Spatial and Temporal correlations in spore detection data

An analysis of 2022 field data for Root Applied Sciences

Sondre Skatter, Dec x 2022

# introduction

With the data from the 2022 installed base Root Applied Sciences would like to understand better how densely to apply the sensors and how often to sample.

# Material and Methods

The 2022 data contained 2752 rows, and after removing entries with ‘days\_sampled‘ either ‘NA’ or 0, then 2371 records remained. The statistic used for this study was:

Y = log10(spores / days\_sampled + 1)

Where ‘spores’ is the estimated raw count of spores from the analysis. This figure was normalized by the days sampled in order to normalize to the volume of air sampled. The log10 was taken to get more consistent results (I tried with an without) and the + 1 so that 0 still is 0 after the transformation.

Because the goal is to understand how densely to sample, both in terms of space and time, the autocorrelation function (ACF) was estimated. Essentially the ACF is the correlation coefficient for our statistic over a distance (either in time or space) as given by the x-axis. So for spatial ACF we would be looking for a drop in the ACF when increasing the, and such a drop would guide us to how densely to place the sensors.

Often when the ACF is computed there is a regular sampling interval, for example in time series data. In this data this is not the case, and therefore I use a couple of computational tricks to get a smoother ACF.

## Computing temporal ACF

To compute the temporal ACF we pair up data points with the same sensor sampled at different times. We bin the time into steps of 1 day so that we have one value for the ACF for 1 day, one for 2 days, etc. Because each measurement is sampled over a range of days, then any measurement *pair* is contributing to several bins. For example if measurement A was May 8 – May 10 and measurement B was May 14 – May 15, then this pair will contribute to bins 4, 5, 6, 7.

For any given bin, we’ll have *n* pair (say A and B for the pair) of measurements, and we will take all A measurements: YA, and all B measurements, YB, and compute the ACF for the bin, *i*, by:

## Computing spatial ACF

To form measurement pairs to compute spatial ACF we limit to measurement pairs of different sensors that overlap in time. The time overlap condition was defined by computing intersection-over-union (IoU) for the sampling time for each measurement pairs. Only pairs with IoU overlap above 0.5 were included in the analysis.

Because the spatial sampling for the ACF was much sparser and more irregular, I used a kernel approach. This means that each measurement pair that has a distance between them will contribute to the statistics via a Gaussian kernel that spreads the point over multiple bins. The sigma for the Gaussian was taken to be 0.15 x the distance. The spatial ACF was binned in steps of 10 meters up to 2km.

With the kernels acting as weights we then first compute the averages and in each bin, *i*:

# Results

## Data overview