Spatial and Temporal correlations in spore detection data

An analysis of 2022 field data for Root Applied Sciences

Sondre Skatter, Dec 11 2022

# introduction

With the data from the 2022 installed base Root Applied Sciences would like to understand better how densely to install 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 to normalize to the volume of air sampled. The log10 was taken to get more consistent results (I tried with and without) and the + 1 so that 0 remains 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 plotted as a function of a distance (either in time or space). For the spatial ACF we would be looking for a drop in the ACF with increasing distances, 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 day bins 4, 5, 6, 7.

For any given bin, we’ll have *n* pairs (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 ourselves 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 kernel 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*:

Then the temporal ACF:

# Results

## Data overview

A temporal overview of the data is shown below in Figure 1 with each sensor forming a row in the image and. It shows the general picture of the growing season as three phases: 1) early spring: no spores detected, 2) mid season: both detections and clears in a salt-and-pepper fashion, 3) late growth season: predominantly, spores are detected everywhere.

Chart

Description automatically generated

Figure 1. Lab test results throughout the growing season. Each row represents a sensor. Red means there was at least one positive test in time period, green that there was only negative results, and gray that no tests were performed in the time period.

For the spatial ACF , we would expect a high level of correlation even at longer distances for zone 1 where most sensors are negative and zone 3 where most sensors are positive. In the first case there are no spores anywhere and in the latter they are everywhere. So in order to answer the question of how densely to place the sensors, the mid zone will be most helpful.

## Temporal ACF

The temporal ACF is shown in Figure 2. We see there is a positive correlation for the first 40 days, then it switches to negative. This switch comes about because an increasing amount of the data for the longer lags are between Zone 1 and Zone 3, which means the opposite state is likely.

Chart, line chart

Description automatically generated

Figure . Temporal autocorrelation function (ACF)

Apart from this, it is worth noting that the correlation is rather low, even at close time periods. This is illustrated also by the noisiness (salt and pepper noise) seen in Figure 1. Presumably, this is to be interpreted as the spore density fluctuating over time, perhaps as a function of the weather. However, another plausible interpretation would be that the repeatability of the measurement is low.

## Spatial ACF

The spatial ACF was computed

1) with all the data (Figure 3), and

2) with only the data from the mid zone, between April 15 to July 15 (Figure 4).

As expected, when all the data is included the ACF stays high even for large distances. We expect this because in the beginning all the sensors report zero level spores, and in the end of the season most sensors detect spores. When including only the mid zone, the correlation for longer distances falls from around 0.4 to less than 0.2.

Chart, line chart

Description automatically generated

Figure . Spatial ACF including all the data.

Chart, line chart

Description automatically generated

Figure . Spatial ACF with only data from Apr 15 – Jul 15

In both cases there seem to be high correlation for very small distances, which drops off to the asymptotic value by the time distances of 250 meters is reached. It appears that most of the drop happens in the first 75, however, as can be seen from the red curve, there is very little data for inter-sensor distances less than 100 meters so we can’t identify a transition point from this data.

To put that in perspective in terms of sensor density, if placing the sensors on a grid where every other row is offset by half a distance (like a 5 on a die), then 1 sensor per acre corresponds to inter-sensor spacings of 72 meters.

# Conclusions and further work

More analysis should be done on the nature and statistics of the measurements. Understanding the uncertainty in each measurement would help disambiguate how much differences between neighboring sensors, or same sensors over short time periods, is due to the measurement itself or the actual spore density. In order to do this, we will need to get the detailed readings for each measurement, i.e. for each cell in the assay the red/green value.

For the next growing season it would be helpful to have more variation in the inter-sensor distance, in particular in the range from 40-250 meters.