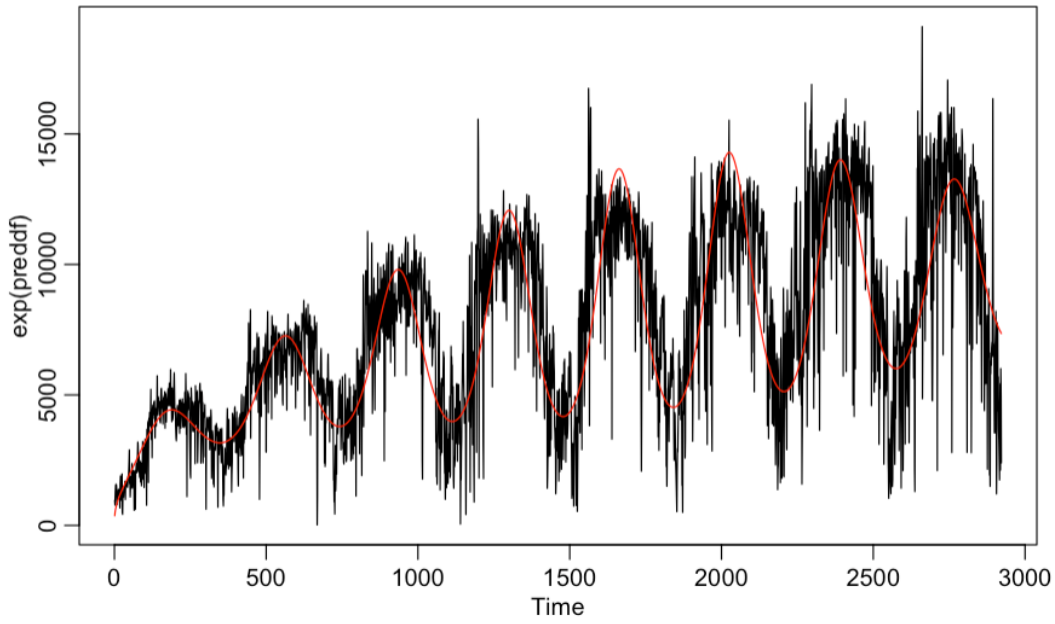

Figure 9: Actual vs Model Simulation (t + 2 cosine trend)

```
Call:
lm(formula = log_total_cust ~ log(t) + Xcos + Xsin + Xcos2 +
    Xsin2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4430 -0.1350  0.0609  0.2312  0.9451

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.168969   0.053330 115.675   <2e-16 ***
log(t)       0.374606   0.007564  49.528   <2e-16 ***
Xcos        -0.392065   0.010533 -37.222   <2e-16 ***
Xsin         0.185559   0.010508  17.658   <2e-16 ***
Xcos2        0.141027   0.010509  13.419   <2e-16 ***
Xsin2       -0.012495   0.010513  -1.189    0.235
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4 on 2916 degrees of freedom
Multiple R-squared:  0.6034,    Adjusted R-squared:  0.6027
F-statistic: 887.4 on 5 and 2916 DF,  p-value: < 2.2e-16
```

Figure 10: Summary of fitting cosine trend

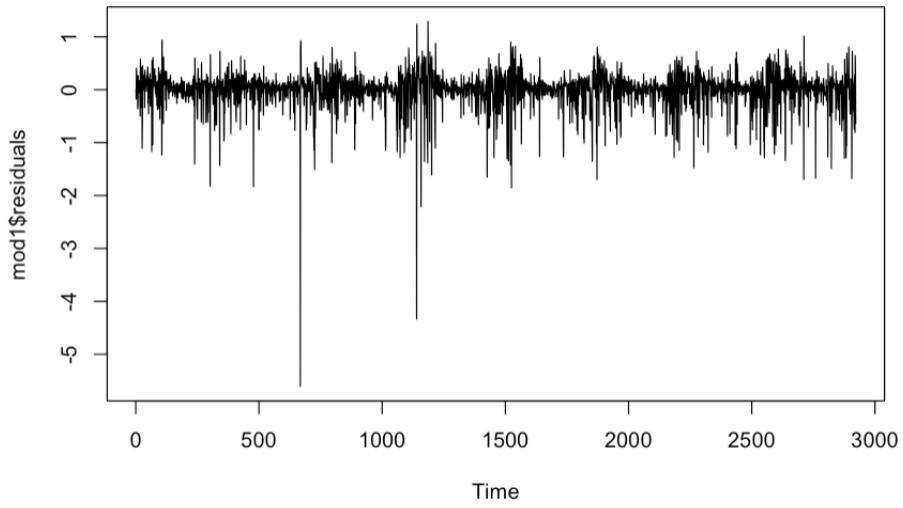

Figure 11: Residuals of ARIMA[1,1,1]

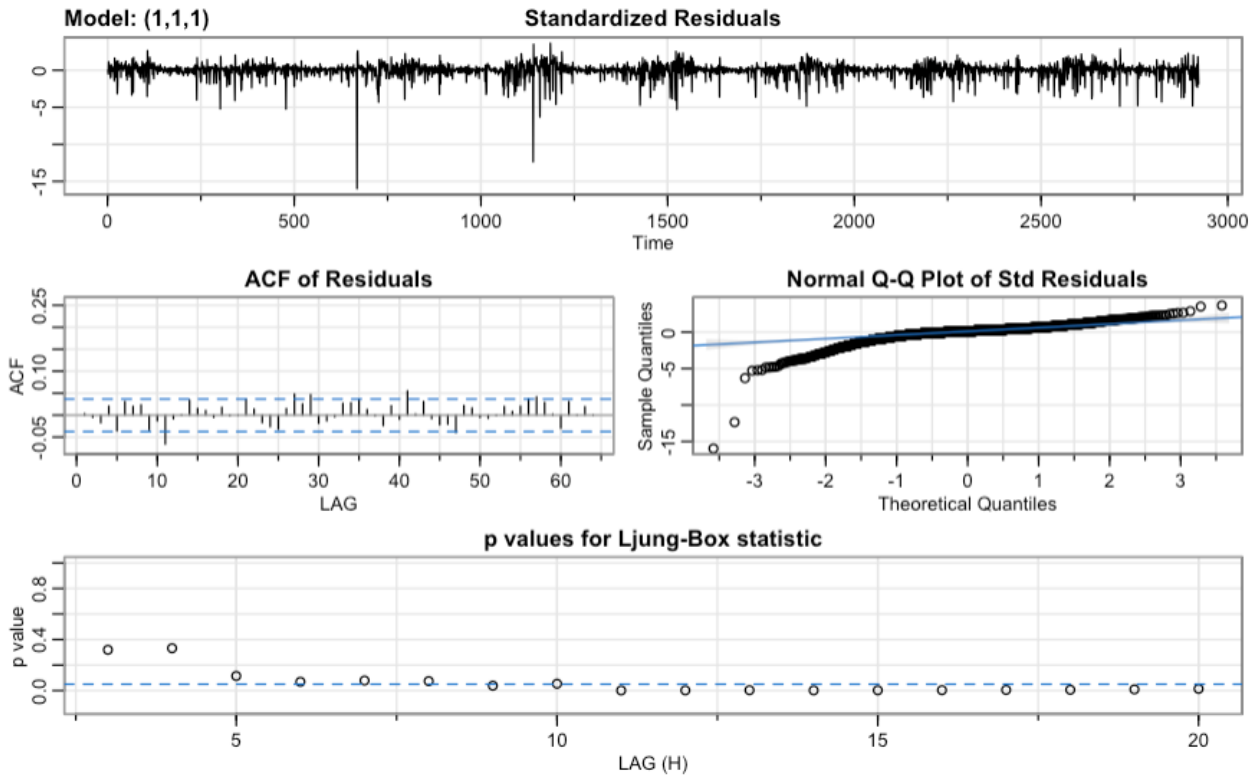

Figure 12: Model Utility tests for ARIMA(1,1,1)

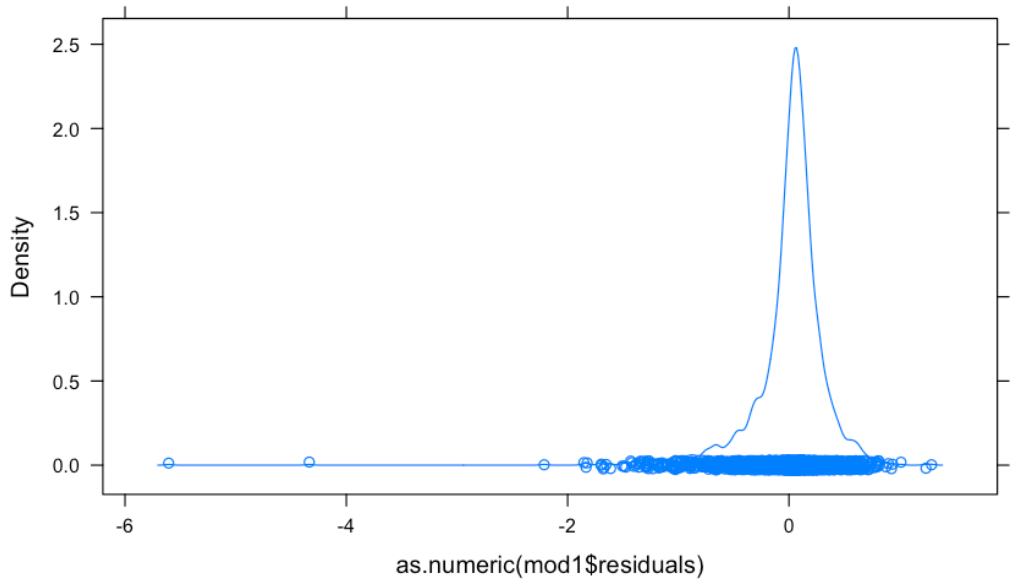

Figure 13: Density plot of residuals from ARIMA(1,1,1)

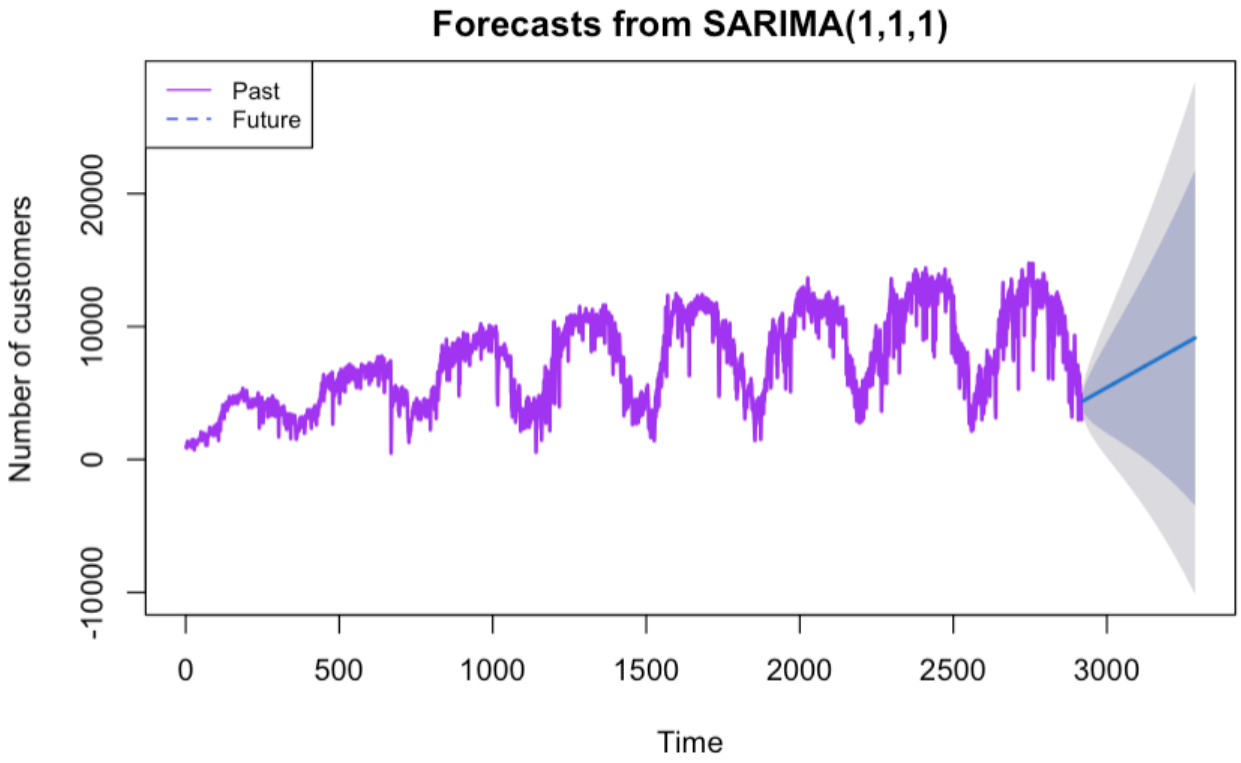

Figure 14: Forecasting with ARIMA(1,1,1)

```
z test of coefficients:

     Estimate Std. Error  z value Pr(>|z|)
ar1 -0.511038   0.058015  -8.8087  < 2e-16 ***
ar2  0.306507   0.032295   9.4908  < 2e-16 ***
ar3 -0.038163   0.020764  -1.8379  0.06608 .
ma1 -0.028493   0.055182  -0.5163  0.60561
ma2 -0.784198   0.050940 -15.3946  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Series: log_total_cust
ARIMA(3,1,2)

Coefficients:
         ar1     ar2      ar3      ma1      ma2
      -0.511  0.3065  -0.0382  -0.0285  -0.7842
s.e.   0.058  0.0323   0.0208   0.0552   0.0509

sigma^2 = 0.1231:  log likelihood = -1083.78
AIC=2179.56   AICc=2179.59   BIC=2215.44
```

Figure 15: Summary of ARIMA(3,1,2)

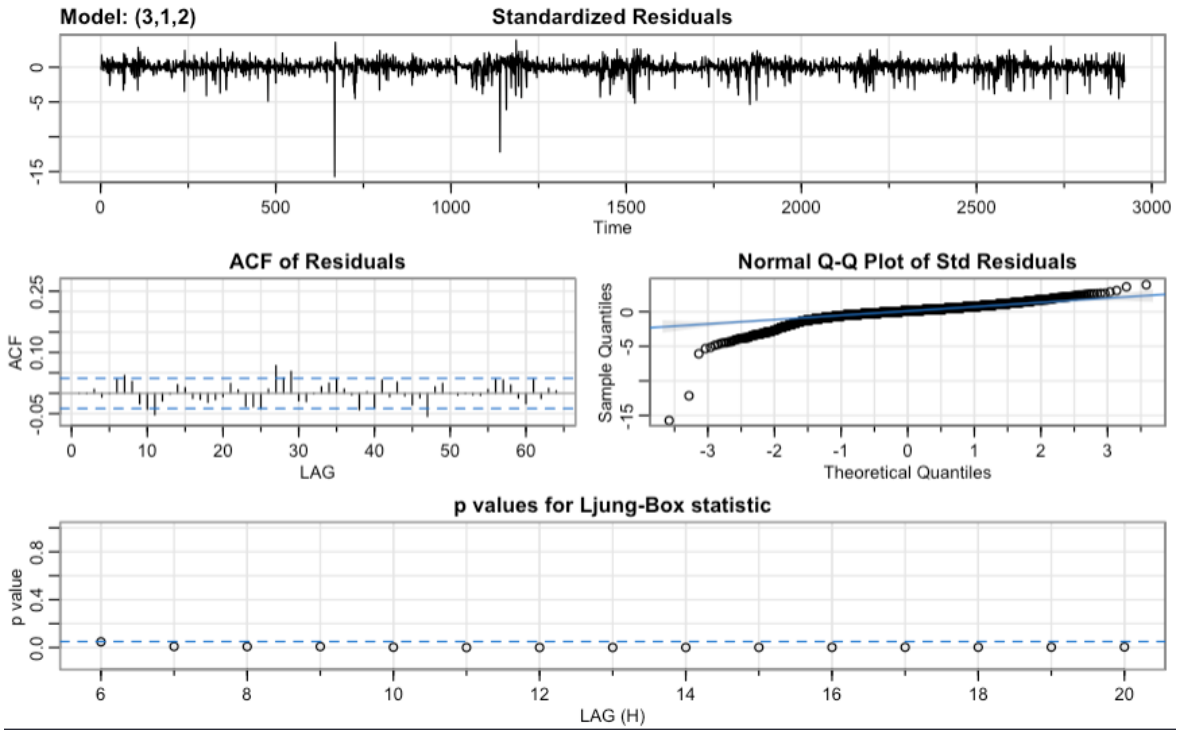

Figure 16: Model utility test for xreg models

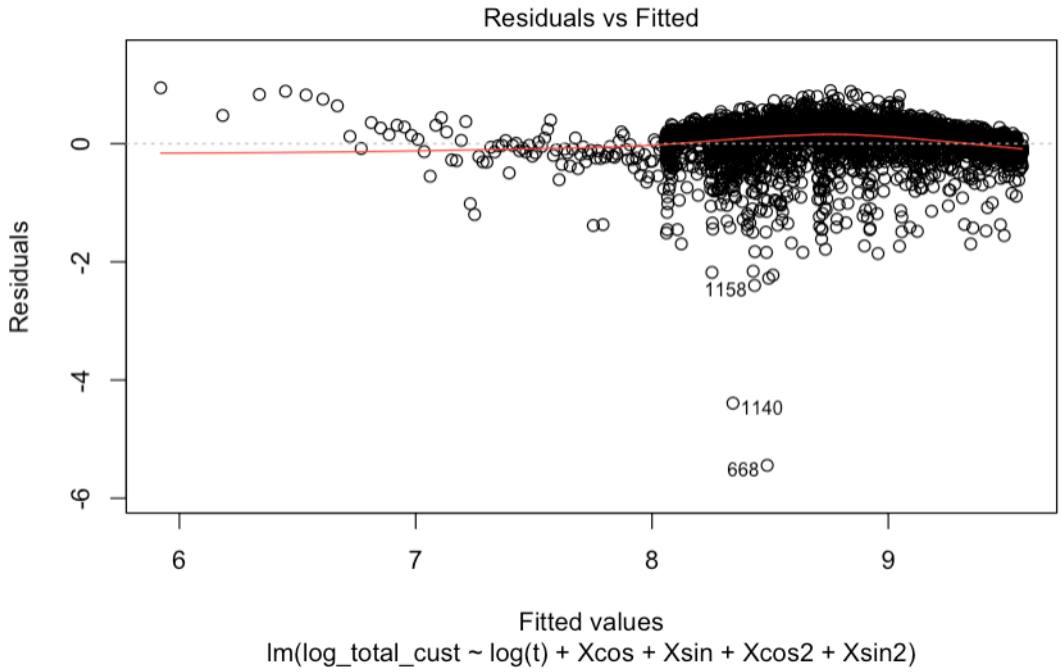

Figure 17: Residual vs Fitted residuals of spectrum model

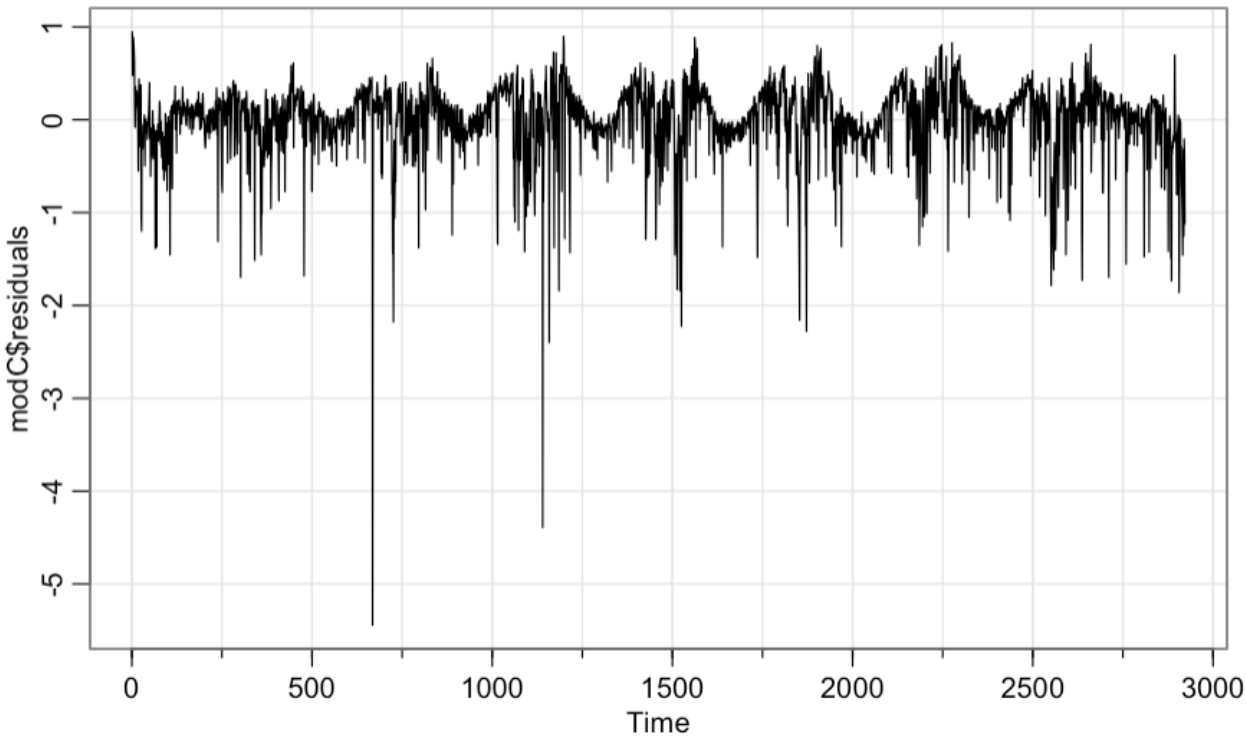

Figure 18: Residuals of spectrum model

```
Call:
lm(formula = log_total_cust ~ log(t) + Xcos + Xsin + Xcos2 +
    temp_max + precip + wind + holiday)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9413 -0.1283  0.0331  0.1932  0.9253

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.291617   0.197344   6.545 7.00e-11 ***
log(t)       0.366539   0.006313  58.063  < 2e-16 ***
Xcos        -0.110570   0.014039  -7.876 4.73e-15 ***
Xsin         0.109027   0.009629  11.323  < 2e-16 ***
Xcos2        0.058720   0.009283   6.326 2.91e-10 ***
temp_max     3.617722   0.137595  26.293  < 2e-16 ***
precip      -0.128663   0.006282 -20.480  < 2e-16 ***
wind        -0.086542   0.015181  -5.701 1.31e-08 ***
holiday     -0.387205   0.051962  -7.452 1.21e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3337 on 2913 degrees of freedom
Multiple R-squared:  0.7242,    Adjusted R-squared:  0.7234
F-statistic: 956.1 on 8 and 2913 DF,  p-value: < 2.2e-16
```

Figure 19: Mixed Model summary

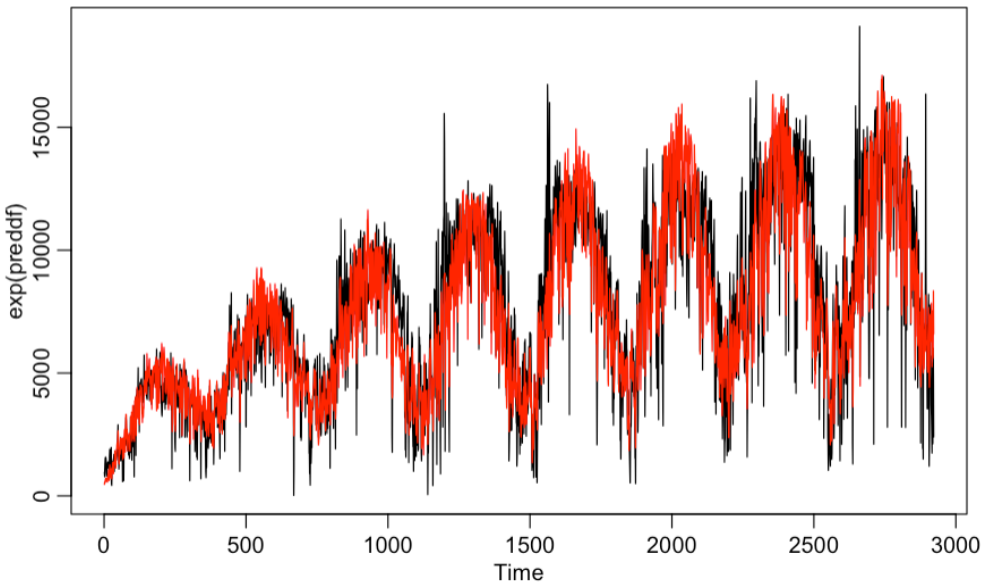

Figure 20: Actual vs Mixed Model simulation