Michigan Outdoor Recreation Search Interest Installment 6

This is my project to analyze and forecast Google search interest for 10 forms of oudoor recreation by people in Michigan. My data is daily data for each form of search interest from January 2021 to March 2025. The search interest terms for this project are atving, boating, camping, fishing, hiking, kayaking, rving, hunting, skiing, and snowmobiling. For more information on how this data is pulled, please see installment 5:

https://dataandoutdoors.com/michigan-outdoor-recreation-search-interest-installment-5/

This installment is my first of two or three attempts/methods to model and forecast search interest based on this data.

Nonstationary Model

The method I employ for this installment is a modification of a method I learned in my micromasters course "Learning Time Series with Interventions" from the Statistics and Data Science Micromasters. This method comes from the engineering discipline and specifically relates to signals. Personally, I have mixed feelings on applying methods designed from physical sciences to data generated by humans. Often, studying human behavior and economic data presents its own sets of challenges. However, in this specific project we use Google trends data which is is noisy and error prone due to sampling and other issues. In my Northern Michigan search interest project, I have resorted to averaging multiple pulls of data in order to mitigate this issue. However, this approach is not feasible in this project. Therefore, the prototype of readings from a faulty error prone sensor is one I embrace for this project (at least for now.)

My initial step is to denoise the data and, subsequently, I use this denoised data to estimate a model for the nonstationary series. The nonstationary series is defined as considering the trend and seasonality of the data.

To do this, I create a hankel matrix with the data. Consider a single data series (say atving). First I eliminate the leapday February 29, 2024 from the data. Then I organize the first 366 values into the first column of the matrix (see figure immediately below). The first column begins and ends with January 1st with the last value January 1, 2022. This reason I chose this value is because my original analysis showed both annual and weekly seasonality in the data so by choosing columns a year in length I can consider the 7th and 365th lag in my eventual nonstationary model.

Once I have my first column, I graduate the date interval by one to create my second column. This column includes data values from January 2, 2021 to January 2, 2022. In this manner, I keep creating columns in the matrix until the final value in my column is the final value in the dataset--March 31, 2025. This final column has values from March 31, 2024 to March 31, 2025. Therefore, the dimensions of my final hankel matrix (for each of the 10 data series) is 366 x 1,185.

$$\left\{\begin{array}{ccccc} 1/1/2021 & 1/2/2021 & ... & 3/30/2024 & 3/31/2024 \\ 1/2/2021 & 1/3/2021 & ... & 3/31/2024 & 4/1/2024 \\ ... & ... & ... & ... & ... \\ 12/31/2021 & 1/1/2022 & ... & 3/29/2025 & 3/30/2025 \\ 1/1/2022 & 1/2/2022 & ... & 3/30/2025 & 3/31/2025 \end{array}\right\}$$

Once I have my hankel matrix (for each of my 10 data series), I now need to denoise the data. Note that Google trends data has considerable measurement error and noise. Pulling the data several times will create different results each time. Similarly, pulling the data from the Google trends web interface will produce different results from the pytrends webscraping package. Therefore, I decide to denoise the data, in this case using Singular Value Decomposition (SVD). I use the python package numpy to perform SVD on the hankel matrix which decomposes the matrix into a product of three different matrices: U, S, and V.

$$H = USV^t$$

The matrix S is the concatanation of a diagonal matrix with 366 diagonal values which is padded right with a matrix of zeros with 819 columns such that the end dimensions of S are 366 x 1,185. These 366 diagonal values are singular values organized in rank order based on size. In order to denoise the data, we want to choose a subset of the top singular values. We choose this number 'd' to be such to comprise 99% of the total squared sum of the singular values.

$$s_1^2 + s_2^2 + ... + s_d^2 > 0.99(s_1^2 + s_2^2 + ... + s_d^2 + ... + s_{366}^2)$$

Once I know the number 'd' of components I want to use, I set all other component values to zero. I can then multiply with U and V to get a denoised matrix H hat using S hat.

$$\widehat{H} = U\widehat{S}V^t$$

The following table shows the number of singular values used for each of the 10 models. At most, the models require 215 out of 366 singular values to account for 99% of the squared values (for hiking). At minimum, the models require 12 singular values (for camping).

	Model	Singular Values
0	Atving	201
1	Boating	94
2	Camping	12
3	Fishing	50
4	Hiking	215
5	Kayking	160
6	RVing	59
7	Hunting	198
8	Skiing	108
9	Snowmobiling	202

Once I have the denoised hankel matrix for all 10 data series, I am ready to estimate my non-stationary models. I do this individually for each of the 10 series using a form of Ordinary Least Squares (OLS). In order to estimate the model, I take the final row of my original hankel matrix H. This is my observations 'y' which is a 1,185 x 1 matrix (after I transpose it). These are the daily values from January 1, 2022 to March 31, 2025.

The predictor or explanatory variables are all the other rows of the hankel matrix, for which I use the denoised hankel matrix H hat. This is essentially the denoised lags of the observations, up to the 365th lag. Therefore, the 'X' matrix is 1,185 x 365 (after I transpose it). Using OLS estimatation, I achieve a model with 364 parameters, which is one for each of the 365 lags (minus 1 as it's a linear combination of the others). This allows me to account for annual seasonality (365th lag), weekly seasonality (7th lag), along with a fairly arbitrary functional form for trend.

Stationary Model

Once I estimate the stationary models above, I extract the 1,185 residuals from the estimation of each of the 10 series. These residuals are considered to be stationary. Unlike the setup I learned in the class project where we estimated a single stationary series and employed an ARIMA model, here I combine all ten series and use an Vector AutoRegressive (VAR Model). The VAR model considers not only lagged values of the same series (say ATVing) but also lagged values of other series (such as camping). Here I consider up to one lag of each series.

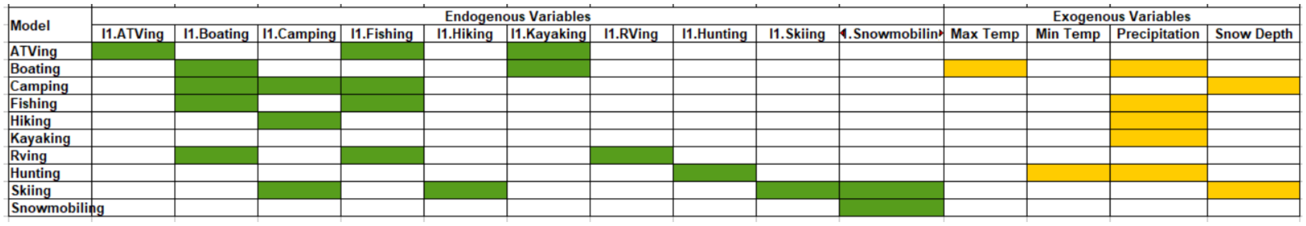
Another modification I employ compared to the class setup is that I consider 'exogenous' explanatory variables. These are four daily weather variables: precipitation, snow depth, maximum temperature, minimum temperature. These series are the average of the values taken from 105 different weather stations throughout Michigan. For more information on how these data series were created, please see installment 3:

https://dataandoutdoors.com/michigan-outdoor-recreation-search-interest-installment-iii/

I do not use raw values of the weather variables in the VAR model. Instead, I normalize the values by mean and standard deviation. The mean and standard deviation are not the overall values but rather taken from the same seven day period of each year for years 2021-2024. So each mean and standard deviation will be calculated over 28 values. The mean then will be subtracted from each number and then the result divided by the standard deviation. The resulting values show to what extent the weather indicator is low or high for that time of year. Also, for snow depth I do not divide by standard deviation because for most weeks of the year the standard deviation of snow depth is zero.

One advantage of the VAR model is you can technically estimate a single model for all ten stationary series. However, the disadvantage of this approach is that different explanatory variables are likely to be statistically significant for each submodel. Therefore, I instead estimate a completely different VAR model for each of the 10 outdoor recreation series. The table below shows which variables are included in each model, which all have a p-value of 0.10 or lower. Included endogenous variable are pictured in green and exogenous variables in orange.

All of the models are relatively parsimonious with at most 5 of the possible 14 variables included in a single submodel. However, exactly what variables are chosen shows some of the limitations of this model. Specifically, the weather variables all have the same coeficient or impact regardless of the time of year. For instance, for fishing you would expect a positive effect for temperature in June (as cold fronts reduce fish activity) and a negative effect for January (as cold weather brings more ice to walk on). In the end, temperature is not included for fishing at all. In fact, three models include no weather variables: ATVing, RVing, and Snowmobiling. It's highly likely that weather impacts these activities, but in some cases only a month or two a year. Further, these effects are likely to be nonlinear. For instance, the impact of snow depth on snowmobiling is likely much larger from 0 to 6 inches (which enables snowmobiling) than it is from 6 to 12 inches (where snowmobiling is already feasible).



Model	Endogenous Variables										Exogenous Variables			
	l1.ATVing	l1.Boating	l1.Camping	l1.Fishing	l1.Hiking	l1.Kayaking	l1.RVing	l1.Hunting	l1.Skiing	.Snowmobilin	Max Temp	Min Temp	Precipitation	Snow Depth
ATVing	🟩			🟩		🟩								
Boating		🟩				🟩					🟧		🟧	
Camping		🟩	🟩	🟩										🟧
Fishing		🟩		🟩									🟧	
Hiking			🟩										🟧	
Kayaking													🟧	
Rving		🟩		🟩			🟩							
Hunting								🟩				🟧	🟧	
Skiing			🟩		🟩				🟩	🟩				🟧
Snowmobiling										🟩				

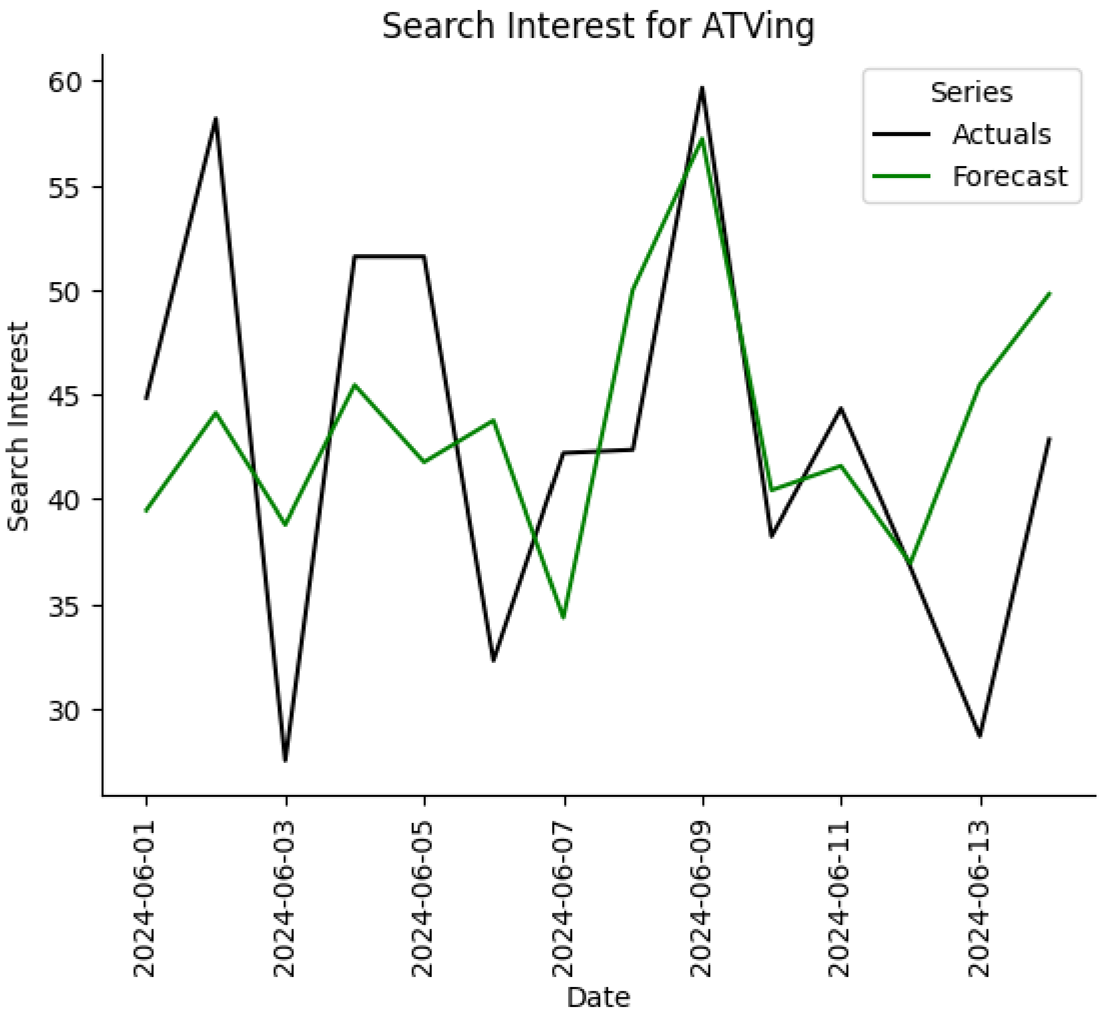
Selected Forecasts

The method above can be used to forecast forward. For instance, I can produce a 14 day ahead forecast. The only actual values for search interest used directly in the forecast will be prior to the forecast starting point. The forecast is made one month at a time where the nonstationary model is forecasted forward for all 10 variables and then this forecast is used as the starting point for the subsequent month's forecast. In each case, the previous 365 (denoised actual or forecasted) values of the same series are used in the nonstationary model forecast. This process is repeated 14 times for the 14 day forecast. Simultaneously, the previous month's stationary values are used to predict the next month's values. For the first forecast period, these will be the actual values. For the second period onward, the previous value with be the forecast from the previous period. However, the actual values of the weather variables will be used throughout.

Below I choose one forecast start point for each of the 10 forms of outdoor recreation. This is to illustrate the forecasting capability of the model for purpose of discussion. This is not intended to be a comprehensive evaluation of the model. A single forecast start point is not representative of the forecasting capability for a model over any unspecified time period. Further, the model was estimated on data on and after the forecast periods, which tends to inflate the forecasting performance of the model. (However, the fact that the model parameters were chosen based on statistical significance instead of cross validation in this case mitigates this issue.) Regardless, in the future when I have one or two additional models to compare I will have a more comprehensive method to evaluate the models.

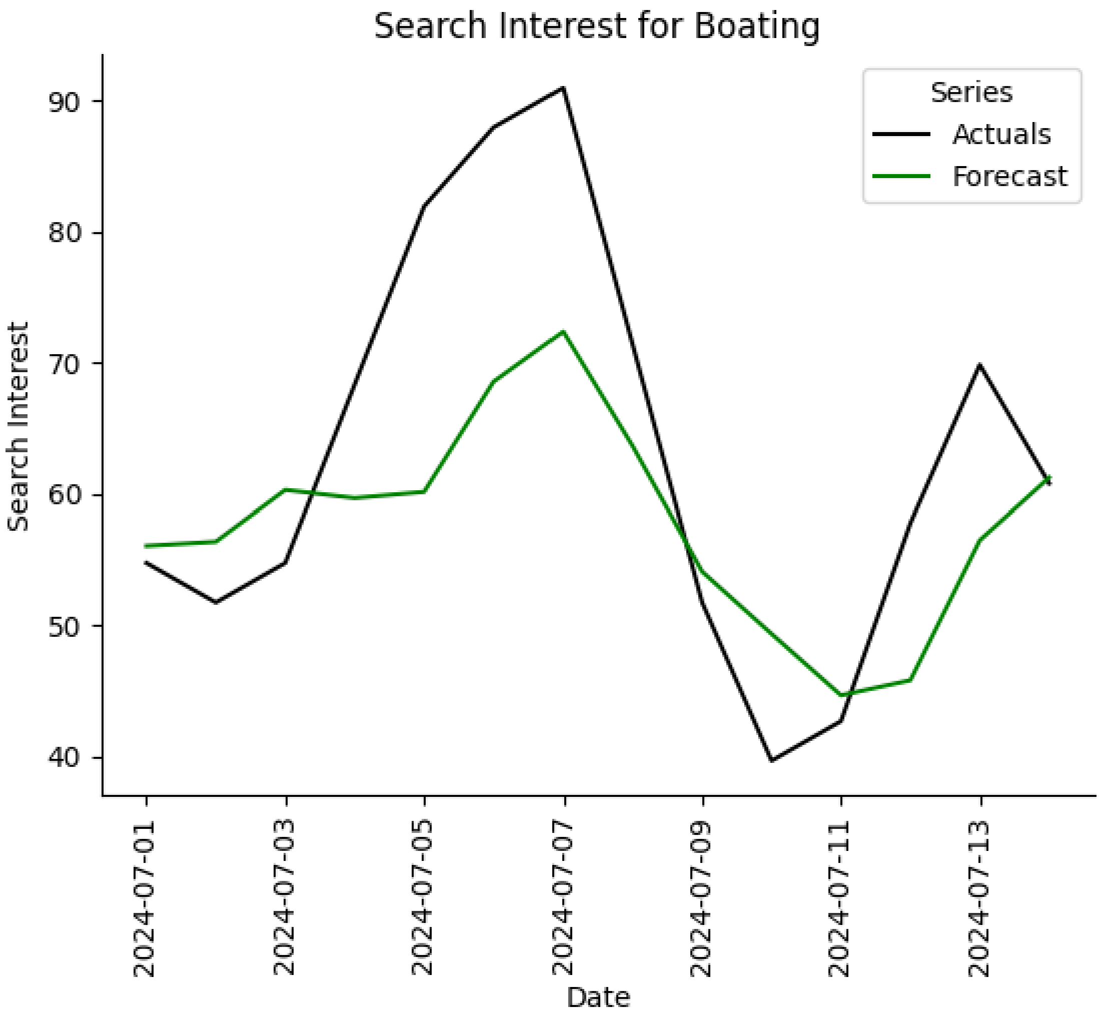
ATVing

Below I created a forecast for ATVing starting June 1, 2024 and compare this forecast to actuals. I chose the first half of June for the forecast because this is the first month that is relatively warm which is extremely popular with ATV enthusiasts. The actuals show quite a bit of fluctuation in the first week that isn't captured by the model. The model appears to fit actuals better during the second week of the forecast.



Boating

The forecast for boating is over the 4th of July weekend. This weekend is very popular for boating. The forecast captures the general weekly seasonality of the actuals but not the heightened impact of the 4th of July holiday. Note that one limitation of this model is that it doesn't explicitly account for holidays.

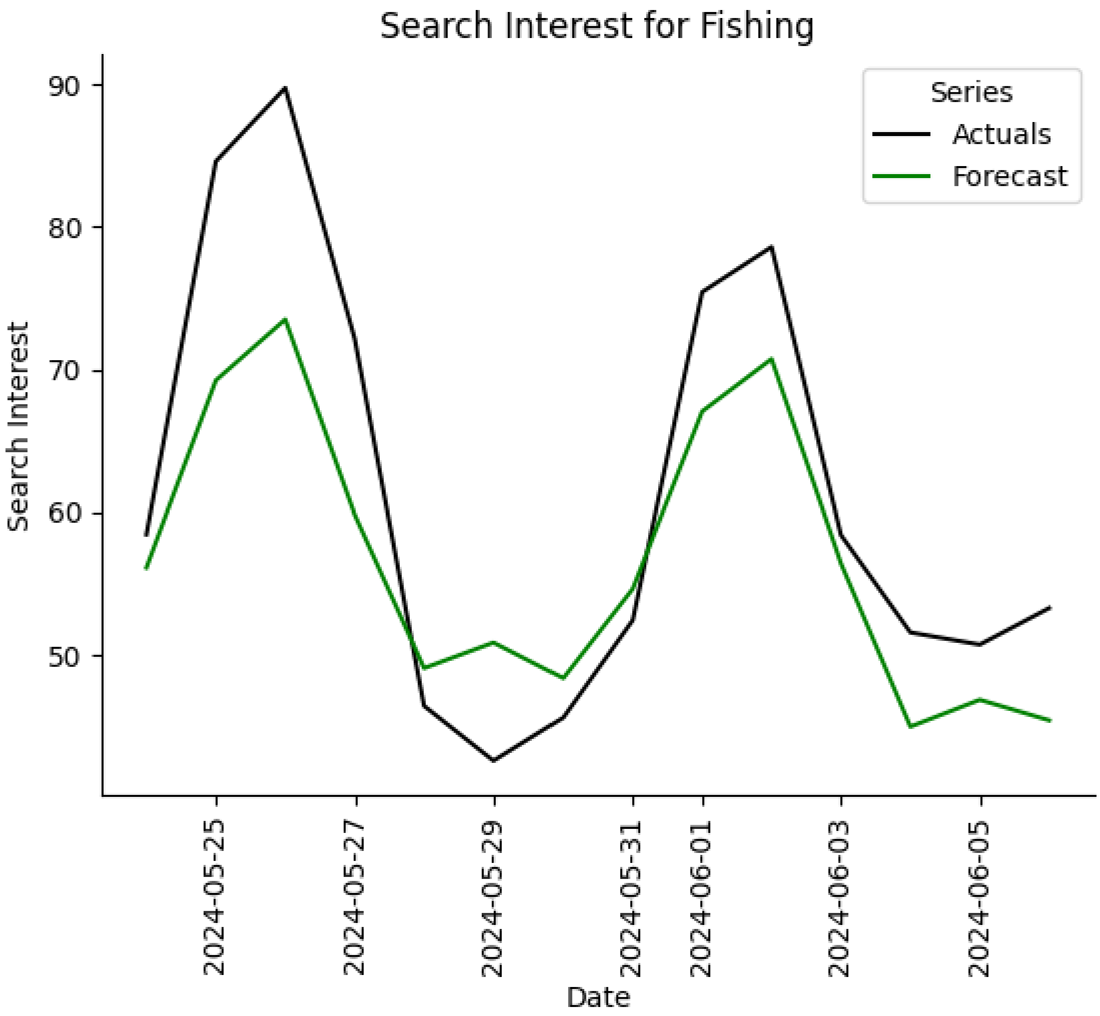


Camping

Unfortunately, for camping I wasn't able to produce a forecast with this methodology. As can be seen above, the hankel matrix for camping only has 12 components compared to at least 50 for the other series. This multicolinearity in the values has made it difficult to fit the nonstationary model. The model for camping will have to remain an area of future research.

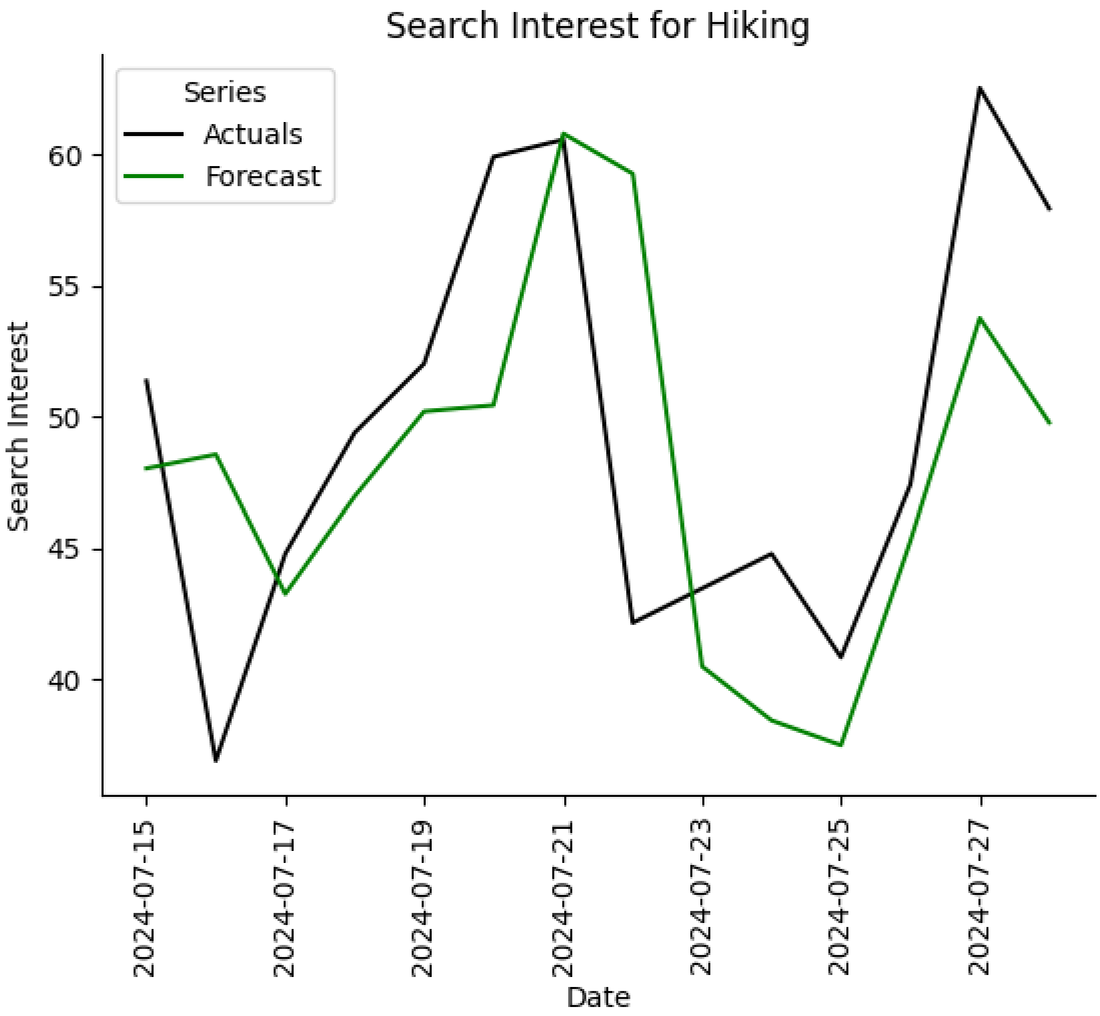
Fishing

I provide a forecasts for fishing over Memorial Day weekend. This is the opening weekend of bass season and popular for fishing anyways. The forecast captures the same basic pattern of actuals but not the amplitude. A limitation of this model is that it doesn't explicitly account for holidays.



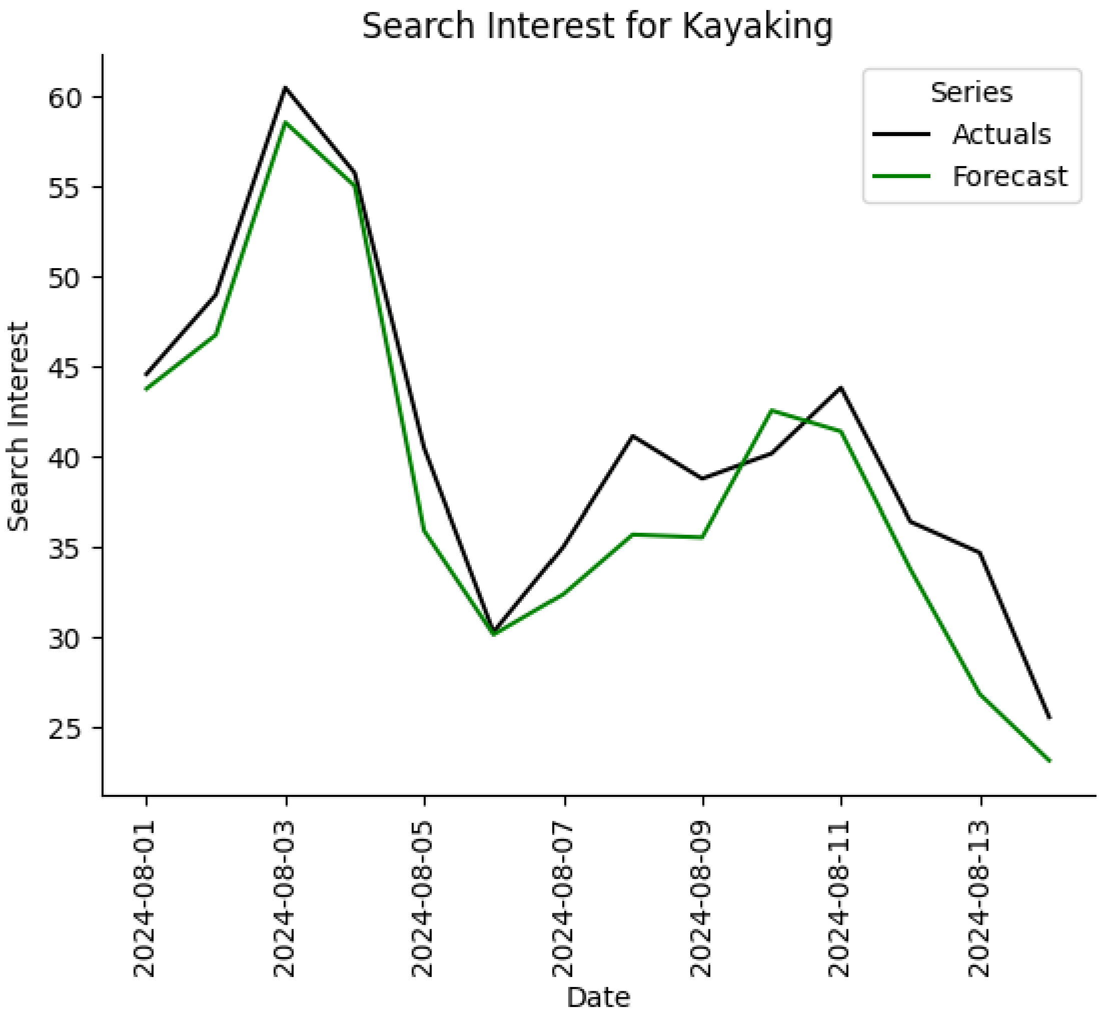
Hiking

I provide a forecast for hiking starting July 15. Hiking is popular during the summer. The forecast is generally very similar to actuals.



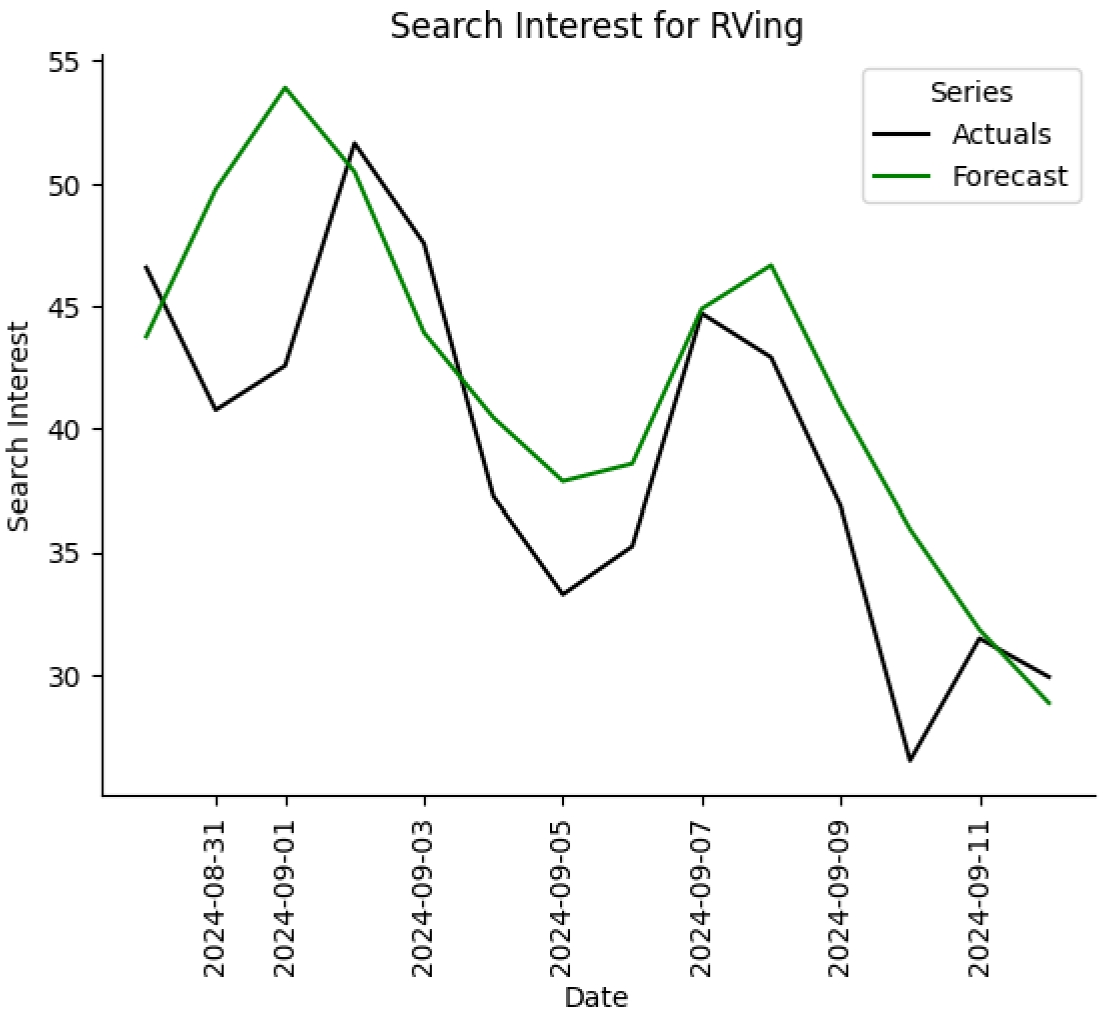
Kayaking

I provide a forecast for Kayaking starting August 1. August is popular for kayaking. The forecast is extremely close to actuals.



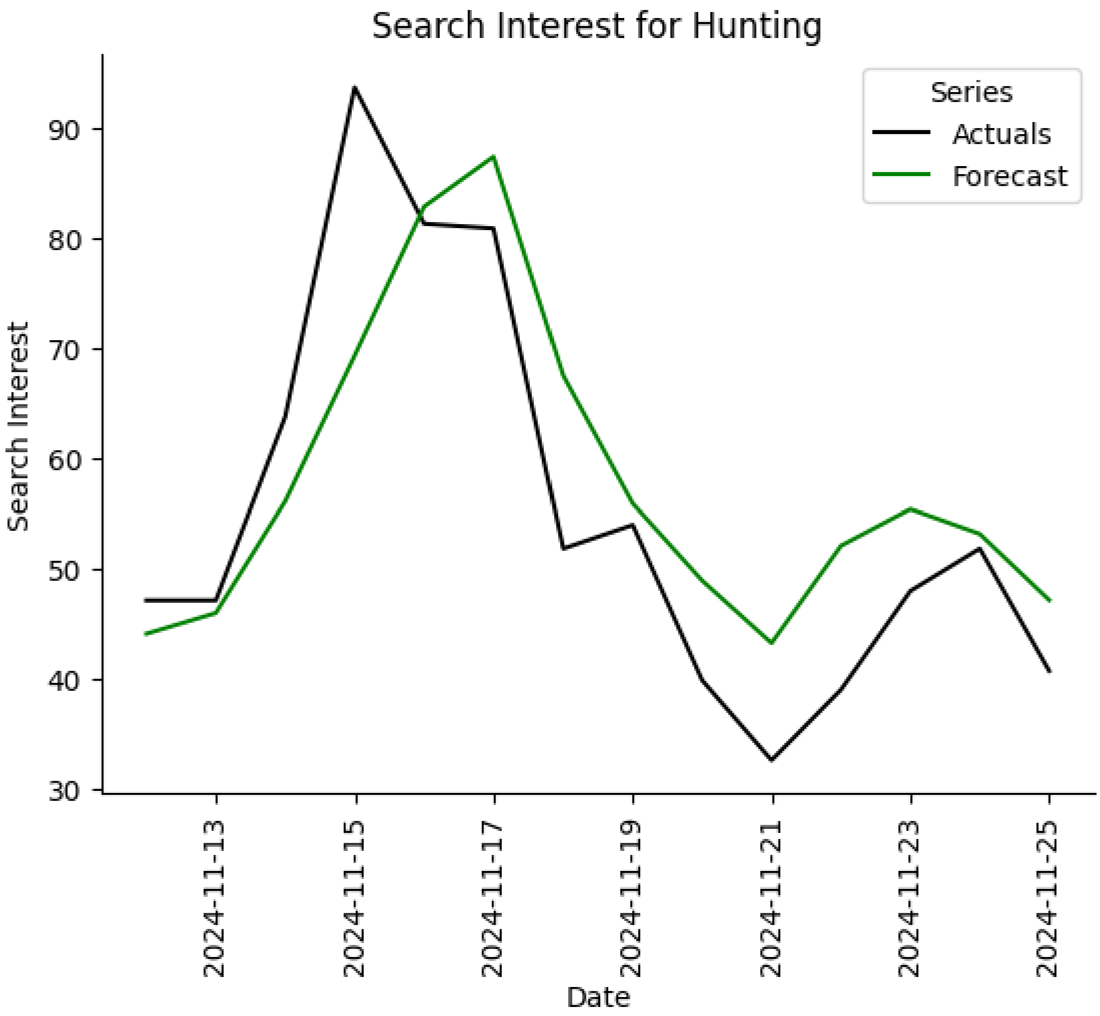
RVing

I provide a forecast for RVing over Labor Day weekend. RVing is popular over summer holiday weekends. The forecast for RVing is reasonably close to actuals. The actuals better reflect the impact of the holiday with a peak on Labor Day (September 2nd). The forecast, on the other hand, does not consider holidays and places the peak on the weekend instead (August 31/September 1st).



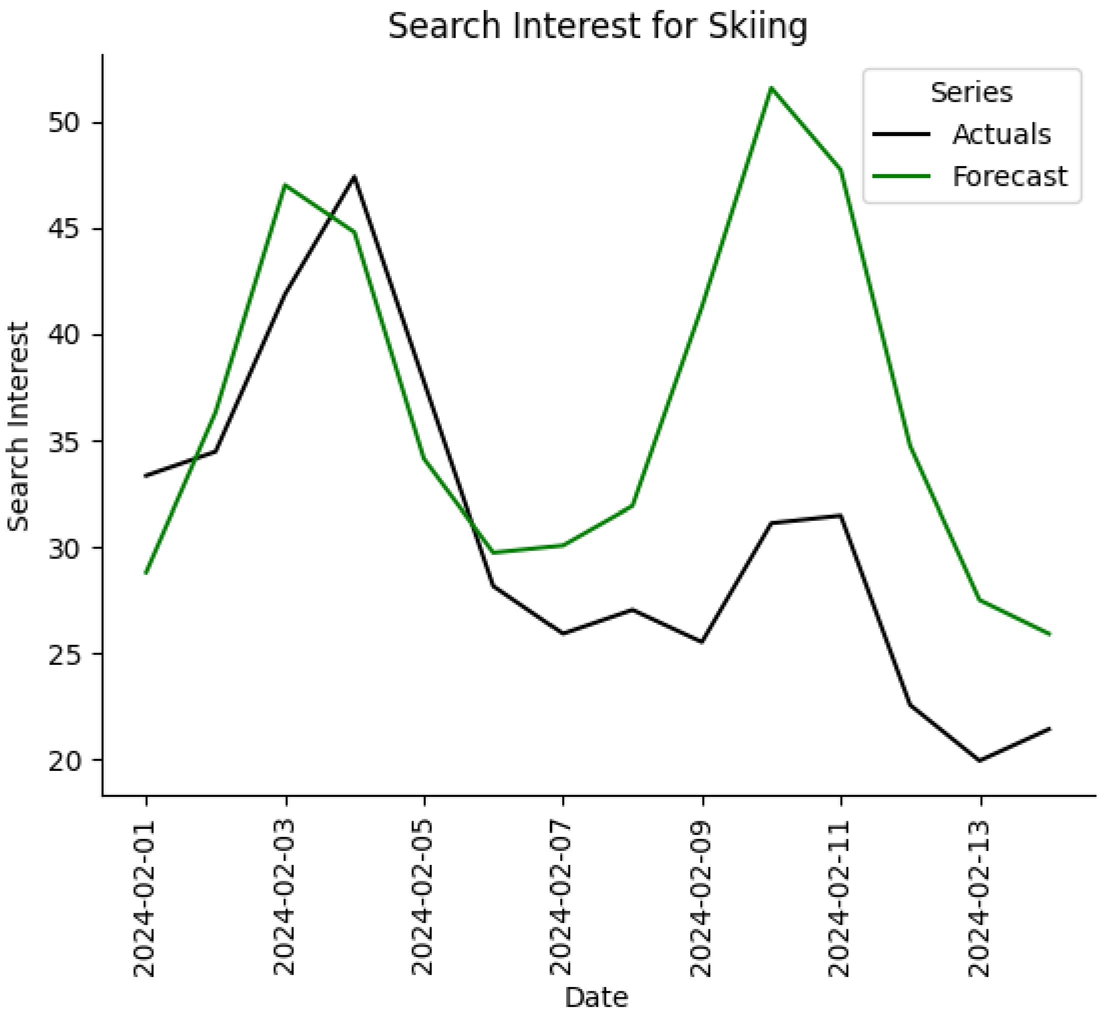
Hunting

I provide a forecast for hunting over November 15th, which is the opening of firearms deer season in Michigan. The forecast is close to actuals. The peak of the actuals better reflects the November 15th date. However, the forecast doesn't consider the signficance of this specific date, similar to holidays. Instead, it shows the highest search interest on the weekend (November 16/17).



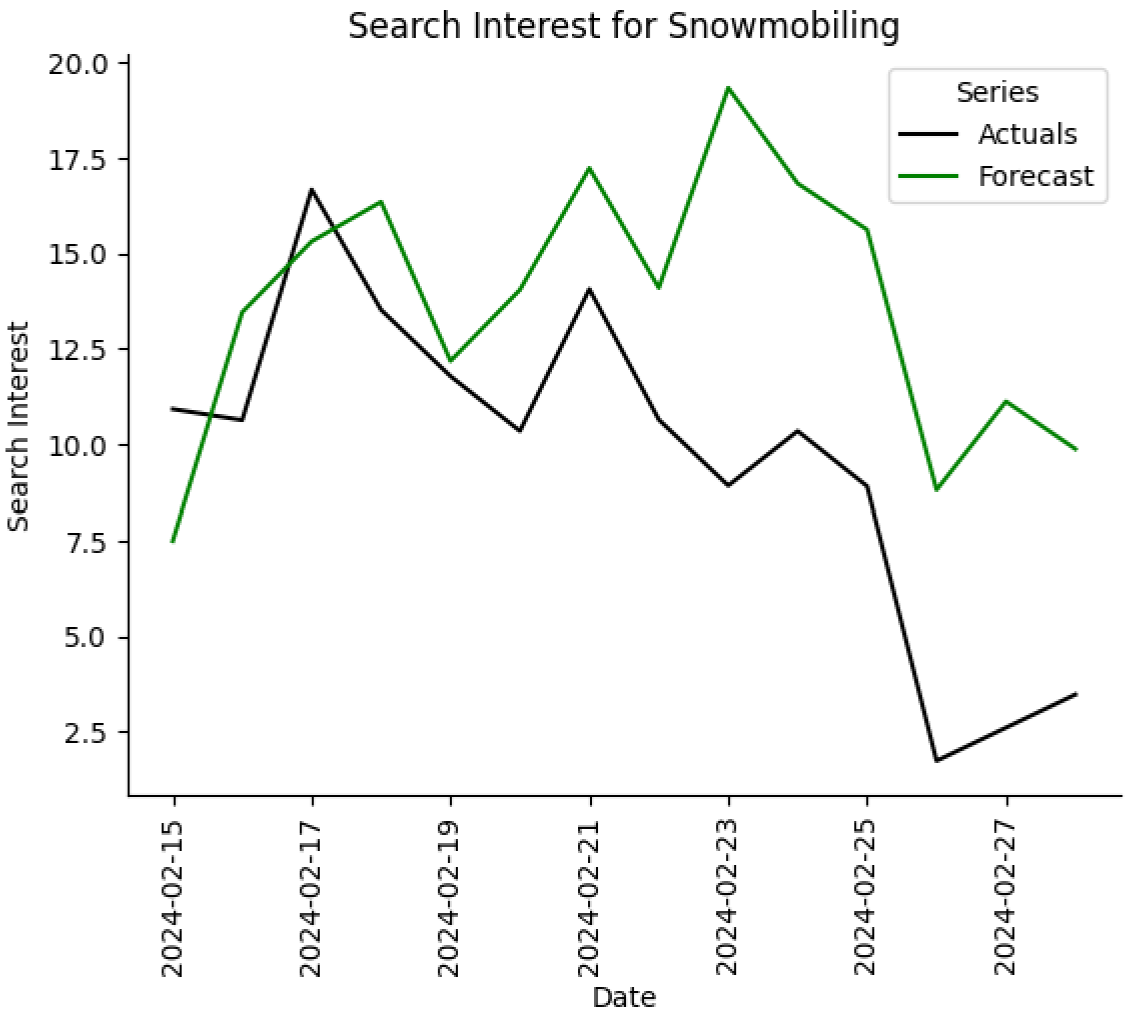
Skiing

I provide a forecast for skiing starting February 1st. February is the month in Michigan with greatest snow coverage. However, in 2024 there was very little snow coverage especially into February. Snow depth is considered in the skiing model. However, there is a single averaged effect for the entire year. Given that snow depth only has a real impact on skiing interest during a few months of the year, this no doubt dilutes the impact of the variable. Also, the impact of having no snow (as was often the case in 2024) is no doubt different than having marginally more snow in other years. Clearly, the model is not full capturing the impact of low snowfall in 2024.



Snowmobiling

I provide a forecast for snowmobiling starting on February 15th. Similar to skiing, the actual interest for snowmobiling was quite a bit less than the forecast. Quite a bit of Michigan was free of snow in February including areas popular with snowmobiling enthusiasts. Note that snow depth is not in the model for snowmobiling.



Conclusion

This is my first model option for forecasting search interest for ten forms of outdoor recreation in Michigan. I plan to complete one or two other options. This option accounts for the sigificant measurement error in Google trends data, trends in the series, seasonality in the series, and the effects of weather variables. The limitations of this model is that the impact of weather is not allowed to vary during different times of year, times of the week, or on holidays. Also, this model doesn't capture the impact of holidays. Finally, this method has not worked for the camping series.