Michigan Outdoor Recreation Search Interest Installment 9

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/

So far, I have attempted three different models to forecast this data. Of these, I've selected the first and third for further analysis.

The first method employs more traditional statistics. I decompose the series into a nonstationary (trend and seasonality) and stationary (residual) series utilizing a Hankel matrix. I then employ a Vector Auto Regressive (VAR) model on the stationary series. More information on this model can be found in Installment 6 at the following link.

https://dataandoutdoors.com/michigan-outdoor-recreation-search-interest-installment-6/

The other method employs a Long Short Term Memory (LSTM) Model a form of recurrent neural network that considers the sequence of the time series and also a constantly adjusted state that changes the impact of the input series on the forecast. More information on this model can be found in Installment 7 at the following link.

https://dataandoutdoors.com/michigan-outdoor-recreation-installment-7/

This installment will compare cross validation results between these two models for each of the 10 series. The final model will be chosen based on this cross validation. It's important to understand that developing each of these models involved making inumerable decisions that could have changed the outcome of the analysis. Therefore, someone else developing either of these or another model or evaluting them over a different time period may have achieved different results.

Training Data

To evaluate the models I choose to use a hold out time period that is later than the training period. Just as with Installment 7, I choose my training data to be January 2021-March 2024. This leaves the final year of the data for cross validation.

Due to computational concerns with the LSTM models, all models will only be trained once on the full training data January 2021-March 2024. In the past, I have employed a more complex method where the model is trained many times based on a sliding cross validation window. An example of this other method can be seen in my Northern Michigan Search Interest project at the link below.

https://dataandoutdoors.com/sample-page/

For the LSTM model, this is the same training data I used to choose the hyperparameters of the model. Therefore, I retrain the models using the same model specifications described in Installment 7.

For the Hankel/VAR model, I originally trained the model on the entire dataset, as is typical of statistical modeling. When I re-estimated these models on the reduced training data, some things changed. I did not worry about loss of statistical significance. However, as we seen for 'camping' in Installment 6, sometimes this model fails to provide reasonable forecasts. Hereafter, I will refer to this as a diverging forecast. Since the VAR model uses lags of other data series in the estimation of the model, I would not want to use the lags of data series with diverging forecasts in another model. Therefore, I adjust the specifications of four of the VAR models in the following ways:

1. Remove lagged fishing from the atving and rving model

2. Remove lagged camping from the hiking and skiing model

Cross Validation Data

There are many ways to perform cross validation in time series data. Generally, the cross validation data should be data the model is not trained on. Furthermore, the cross validation data should occur after the training data in the series.

I choose my cross validation data from April 2024-March 2025, which is after the training data. However, given the purpose of the model is to perform 14 day forecasts, I wouldn't want to create a single year long forecast to evaluate the model.

I decided to perform cross validation on multiple 14 day forecasts. This is possible by using a sliding 14 day window. As discussed in the training data section, each sliding 14 day window will be forecasted using the exact same models estimated on the January 2021-March 2024 training data.

For computational reasons, I decided not to slide this window one day at a time. Instead, I chose to slide the window seven days at a time, to coincide with weekly updates to the forecast. In other words, I will start one 14 day forecast, then another a week later, then another a week later, then another a week later. Each forecast will overlap the one before and after it by a week. The forecast will be compared to actuals and used to calculate a Mean Squared Error (MSE) for each time window.

The first sliding window is April 1-14 2024. The final window is March 17-30 2025. For each model and series, the MSE from each of these windows is averaged into a single MSE.

This is the same method employed in Installment 7, which can be visited for more details.

Results

The table below shows the results from the cross validation. The series column lists the various data series. The LSTM column lists the MSEs for the LSTM model. The VAR column lists the MSEs for the Hankel/VAR model.

If the MSE is over 200, then I list the model as having a diverging forecast. This occurs for the Hankel/VAR model because sometimes the nonstationary model does not produce a reasonable forecast. This only occured for camping when using the full data, but it also occurs for fishing, rving, and hunting when using the reduced training data.

If one were determined to use the more traditional statistical model, you could try different specifications for this model. For instance, I use a large number of lags (365) in the nonstationary model in an attempt to capture seasonality. Often, this was successful but it also leads to computational challenges related to multicollinearity and matrix inversion. Even when shortening the number of lags for the camping model, I struggled to get a working model. However, there are many other ways to create a nonstationary model than the Hankel matrix method employed here.

Based on the table below, the best version depends on the series. Whichever model leads to the lowest MSE is chosen as I didn't have a reason to override any of the results. In many cases when the MSE were close, the winner was my preferred option anyways. The LSTM model is chosen for camping, fishing, hiking, rving, hunting, skiing, and snowmobiling. The Hankel/VAR model is chosen for atving, boating, and kayking.

It's important to note that while these sorts of comparisons between different models are interesting and popular, often results can be driven as much by individual modeling decisions for each model. During the original installments for these models, I investigated and made several different decisions. However, there are inumerable possible iterations of these models which means that different approaches or evaluation over different time periods could easily change outcomes.

	Series	LSTM	VAR	Chosen
0	Atving	142.7	90.0	VAR
1	Boating	17.1	17.0	VAR
2	Camping	17.2	Diverges	LSTM
3	Fishing	38.2	Diverges	LSTM
4	Hiking	31.7	100.6	LSTM
5	Kayaking	26.3	11.3	VAR
6	Rving	22.9	Diverges	LSTM
7	Hunting	20.1	Diverges	LSTM
8	Skiing	23.3	24.6	LSTM
9	Snowmobiling	6.0	6.3	LSTM

Conclusions

This was a cross validation comparison between two very different models for estimating search interest for outdoor recreation in Michigan. One model is a more traditional model that involves time series decomposition and statistical models. The other utilizes a relatively esoteric form of a neural network.

These types of comparisons of disparate modeling methods are popular and tend to highlight differences between popular approaches while providing opportunity for education and learning. Sometimes, these comparisons can be ideologically driven as some will try to prove a point about the benefits of one method or the problems with another.

However, frequently results are driven as much by individual modeling decisions within each model than the basic model type. A skilled modeler can usually get a satisfactory results with various types of models. I myself dedicated some time to developing these models, given limitations of the free time I wanted to devote to them. However, there's obviously all forms of improvements I could continue to make.

In this case, the LSTM model was chosen 7 out of 10 times. However, for four of the series, the Hankel/VAR method didn't even have a working model. Had I devoted the time to correct this issue, the end result would have likely been more even. In fact, of the six series the Hankel/VAR model could have been chosen it won three times.

Regardless, I feel I have a working model for all ten series I can use going forward. My next step will be to discuss sourcing weather forecasts from accuweather.