TS9

Dylan Hayashi

MSDS 413, Fall 2022, Section 55

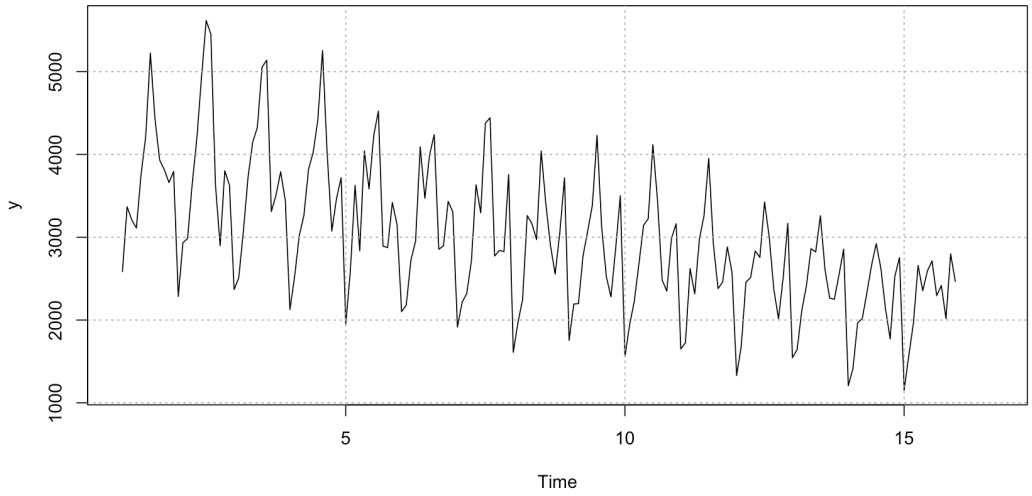
Northwestern University, Time-Series Analysis & Forecasting

November 21, 2022

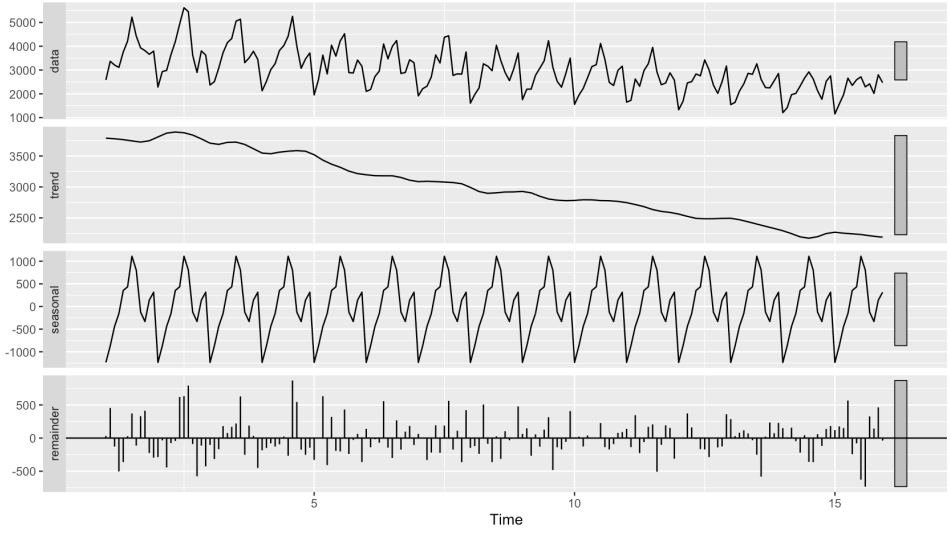
1. Exponential Smoothing

1.1 EDA

The dataset for Australian wine contains a sales variable and is indexed by month, which is then converted to years. To be time series data, they must be sequential observations of a stochastic variable over constant time intervals. The sales variable has variance of 760461.12, indicating it is stochastic. The timeseries is indexed by unique, consecutive months, each with an accompanying observation of sales. This is demonstrated by the vectors sales, Month, and unique(Month) all having the same length of 188. Given that the index contains consecutive months, the data are over constant intervals. Thus, the data fits the definition of time series.

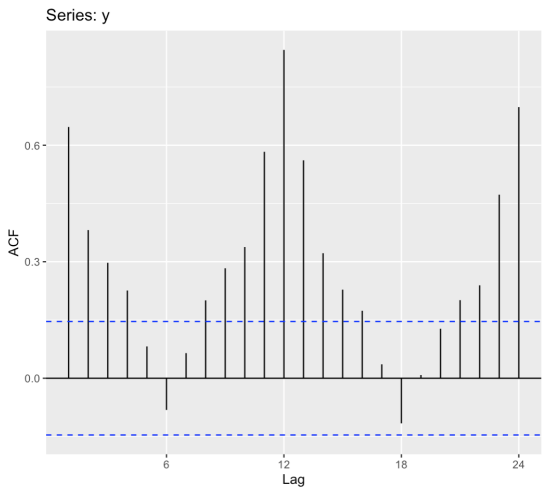


The time plot shows a negative trend. It also has a seasonal component that changes in magnitude in proportion to the level of the series. This indicates that a multiplicative exponential smoothing model is appropriate. Decomposition of the variable, shown below, does not affirm that a multiplicative model is appropriate, as the magnitude of the seasonal component doesn't change. However, it does affirm the negative trend in the data.



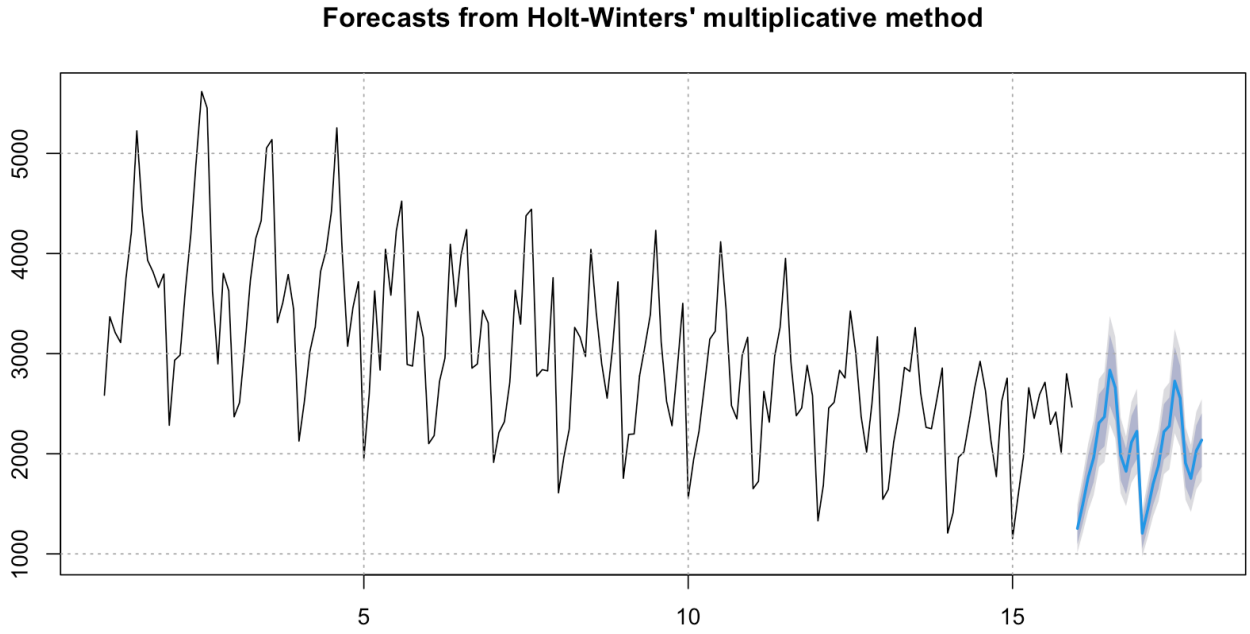
Changes in the magnitude of the seasonal component are likely associated with changes in the overall variance of the variable. A McLeod-Li test returned a set containing all tested lags, indicating that the series has non-constant variance. This is not itself indicative of changes in the seasonal component of the data, but it is related.

The ACF plot to the right shows clear auto-correlation, with numerous spikes above the significance line. There is also a pattern in the spikes indicative of seasonality, as shown in the original timplot.



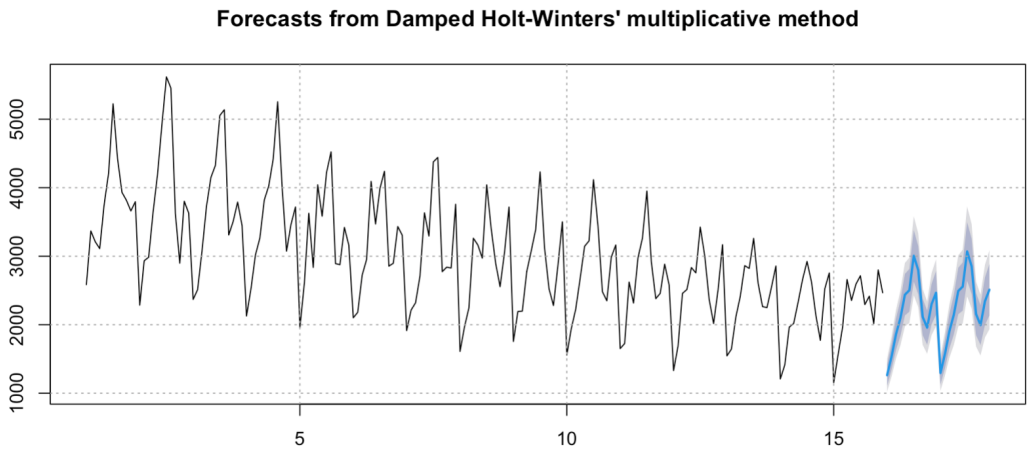
1.2 Forecast

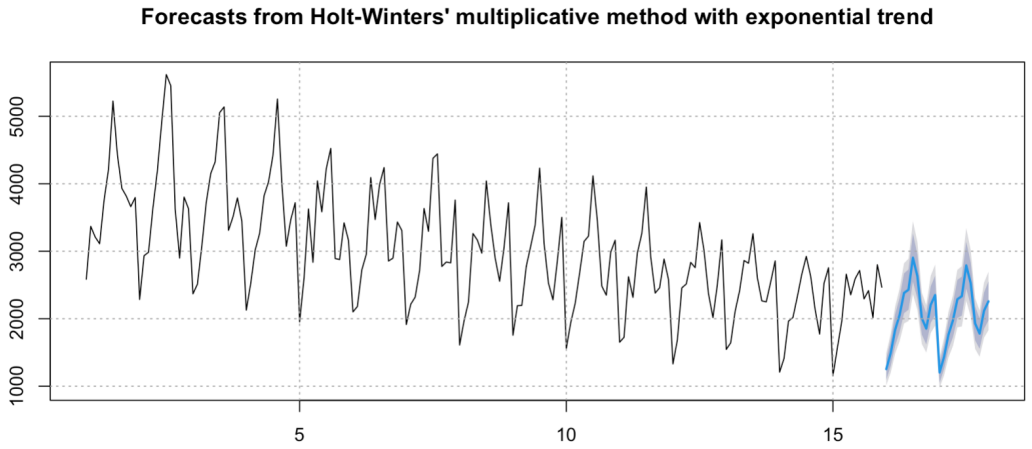
Below is a forecast of sales over the next two years by a multiplicative Holt-Winters' model (1). I think this forecast looks quite reasonable; it seems to accurately describe the seasonal pattern both in steps and in magnitude. The forecast also captures the overall downward trend in the data and has small confidence intervals.

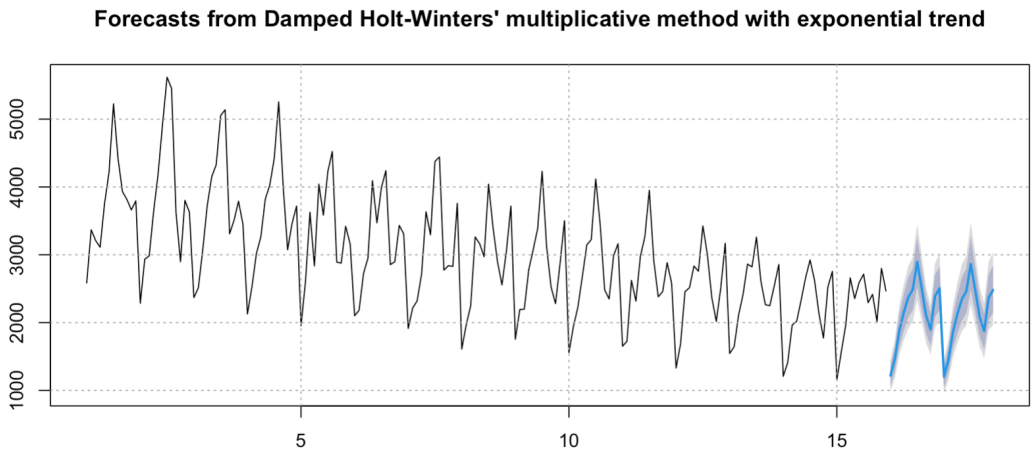


1.3 Experimenting with Trends

Below are two year forecasts produced by multiplicative Holt-Winters models with exponential trend (2), dampened trend (3), and both exponential and dampened trend (4). Both dampening and exponential trend increase the y-range of the forecast and the width of the confidence intervals. Overall, these result in less precise forecasts. I prefer the simple Holt-Winters' model (1) because it provides the most precise forecasts.







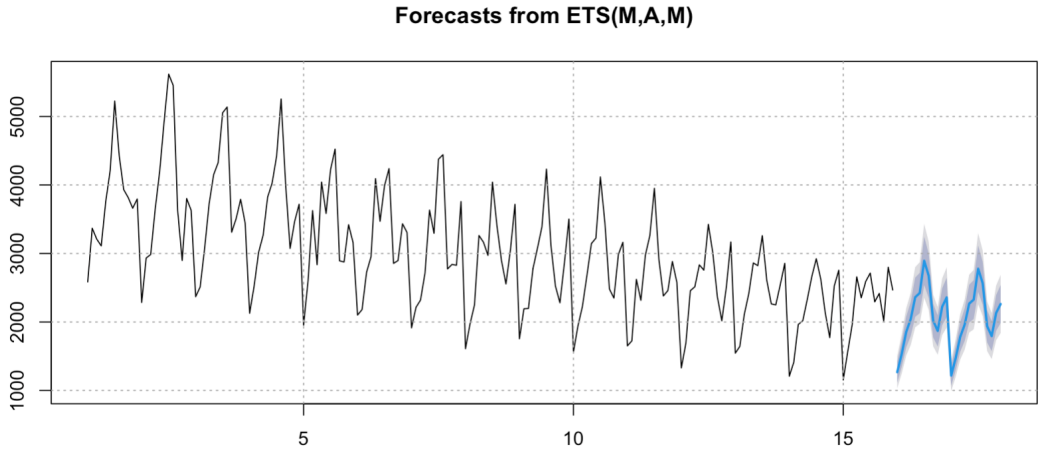
1.4 Error Comparison

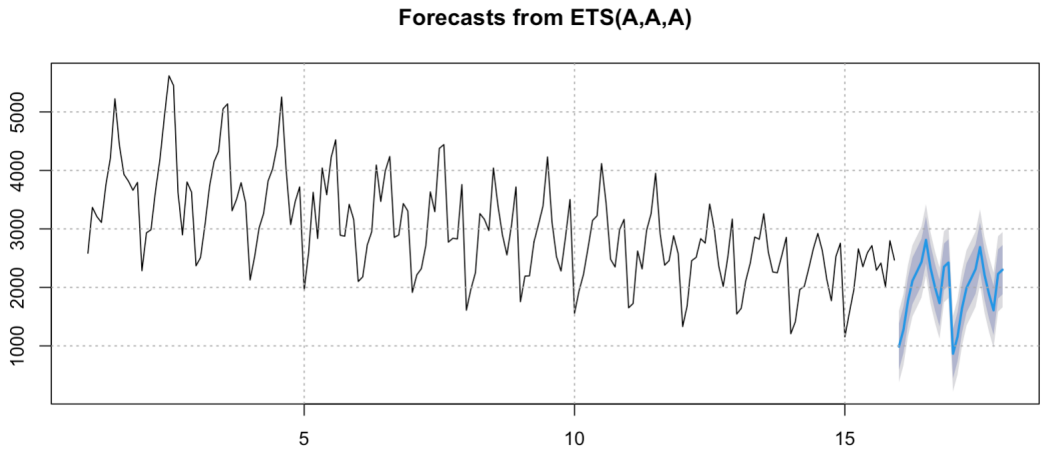
Below are accuracy statistics for the four models. The RMSE of the first model is lowest. Notably, all other error statistics, except for ME, are lowest on the first model. These values indicate that the first model is most accurate.

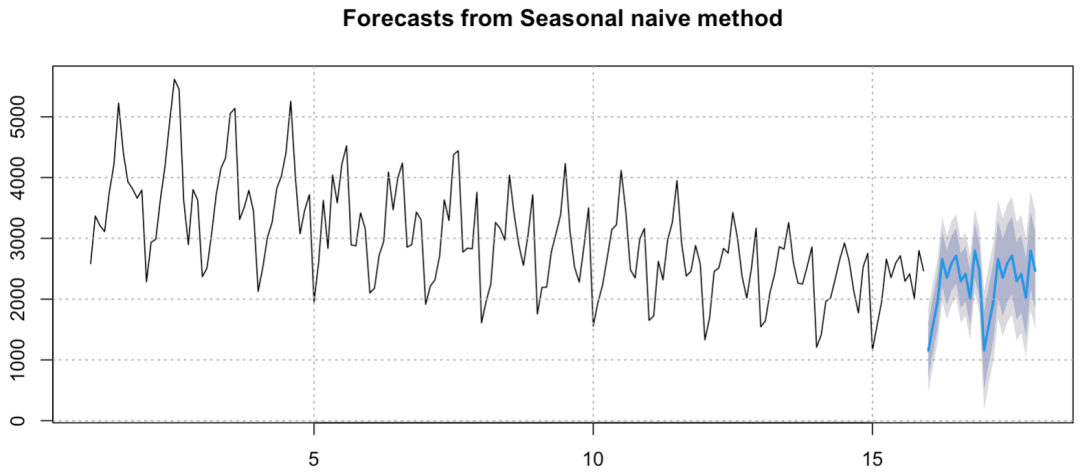
Model	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
fit_multi	33.360	281.119	217.914	0.500	7.175	0.796	0.061
fit_multi_exp	-6.878	288.956	223.317	-0.894	7.383	0.816	0.073
fit_multi_damped	-18.849	290.488	227.826	-1.276	7.560	0.832	0.080
fit_multi_exp_damped	-34.459	286.622	226.672	-1.931	7.617	0.828	0.061

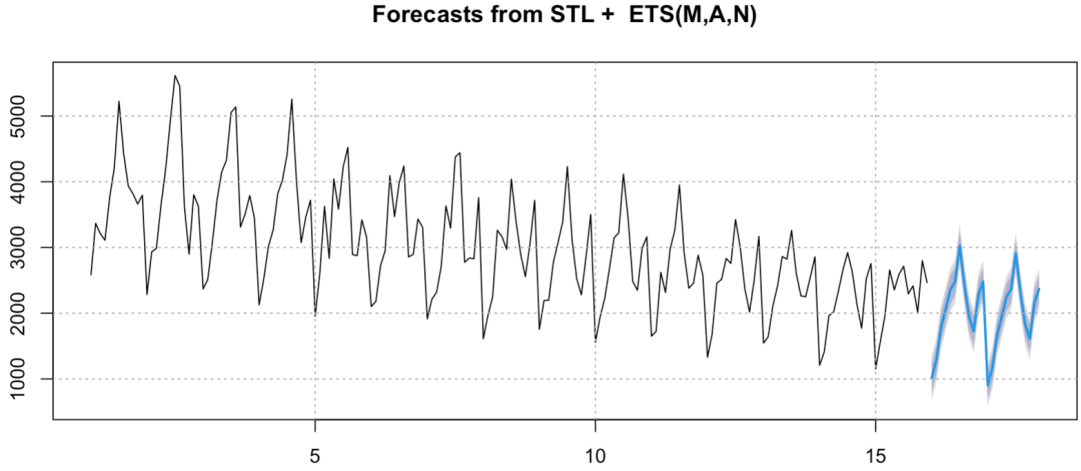
1.5 Model Creation

Below are the forecasts made by an ETS model (5), an additive ETS model fitted on Box-Cox transformed data (6), a seasonal naive model fitted on Box-Cox transformed data (7), and a combination model with STL decomposition on Box-Cox transformed data with an ETS model fitted to seasonally adjusted data. (8).









1.6 Model Comparison

Model	RMSE
fit_multi	281.119
fit_mam	287.464
fit_ana_box	295.900
fit_naive	350.264
fit_stld	241.232

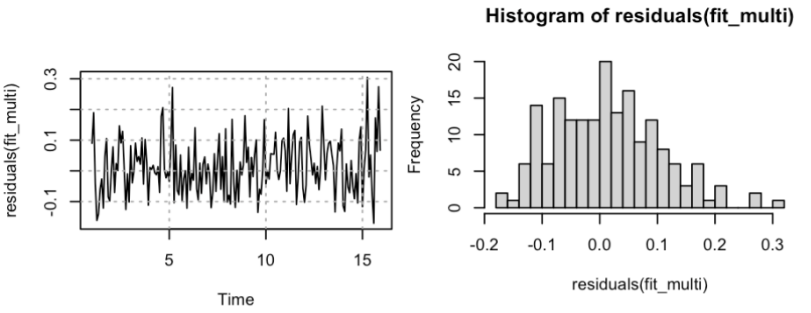
The most appealing of the forecasts above is that of the decomposed and seasonally-adjusted ETS model (8). Its trend line is quite similar to those of the other models, but it features the smallest confidence intervals. This model has the lowest RMSE, as shown in the table to the right.

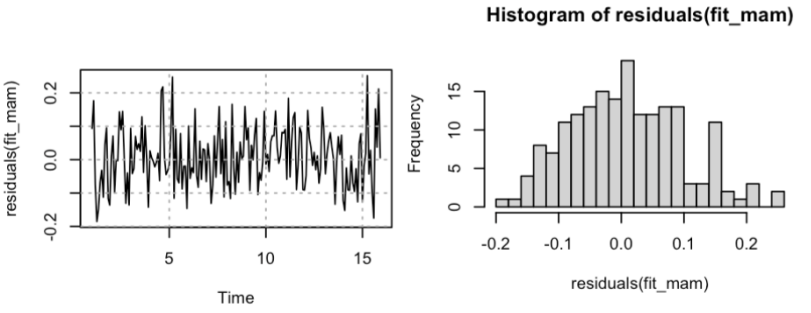
The forecasts produced by the ETS (5) and additive ETS (6) models look similar to those from the previous section. Their magnitude, frequency, and trend seem to match those of the data well. The naive model (7) forecasts the most recent seasonal pattern, which does not look representative of the regular pattern, as it contains a higher number of steps around the peak. Additionally, the confidence intervals around the naive's forecasts are extremely wide relative to the others.

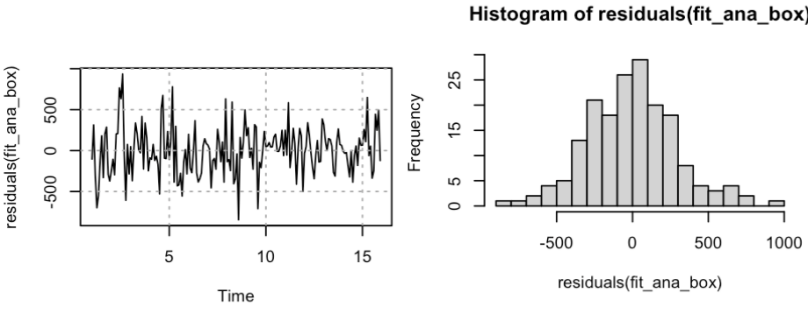
Time plots of the models' residuals, shown on the right, have generally similar features. They appear to be mean-zero and have volatility that is close enough to constant. These features are indicative of stationarity. The time plots also show some similar clusters of variance, notable around years 2, 5, and 16. Sinusoidal patterns are most easily observable in the residuals of the simple Holt Winters multiplicative (1), additive ETS (5), and seasonally adjusted ETS (8) models. These are patterns of variance that the models were not able to explain.

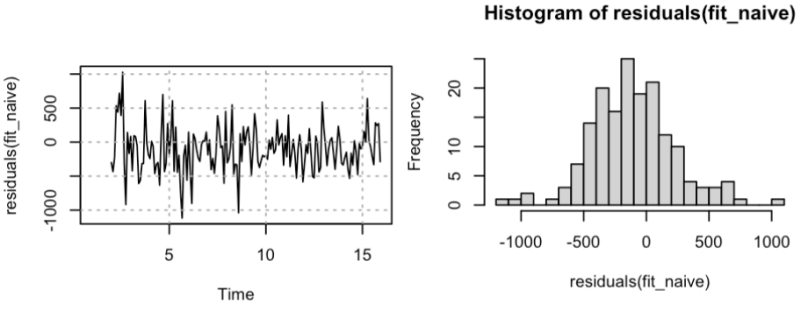
The histograms of the models' residuals depict varying degrees of normal approximation. The residuals of the naive (7) and additive ETS (6) models have the least skew and kurtosis. The residuals of the multiplicative Holt Winters (1) and multiplicative ETS (5) are quite flat and don't have much of a bell shape. Lastly, the seasonally adjusted ETS (8) model's residuals have a good kurtosis and slight right skew, but are overall desirable.

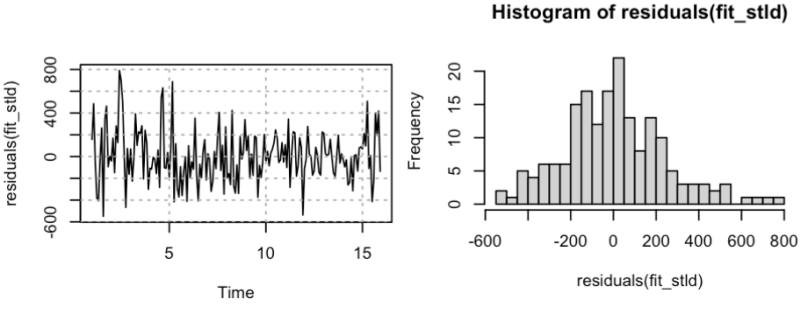
I prefer the seasonally-adjusted ETS (8) model to the rest. As stated in the beginning of this section, I think the forecast of this model accurately reflects the seasonal pattern and trend of the data. It also features the smallest confidence intervals and has the lowest RME. Its residuals are also on par with the rest, indicating it is as good a fit as the others.







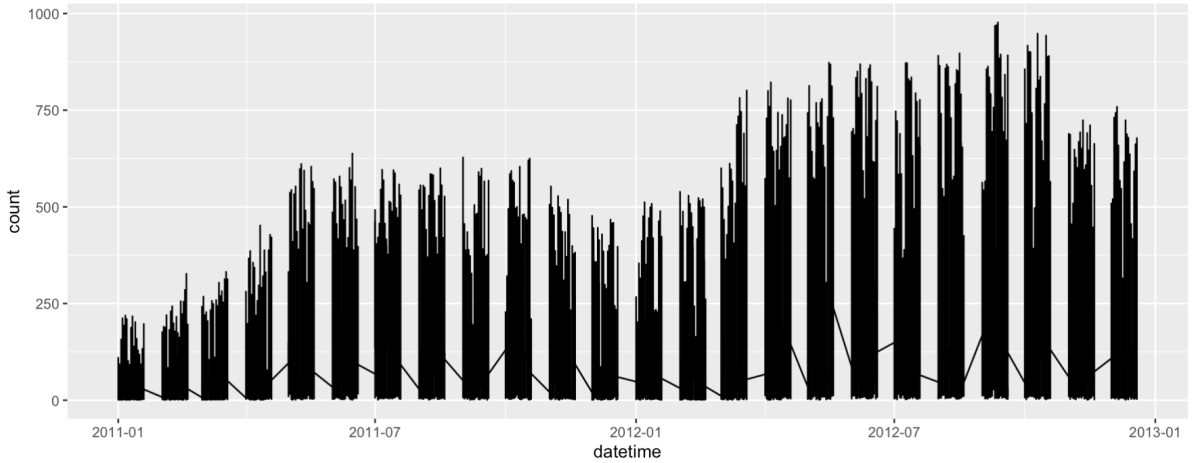




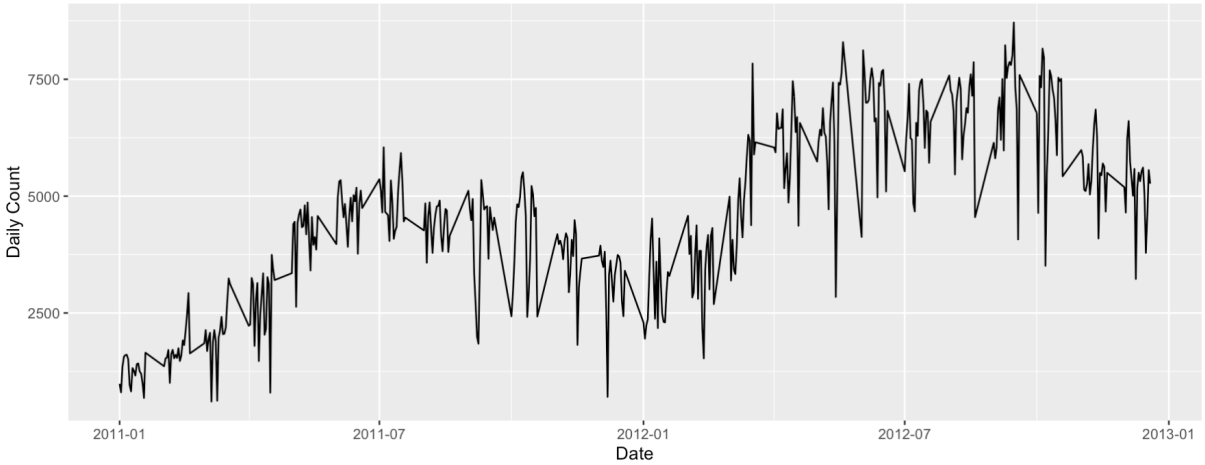
2. Prophet I

2.1 EDA

The bike dataset contains three variables which are important to this analysis: datetime, count, and holiday. Datetime contains hourly timestamps from 2011 through 2012. The count variable indicates counts of bicycles at each timestamp, while the holiday variable is a dummy variable indicating the presence of a holiday. Below is a timeplot of counts over datetime.

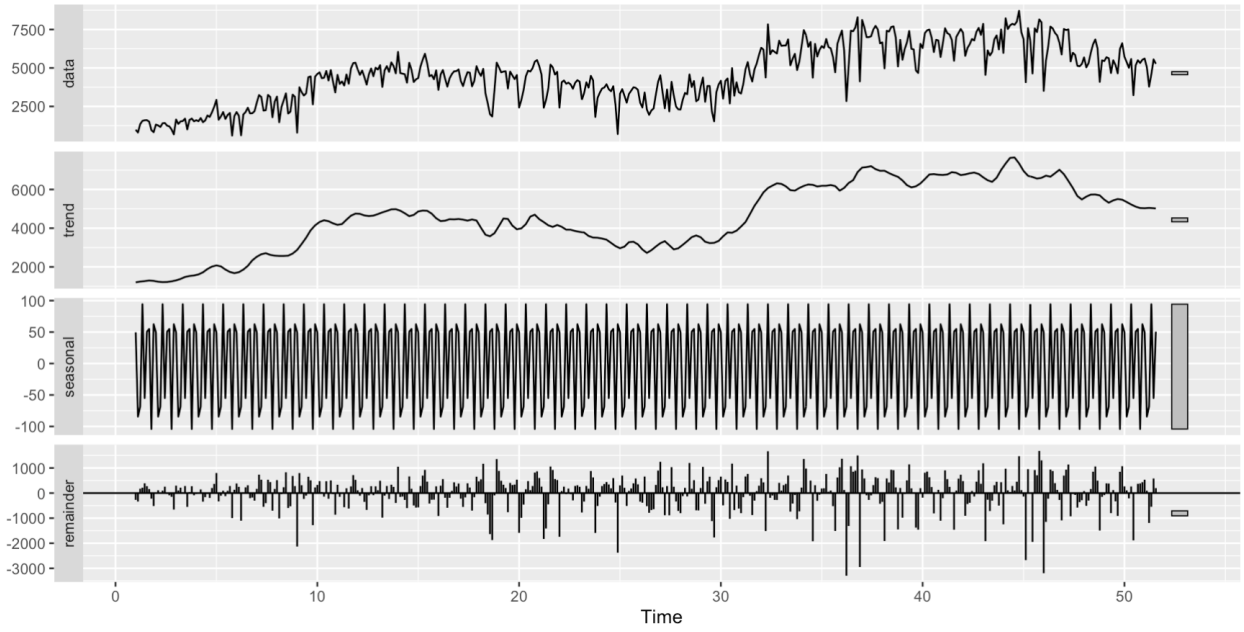


Because the variation in hourly observations is so high, and because there are so many observations, the lines get scrunched, making the time plot very hard to look at. Summing the total count of bikes per day results in a much more tractable series. Prior to analysis, the data should be confirmed to be time series data. The sum(count) variable has variance of 3492189.693, indicating it is stochastic. The vectors sum(count), datetime, and unique(datetime) have equal lengths of 456, indicating that the data are indexed by unique times each with an observation of the variable of interest. Thus, the data are time-index observations of a stochastic variable, fitting the definition of time series.

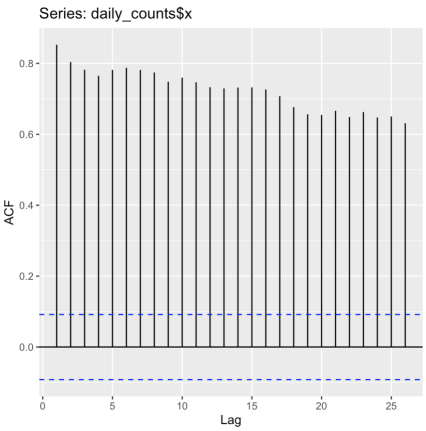


Of additional interest are two other variables: season and holiday. For each day in the date series, these variables describe what the current season is and whether the day is a holiday. The season variable takes four values: 1, 2, 3, and 4. The holiday variable takes the value 0 or 1, with one indicating a holiday. When aggregated by day, these variables have lengths of 456, keeping with the dimensions of the data.

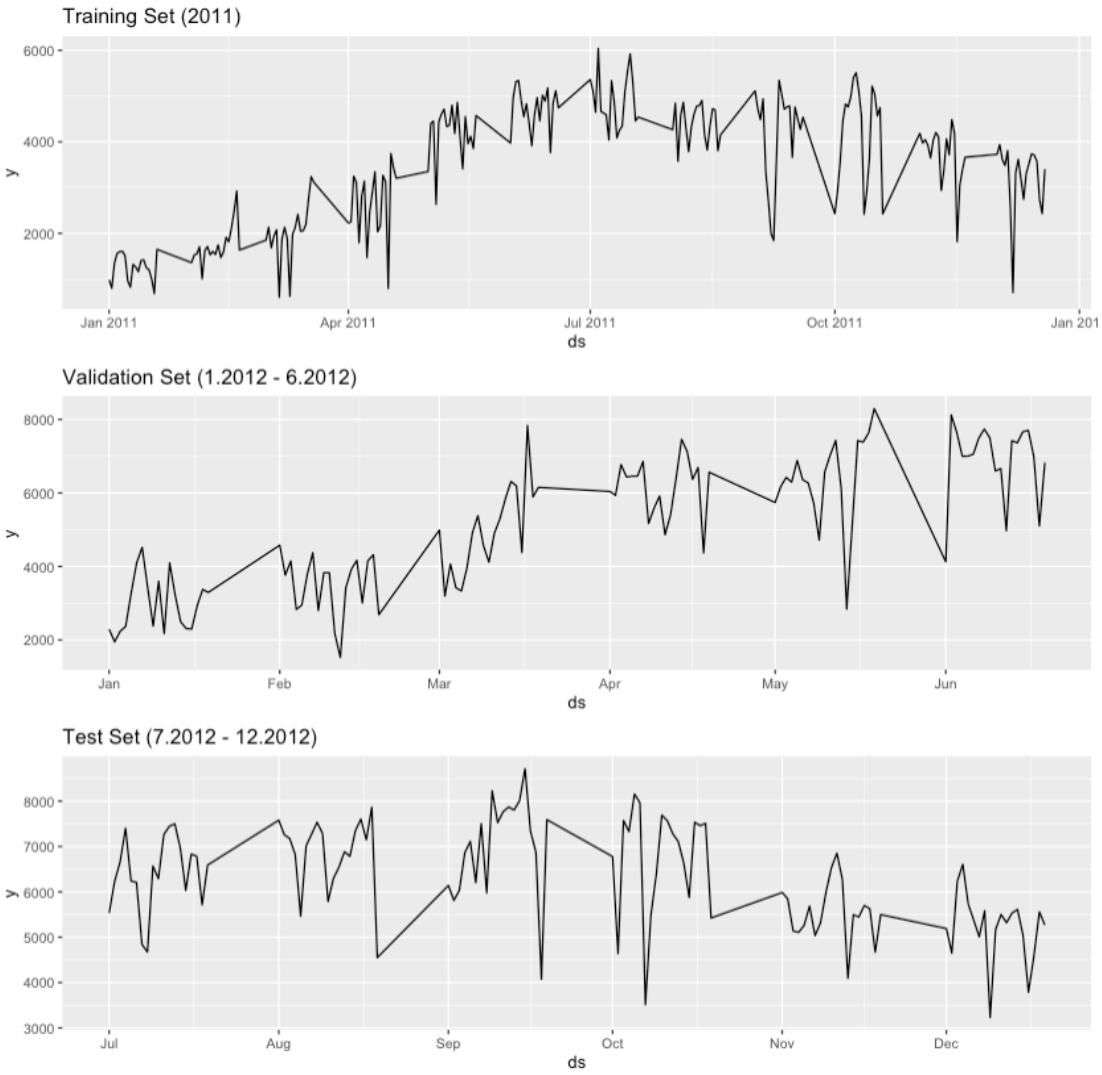
Below is decomposition of daily counts with frequency determined by the findfrequency function (p = 9). The decomposition extracts the trend of the variable well; it is clear and well-smoothed. The seasonal component does not have an easily discernible or simple pattern. Changing the period length made the patterns less discernible. The residuals do not provide much insight either, they seem to roughly match the clusters of variance in the line plot, indicating that the decomposition didn't explain much outside the trend line.



The daily count variable has significant auto-correlation, as demonstrated by the ACF plot on the right. Every tested lag shows a spike above the significance line, and the spikes decrease slightly over time. These features are indicative of a trend in the data, corroborating the trend in the decomposition. There appears to be a wave-like pattern in the peaks of the spikes, but it is harder to discern seasonality than in an ACF plot where the wave-function goes above and below the line y = 0.



For model building and validation, the data are divided into three subsets: training, validation, and testing. The training set contains observations of 2011, while the validation and tests sets contain the first and second half of observations from 2012. The time plots are shown below.



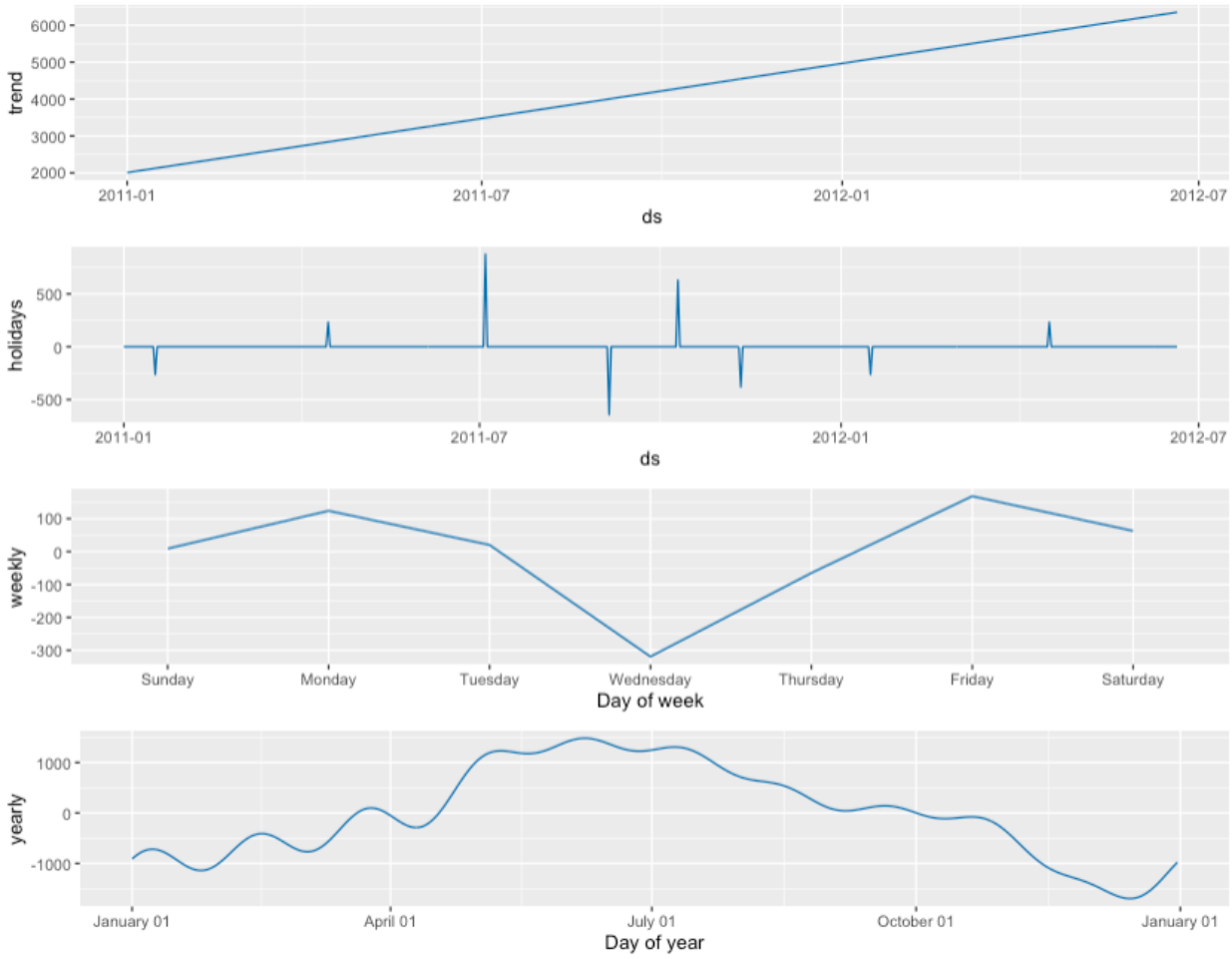
2.2 Holidays

The following holidays have been labeled on the dataset's holiday variable:

```
2011-01-17 Martin Luther King
2011-04-15 Emancipation Day
2011-07-04 Independence Day
2011-09-05 Labor Day
2011-10-10 Columbus Day
2011-11-11 Veterans Day
2012-01-02 New Year
2012-01-16 Martin Luther King
2012-04-16 Emancipation Day
2012-07-04 Independence Day
2012-09-03 Labor Day
2012-10-08 Columbus Day
2012-11-12 Veterans Day
```

2.3 Model Fitting

Below are the components of a prophet model fit on the training data. The trend component shows the model predicts that daily counts will increase by an average of 2000 over the first six months of 2012. The holidays plot indicates that there are spikes associated with holidays, the largest in magnitude being +900 on Independence Day (7/4/12), -700 on Labor Day (9/5/12), and +650 on Columbus Day (10.10.12). The model identified a weekly seasonal component that peaks around the weekday and drops in the middle of the week. There is an additional yearly seasonal component that peaks during the summer months and reaches its bottom in the early winter.

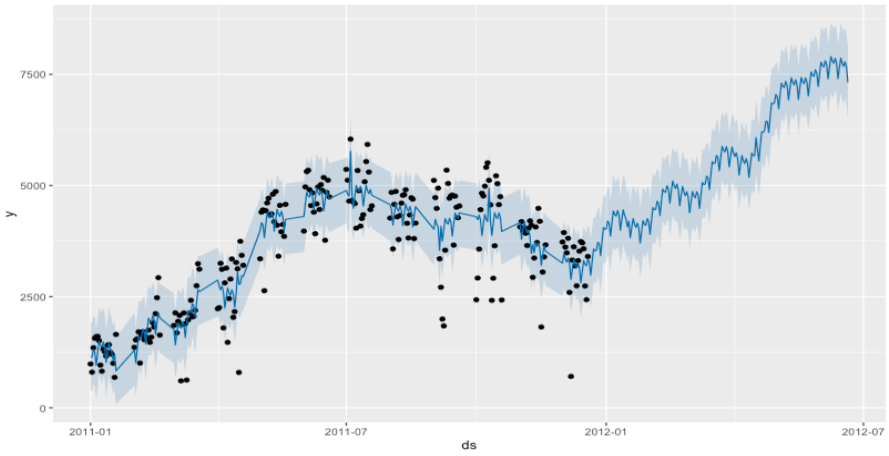


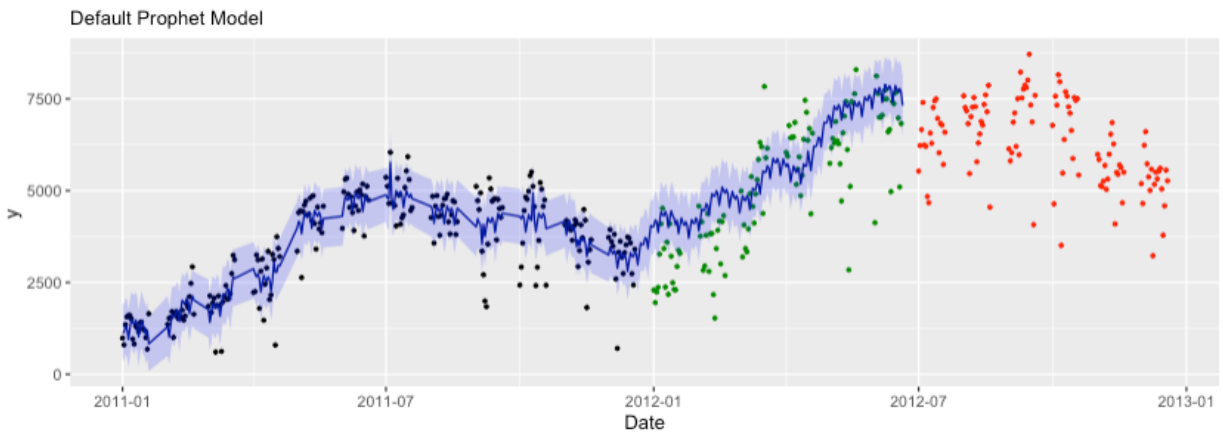
2.4 Model Forecast

Below are time plots of the model's fit and forecast. The fit line in the first plot travels from clusters of observations in a way that looks reasonable, as do the confidence intervals. This indicates a good fit of the data. As discussed in the previous section, the model predicts that daily counts will increase over the first six months of 2012. The plots below indicate that the level is predicted to increase to an average of 7500.

There are three main sinusoidal patterns that the model has identified in the data. The first is that of the weekly seasonality, which can be seen in the smallest peaks and valleys of the prediction line. The second is that of the yearly seasonal component, shown in the component plot on the previous page. In the second plot below, there are two periods, showing two peaks during the summers and two valleys during the winters. The third can best be seen in the forecast line in the first plot. Between January and April of 2012, the confidence intervals show a sinusoidal pattern with frequency of about 1.5 months, showing three peaks and three valleys. This is actually part of the annual seasonal component, but I think it has a distinct enough pattern to be possibly another seasonal component that this model does not include.

The prediction line has the right general upward trend, although I think a lower initial level and higher slope would be a better fit. I don't think the smaller sinusoidal pattern shown in the prediction line matches the green observations very well, although it is hard to analyze.





2.5 Model Accuracy

The root mean squared error (RMSE) of this model was 1292.66. RMSE, which is equal to the average squared residuals of the model, is a goodness of fit metric. Mean average percent error (MAPE), which for this model was 27.57, is another goodness of fit metric. MAPE is the average residual of the model expressed as a percentage of the observation. These measures of accuracy are useful for assessing the relative accuracy of multiple models.

3. Prophet II

3.1 Parameter Matrix

Below is a matrix that contains all possible combinations of settings for five different parameters. They are:

1. Changepoint flexibility values: (0.05, 0.25, 0.5)
2. Seasonal pattern strength: (1,10)
3. Holiday pattern strength: (1,10)
4. Capacity: (7000, 8000)
5. Growth (logistic)

	cps	sps	hps	capacity	growth	RMSE	MAPE
1	0.05	1	1	7000	logistic	1207.026	25.42505
2	0.25	1	1	7000	logistic	1201.515	25.23821
3	0.50	1	1	7000	logistic	1198.301	25.11132
4	0.05	10	1	7000	logistic	1207.528	25.43641
5	0.25	10	1	7000	logistic	1201.479	25.23605
6	0.50	10	1	7000	logistic	1199.271	25.14928
7	0.05	1	10	7000	logistic	1207.218	25.43288
8	0.25	1	10	7000	logistic	1201.053	25.22060
9	0.50	1	10	7000	logistic	1198.342	25.11404
10	0.05	10	10	7000	logistic	1207.639	25.44623
11	0.25	10	10	7000	logistic	1201.126	25.22333
12	0.50	10	10	7000	logistic	1199.283	25.15136
13	0.05	1	1	8000	logistic	1280.935	27.37960
14	0.25	1	1	8000	logistic	1266.452	26.96650
15	0.50	1	1	8000	logistic	1261.135	26.80616
16	0.05	10	1	8000	logistic	1280.798	27.37501
17	0.25	10	1	8000	logistic	1268.824	27.03534
18	0.50	10	1	8000	logistic	1258.953	26.74041
19	0.05	1	10	8000	logistic	1281.946	27.40641
20	0.25	1	10	8000	logistic	1265.966	26.95247
21	0.50	1	10	8000	logistic	1259.677	26.76245
22	0.05	10	10	8000	logistic	1281.239	27.38665
23	0.25	10	10	8000	logistic	1265.771	26.94619
24	0.50	10	10	8000	logistic	1260.182	26.77736

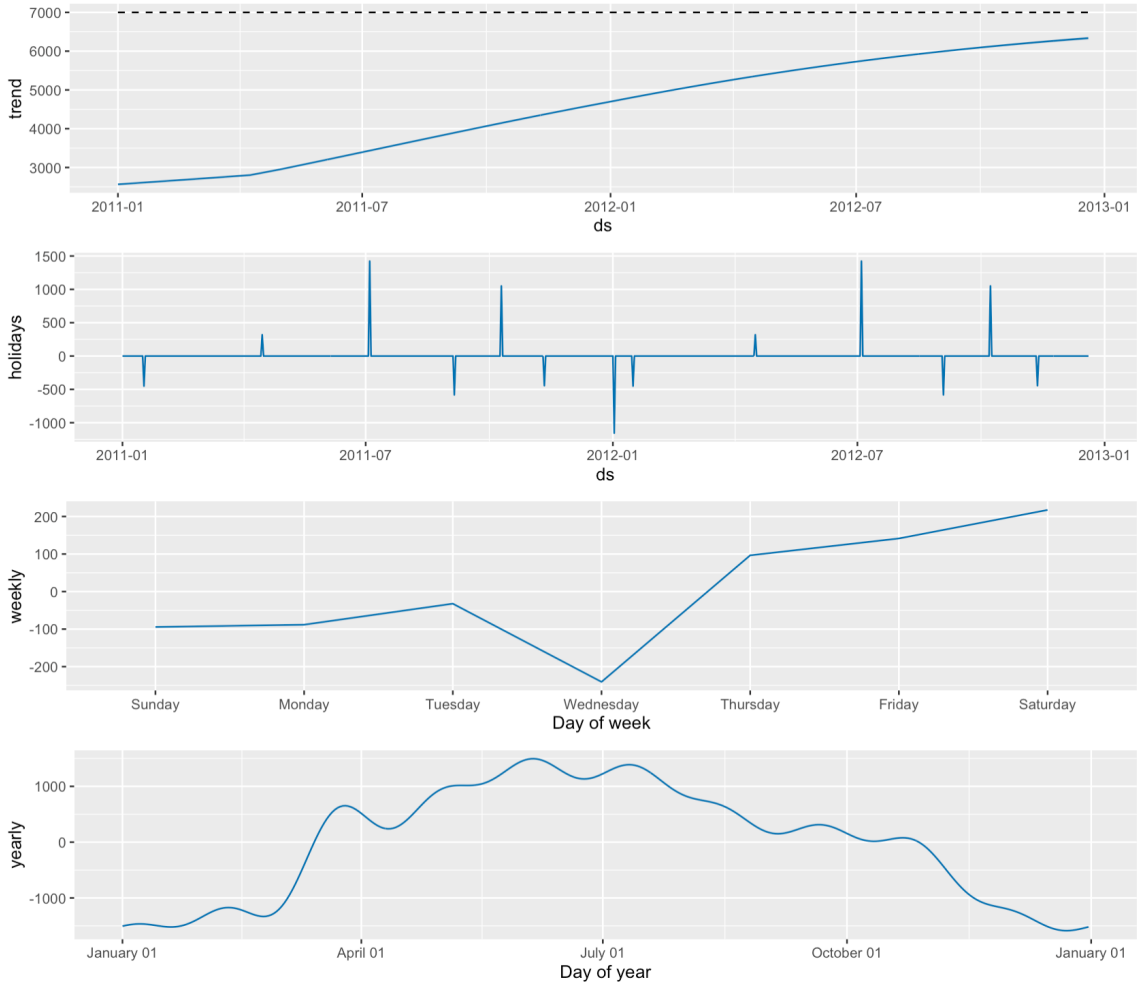
3.2 Optimal Parameter Selection

The matrix on the previous page also includes the RMSE and MAPE of each of the parameter combinations. These error measurements indicate that the third combination of parameter values has the lowest error measurements, shown below. These parameters will be used to retrain the data.

	cps	sps	hps	capacity	growth	RMSE	MAPE	RMSE	MAPE
3	0.5	1	1	7000	logistic	1198.301	25.11132	1198.301	25.11132

3.3 Retraining

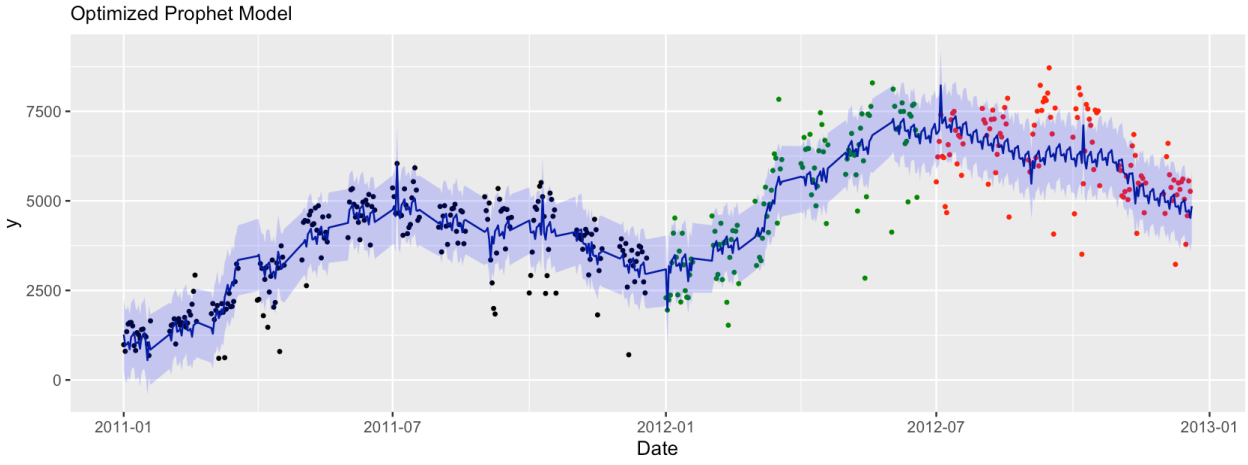
Below are the components of the retrained model, which do show changes relative to the previous model. The trend line in the retrained model is more nuanced, it shows an inflection point and a curve, rather than just a straight line. The holiday component shows that the retrained model's most significant holidays are the same as the first. The weekly seasonal component of the retrained model has shifted up by about 100, and the weekend effect is concentrated in the beginning of the weekend. The overall range of the annual seasonal component has stayed about the same, but the transition between peaks and valleys occurs more quickly in the retrained model, making the slopes of the seasonal transitions more steep.



3.4 Model Forecast

Below is the retrained model's prediction for the second half of 2012. The retrained model shows the three same sinusoidal patterns I described in the first model's forecast. There is a weekly seasonal component, which is shown as the smallest and most frequent peaks and valleys of the prediction line. The second sinusoidal pattern is that of the yearly seasonality component, which shows 2 periods in the whole plot, with peaks during the summer and valleys during the winter. Similar to the previous model, the third sinusoidal pattern of this model lies within the annual seasonal component. It is similarly distinguishable enough that I think it could be another component that the model doesn't include. For the retrained model, this wave has a frequency of 3 months, with peaks around July and October, and valleys around September and December. It is notable that the frequency of this pattern in the first model was 1.5 months, so it has doubled.

In my analysis of the first model's forecast, I said that the forecast could be improved with a lower initial value and a steeper slope. The retrained model's fit line for the first half of 2012 matches this description. I take this as an indication that my notion was correct and that the retrained model may have better predictions. Overall, I do think the predictions of the retrained model are better than those of the first; the retrained prediction line is better centered around the observations, the third sinusoidal pattern is less obvious and seems more fitting of the data, and the slope of the retrained model's forecast line seems better fitted than that of the first model.



3.5 Model Accuracy

The parameters of Prophet II were initially chosen because they showed the lowest RMSE and MAPE values of all tested parameter combinations. These parameters offered improvements of about 10% in these measurements relative to Prophet I, their respective values are shown below.

Model	RMSE	MAPE
Prophet I	1292.66	27.57
Prophet II	1198.3	25.11

Analysis of Prophet II's accuracy over multiple time horizon's produced results similar to those above. The average RMSE was 1228.8, the average MAPE was 21.13. Notably, the plots of these errors show sinusoidal patterns, shown below. This is also true for other measurements, including MSE, MAE, MDAPE, and SMAPE. I perceive this to mean that the model failed to capture some sinusoidal pattern in the data, resulting in the pattern showing up in the model's errors.