Factors Influencing the Performance of VisitorCounts

Dylan R. Way
Department of Mathematics, Western Washington University

January 20 2021

Summary:

This vignette will demonstrate some of the leading factors affecting the accuracy of forecasts produced with this package using some visitation forecasts, with varying data, of two parks of the United States National Parks Service (“NPS”).

Two principal factors concerning the quality of the forecast of park visitation are (1) the amount on-site visitation data and (2) the time periods in which such data are available.

On average, as will be demonstrated, more observations will yield a more precise forecast. No fewer than two one-site observations must be available for a forecast to be generated, but we have found that the amount of error is, on average, greatly reduced when a third observation is made available, and again with a fourth. Thereafter, the the change in the level of error over the number observations available begins to level off.

The time periods from which such on-site data are accumulated can also have a significant impact on the performance on the model, but the ideal time periods for this purpose vary from park to park. However, we have seen that the performance typically improves when the visitation data are taken from the busiest time period(s) of the season.

The actual R code used in finding these results is not shown here, except that which is necessary for acquiring the data set itself, as shown here:

To access this, first, load the package and the accompanying data set:

```r
library(VisitorCounts)
data(park_visitation)
```

Included in this package are the full visitation data for 20 National Parks for the years 2005-2016, together with the corresponding photo-user-days (“PUD”) data from Flickr, both of which are aggregated into monthly data. Photo-user-days represents the number of Flickr accounts that posted at least one photo to the social media site on a given day, within a defined geographical area. Each national park is identified by a four-character identifier. Use the R code below to retrieve the data for individual parks. Yellowstone and Acadia, the two parks used herein, are identified by “YELL” and “ACAD,” respectively.

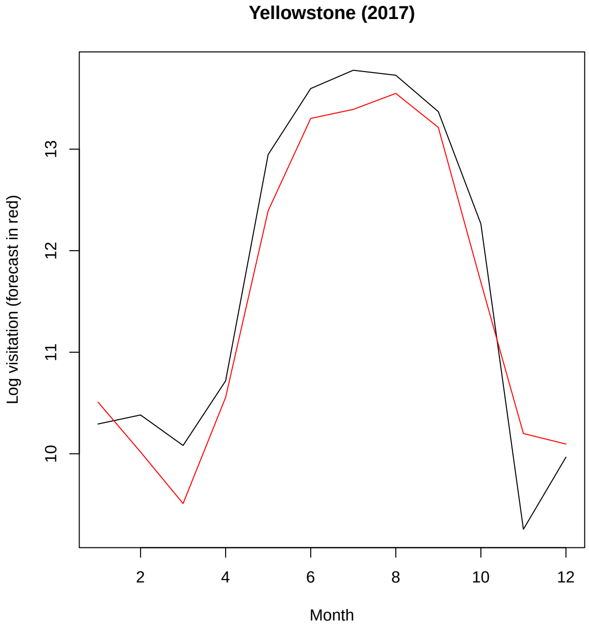
```r
PARK1 <- park_visitation[which(park_visitation$park=="PARK"),]
```

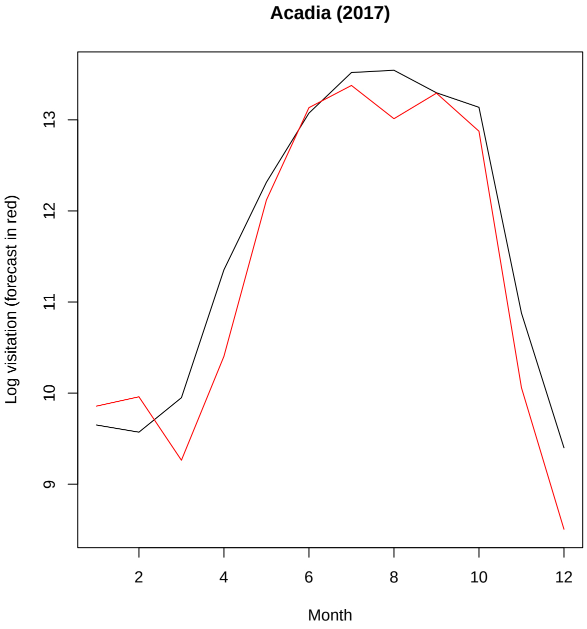
Demonstration:

For this purpose, the data from the first 12 years are used as the training set, and the subsequent 1 year is used in testing.

Forecast with Full On-Site Visitation Data

Using the log of the values of the full NPS training data in model training (144 monthly observations over a 12-year period), we generate the following forecasts for the next 12 months, overlaid with corresponding actual log visitation:

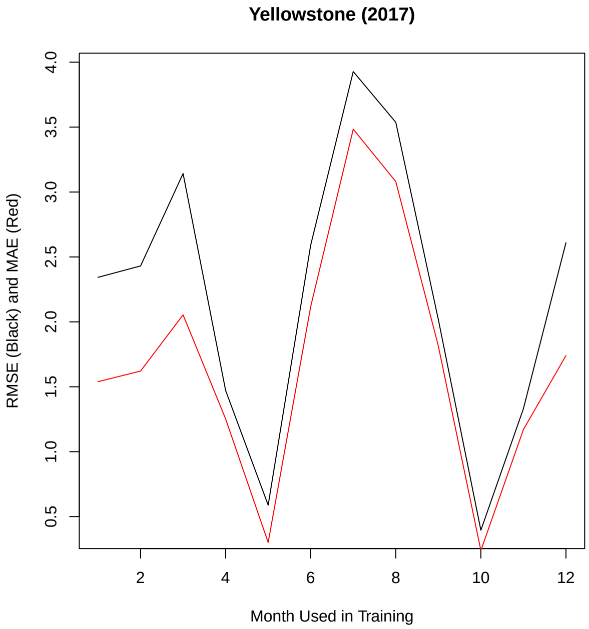

Yellowstone (2017)

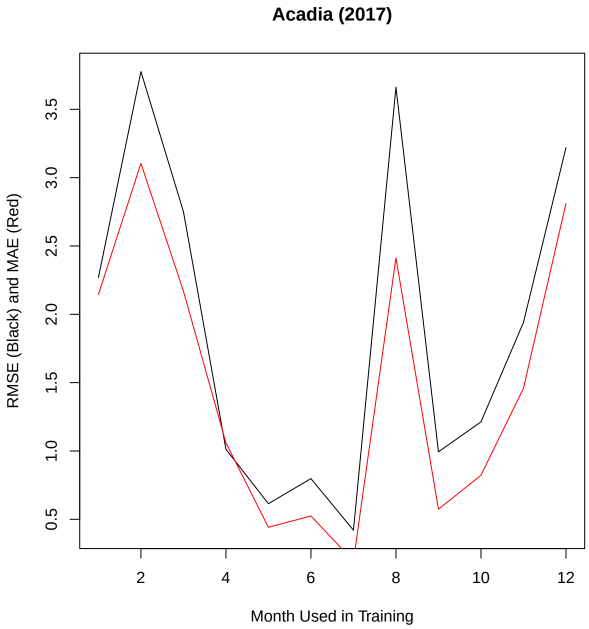

Acadia (2017)

We can see that, in these two cases, the forecast appears to be quite accurate.

Forecast Error Depending on Month for Which On-Site Data is Available

In many practical cases, the on-site visitation data set is very incomplete. Since the model still relies on incomplete data (together with fully-available social media data), we must see which combinations of data are the most useful for this purpose. The following is an evaluation of the level of error in each of 12 forecasts, wherein each forecast is developed using a one-month observation from each of the 12 years in the training set. The y-axis represents which month of each year was used in the corresponding forecast.

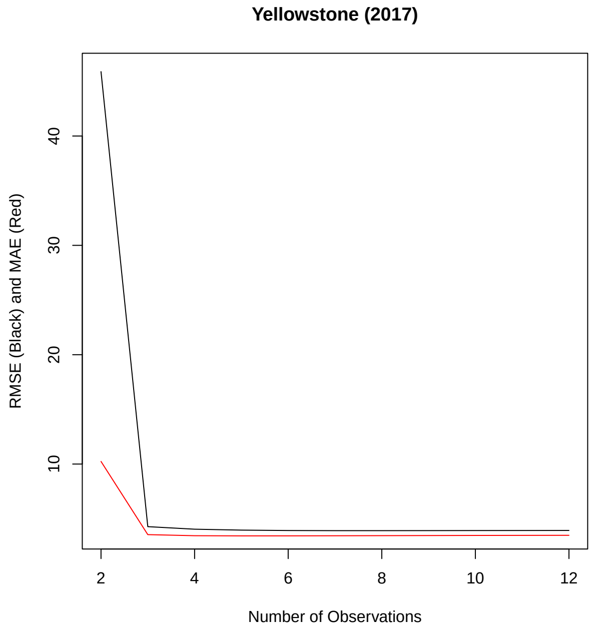

Yellowstone (2017)

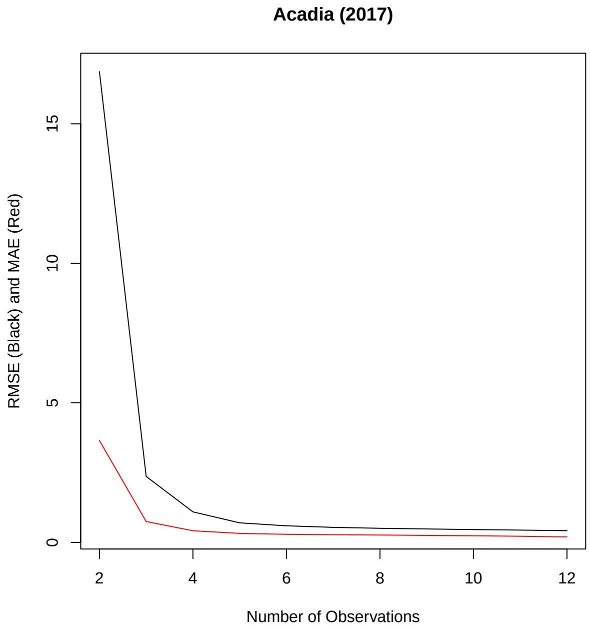

Acadia (2017)

The root mean square error ("RMSE") and median absolute error ("MAE") appear to vary greatly. These findings are far from consistent, but generally the forecast will be most precise when the data are derived from the months for which the forecast using full on-site data were most accurate. For example, the full on-site forecast for Yellowstone was nearly perfect in November, the 10th month of the year, and the model given on-site data of all for the November of each year of the training set had the least error. In practice however, this full on-site data set is not available (hence the need for this package), so it is not clear which data set would produce the most accurate data. However, as stated before, on-site visitation is generally most useful for this model when taken from the busiest time periods of the season. In the case of National Parks, this is usually the summer months of each year.

Forecast Error Depending on the Number of Observations of On-Site Data Available:

The performance of the model is more consistently dependent on the number of on-site observations than the time periods for which such observations are available. As would be expected, a larger number of observations yields a more accurate forecast.


Yellowstone (2017)


Acadia (2017)

The models from which the RMSE and MAE values above are produced with on-site data exclusively from the the month of July. The y-axis shows the number of observations actually used in the model, but since there are many combinations of on-site data for each number of observations, the MAE and RMSE shown herein are those same values, averaged over the models for all possible combinations of the data. For example, in the case of three observations, there are $\binom{12}{3} = 220$ possible combinations of the data from the month of July over the 12-year period of the training set, so the RMSE and MAE values shown in this graph are the mean of those 220 values.[1]

As stated before, it is not possible to produce a forecast with only one observation, so the graphs plot the forecast performances with 2 to 12 observations.

As is shown here, the change in the level of error between two and three observation is very significant, and to a lesser extent, the same can be said for the change of the same between three and four observations. Thereafter, the marginal gains are less significant, and the amount of error between the different amounts of observations begins to level off. While only observations from July are shown here, we have observed the same behavior with other months as well.

[1] To be more specific, for the RMSE calculation, the square root of the mean of all the mean square error (MSE) values is calculated.