Michigan Outdoor Recreation Search Interest: Fourth Installment

The purpose of this installment is to determine whether the search interest data introduced during the first two installments and the weather data introduced during the third installment are correlated. The general idea is that people will search for outdoor recreation search terms differently in some weather conditions than others.

As in most cases of time dependent data, even something as simple as determining correlations can be a challenge. Frequently, current results are based not only on current values, but values in the past. And, since these values are also correlated across time, correlation with past values are not necessarily direct. These concepts are known as autocorrelation and partial autocorrelation.

The other issue relates to trend and seasonality concerns. As I've shown in previous installments, both outdoor recreation search interest and weather are extremely seasonal. Therefore, it's possible that they are spuriously correlated such that they are both actually being driven by the seasons not each other. However, in this case it's quite certain that weather patterns in different seasons are a primary driver behind seasonal variation in outdoor recreation. Some also exhibit a trend which can also lead to spurious correlation, i.e. both indicators just happen to be trending accross time without any actual relationship.

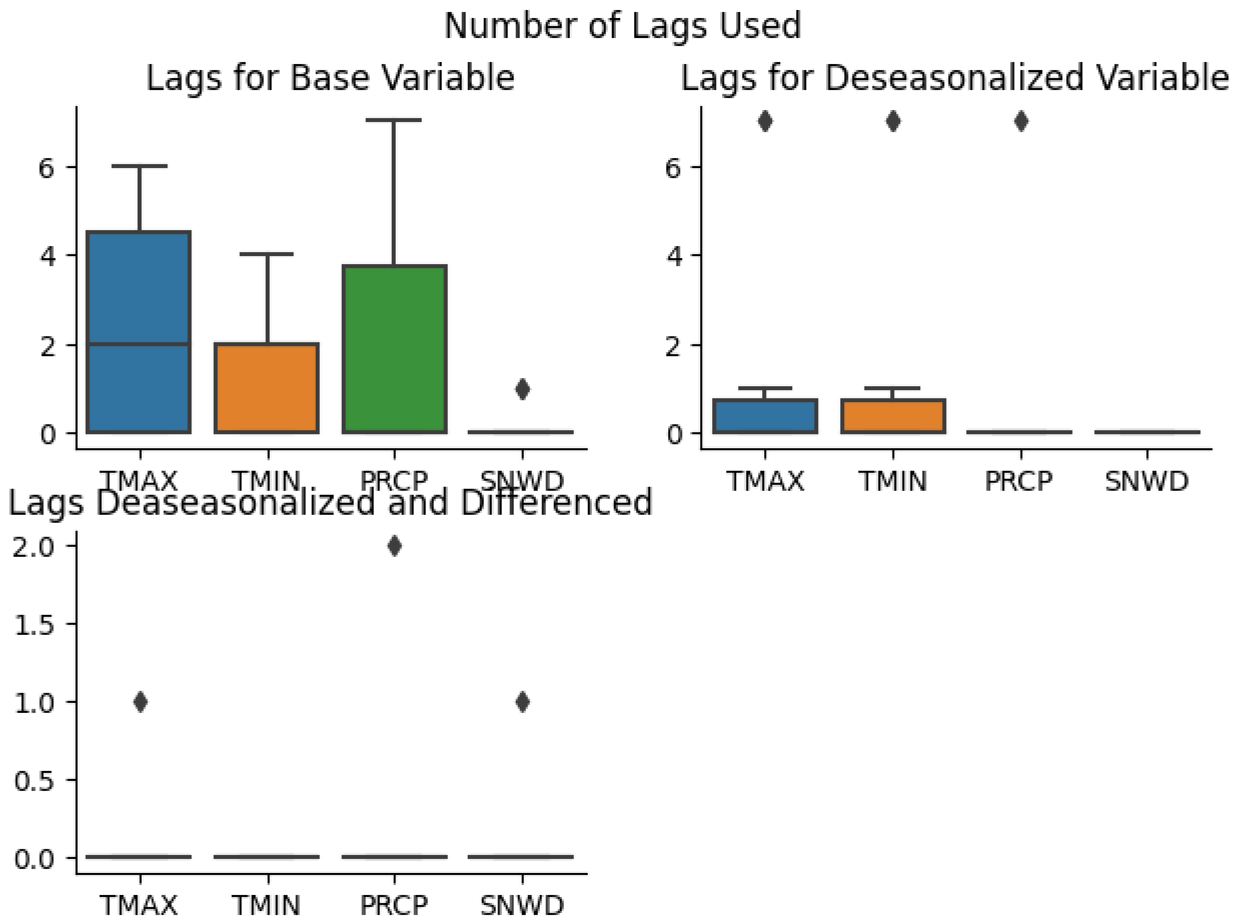
Note there are 10 search interest keywords: atving, boating, camping, fishing, hiking, kayaking, rving, hunting, skiing, and snowmobiling. There are four weather variables: maximum temperature, minimum temperature, precipitation, and snow depth.

The method I employ is to run linear regressions of each of the 10 search interest terms on each of the four weather variables. I run eight regressions for each of these 40 pairings. The first regression will be a regression of search interest on weather of the same day. The second regression will be a regression of search interest on current weather and the first lag (the previous days weather). The third regression will be a regression of search interest on the current weather, the first lag, and the second lag. I keep adding lags until I reach seven lags. Of these eight models, I choose the model with the best (lowest) Baysian Information Criteria (BIC). A full discussion of BIC is beyond the scope of this writing; however, this criteria rewards better model fit while penalizing the addition of more lags.

In this manner, I achieve the optimal number of lags for each of the 40 pairings between search interest and weather variables. However, I must also consider stationarity concerns. I choose these models using data transformed in three different ways. The first is the level (untransformed) values. The next approach I employ is that I use deseasonalized data. The final approach uses data that is both deseasonalized and differenced. This approach is most 'pure' from the statistical point of view. However, period to period noise and data errors tend to predominate the results.

In the figure below, I have boxplots for the number of lags used for the four different weather variables depending on whether the base values, deseasonalized values, or the deasonalize and differenced values are used. Even when using the base values, the median number of lags is zero for all but maximum temperature for which the median number of lags is two. In other words, in most cases, only the current value of weather is used and not any past values.

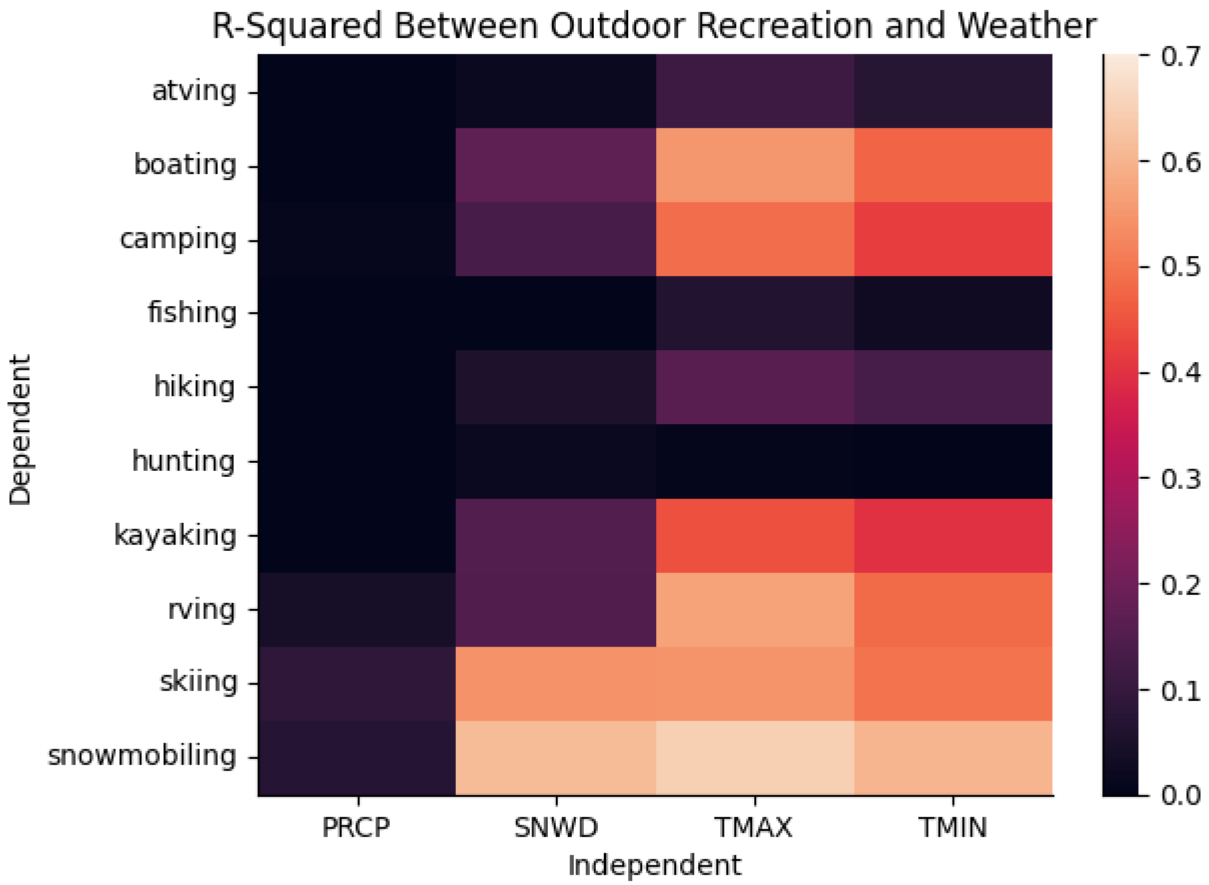
For the deseasonalized values, almost all of the regressions use at most one lag. In the case of deseasonalized and differenced values, almost all of the lags are zero and at most two.



Base Values

Below is the r-squared between various outdoor recreation search terms and the weather variables. Unsurprisingly, many of these r-squared values are fairly high given that both the search terms and weather are seasonal. R-squared for precipitation is generally low. Precipitation levels are not extremely seasonal, though variation for precipitation is generally higher in the winter. Also, r-squared for atving, fishing, hunting, and, to a lesser degree, hiking is generally low. These activities generally occur during more than one season.

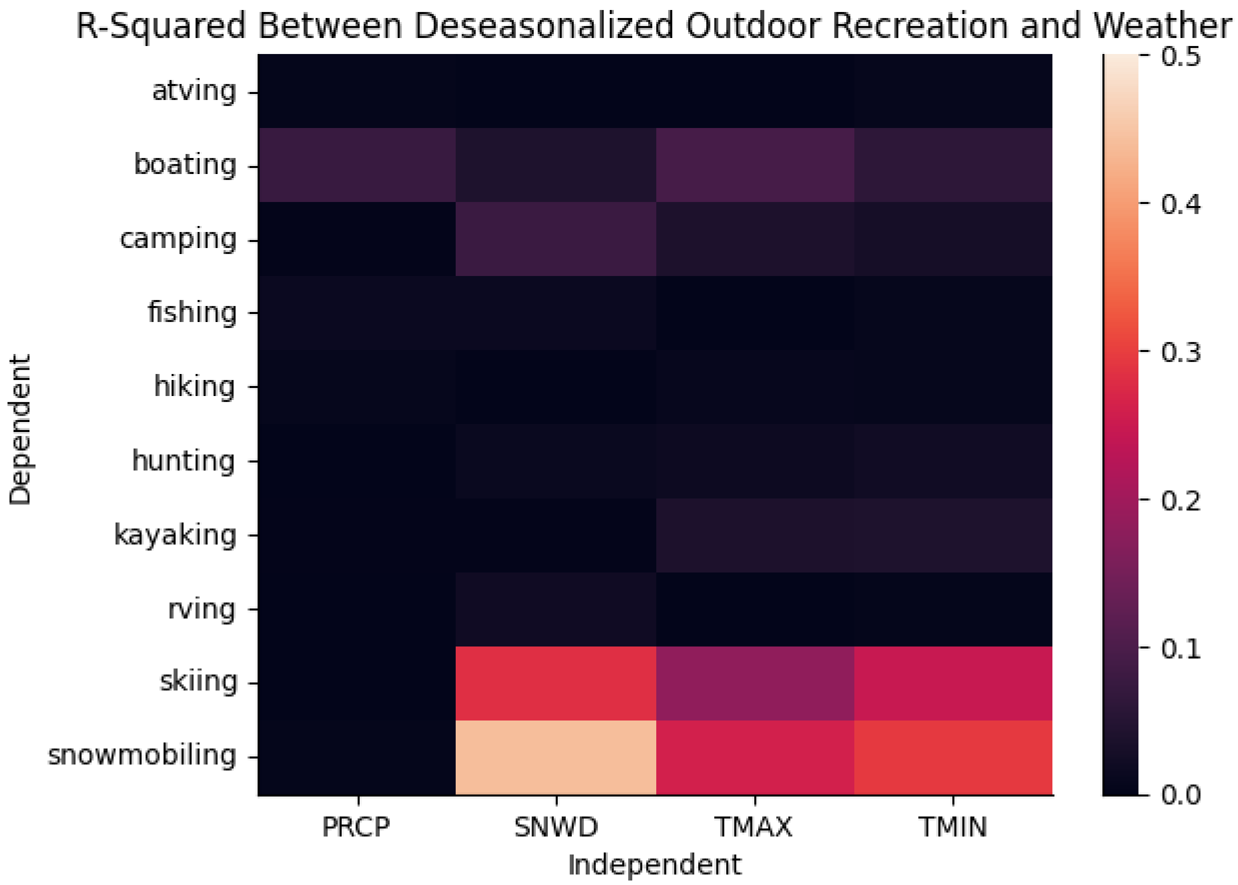
The highest correlations for summer sports are boating, camping, kayaking, and rving and these activities are correlated with temperature. Snow sports, skiing and snowmobiling are correlated with both temperature and snow depth.



Deseasonalized Values

One issue with using the base values is they are not stationary. All of the search interest and weather variables are highly seasonal. In general, seasonal variables will be spuriously correlated. In the case of outdoor recreation, however, there's little doubt that seasonal variation is mostly due to weather. Nonetheless, it's interesting to look at whether there is a change in search interest based on unseasonal weather. In other words, is search interest relatively high for a certain time of year when the weather is relatively different for that time fo year?

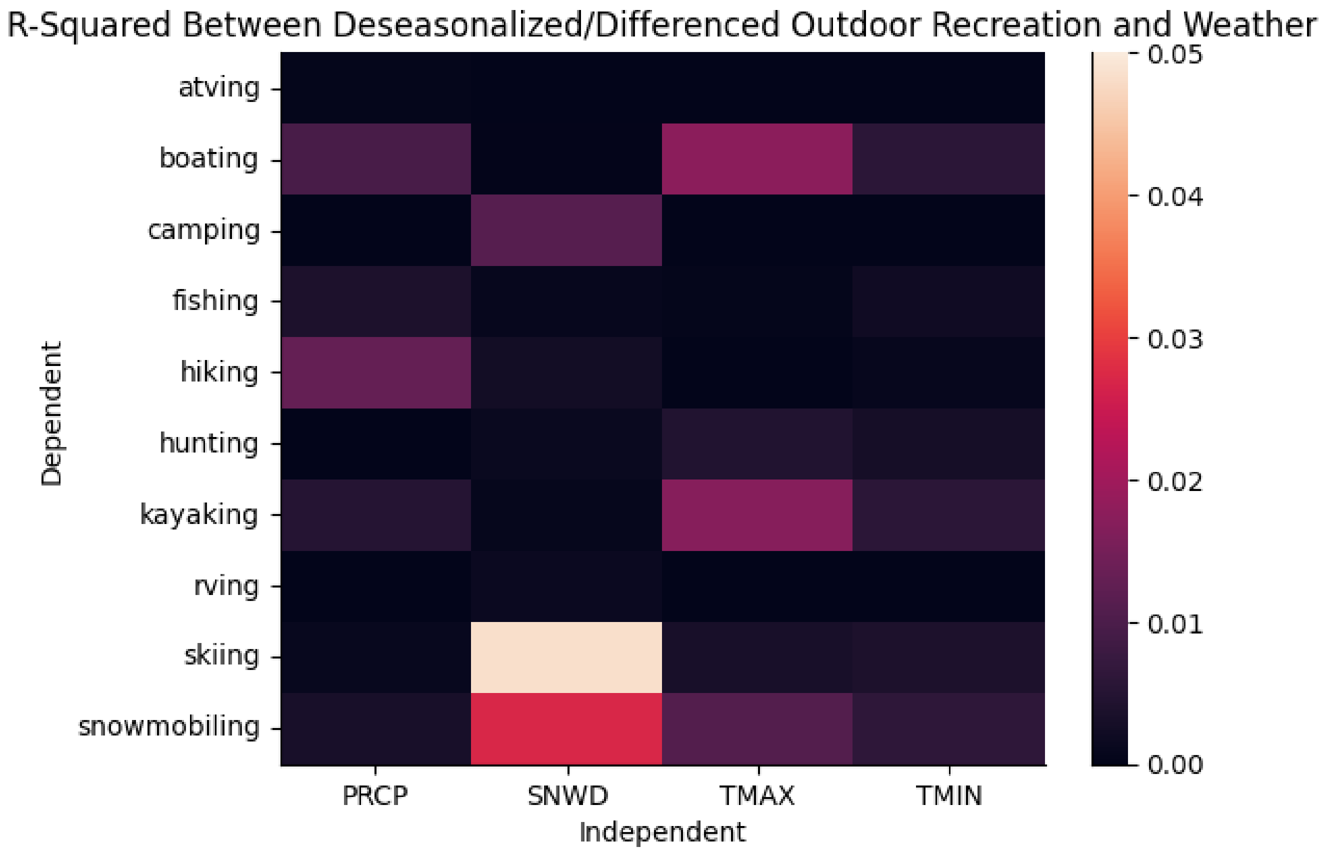
In general, the only search interest variables that still have fairly high levels of correlation are winter sports: skiing and snowmobiling. However, for boating and camping there is some impact from weather data, under 0.1 r-squared.



Deseasonalized and Differenced Values

Even after deseasonalizing the values, there may be long term trends leading to spurious correlation between different variables. The standard method for accounting for this issue is by differencing the data by one period. After differencing the data, the r-squared is at most 0.05 for each weather variable. The highest impact is from snow depth on skiing and snowmobiling.

It's not entirely suprising that these r-squared values are very low. In order for the weather variables to explain search interest in this case, then changes in weather from the previous day would have to explain changes in search interest from the previous day. However, it seems less likely that one would search for a term simply because the weather is different than yesterday, even if there is an impact in general. Also, google trends data suffers from a fair amount of measurement error. Differencing the data tends to increase the percentage of variation due to measurement error and other random factors.



Conclusions

One of the greatest drivers of search interest for outdoor recreation is seasonality, largely due to weather. However, it's generally to a smaller degree that unseasonal weather patterns impact unseasonal search interest outside of recurring seasonality. Finally, changes in search interest from the previous day are generally not explained by changes in weather from the previous day.

While the general ideas concerning the relationship between weather and search interest are described above, there are obviously ways to improve upon this framework. For instance, in many cases these impacts are seasonal in themselves. Changes in temperature during the middle of winter are unlikely to impact search interest for boating. Changes in temperature are likely to have the opposite effect for summer fisherman vs ice fisherman. Aggregating the data for the entire year dilutes the results we see.

We also see a tradeoff between concerns such as stationarity and capturing the impact of weather variables. Differencing is fairly aggressive in removing trend and stationarity issues while also eliminating variation of interest. Due to the relatively short time frame of the sample, I would consider not detrending the data or detrending the data differently due to the strong possibility that changes in search interest are actually caused by weather not serial correlation. However, some forms of search interest no doubt are trending due to impacts of COVID-19 on interest in outdoor recreation.