Spatial covariance

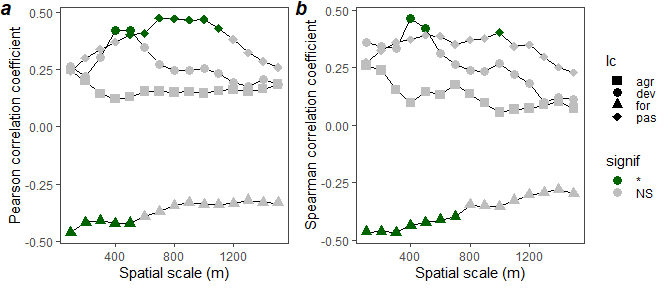
Kevin Li

12/20/2021

## Pearson vs. Spearman correlation revisited

I revisited the correlation tests over different landscape scales. Mainly I wanted to see if transforming the data before performing the Pearson correlation would help clarify some patterns. Avelino also transformed the data by taking the Arcsine square-root of the proportion of infected leaves.

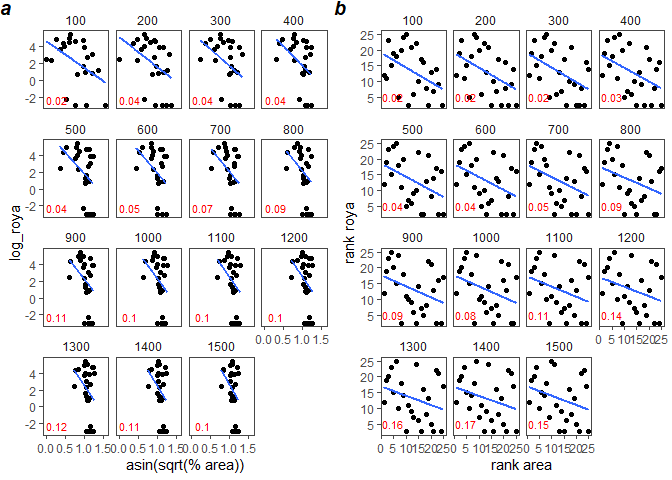
Here, I am taking the natural log of the maximum infected leaf counts, since we can’t calculate a proportion like Avelino. In cases where there were 0 leaves, I added 0.05 first before taking the log. This is because the lowest non-zero farm maximum was 0.1 (averaged over all plants), so I’m taking half of that as “zero.” I am also transforming the landscape variable using Arcsine square-root, because this transformation is appropriate for proportions.



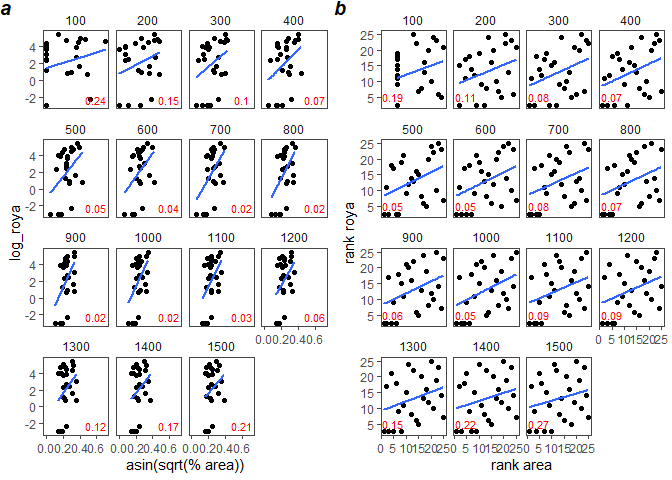
**Figure 1** Comparision between Pearson’s correlation coefficients of transformed variables (log(roya)) and arcsin(sqrt(% area)) **(a)** and Spearman’s correlation coefficients of original variables (# roya leaves and % area) **(b)**, for a range of spatial scales.

The results for the Pearson correlation of the transformed variables match better with the Spearman results than before when we did it on the untransformed data. I guess this makes sense since the data are now closer to normal. For some reason, I got different significant p-values for the Spearman’s correlation tests than Zach, but the shape is similar. I’m not sure if it’s because we’re using slightly different cutoffs, or datasets.

From what I’ve read, the Pearson correlation coefficient is appropriate for data you expect to be linearly related, while the Spearman is appropriate for data that you expect to be monotonically related. Spearman is less sensitive to outliers, though. In Figures 2 and 3, I plot the data that is being used to calculate the correlation - the transformed variables for the new Pearson values and the ranks for Spearman. Maybe there’s some reason for concern about the formerly 0 roya values and the very low roya farm. They seem to be outliers in the Pearson plots.

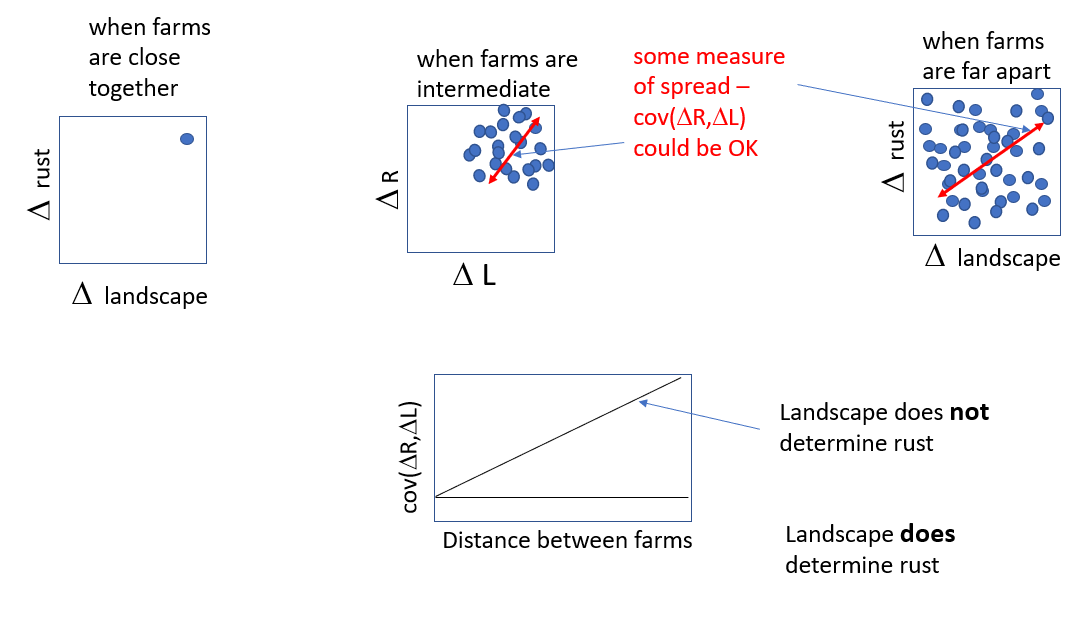


**Figure 2** Correlation between maximum roya and % forest within a certain radius. Each panel title indicates this radius. Pearson correlation coefficient **(a)** is shown between log transformed roya and arcsin square-root transformed % area. Spearman correlation coefficient **(b)** is shown between the untransformed roya and % area ranks



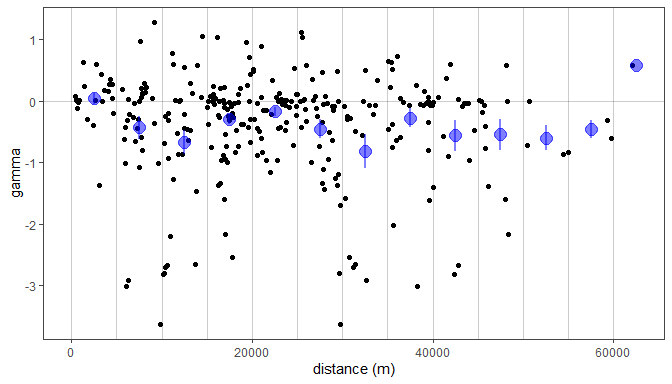
**Figure 3** Correlation between maximum roya and % pasture within a certain radius. Each panel title indicates this radius. Pearson correlation coefficient **(a)** is shown between log transformed roya and arcsin square-root transformed % area. Spearman correlation coefficient **(b)** is shown between the untransformed roya and % area ranks.

## Landscape + roya variables in variograms against distance (John’s package idea)

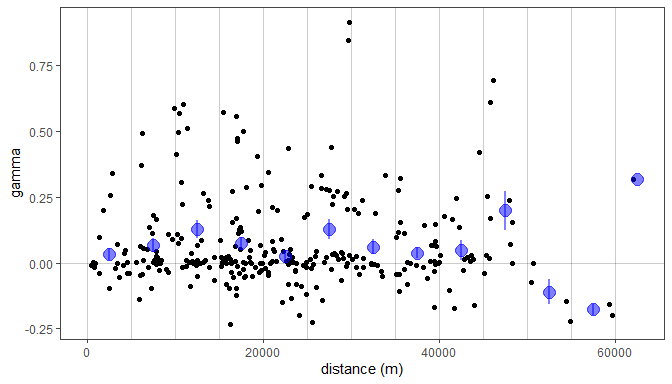
Here is John’s idea: 

I think a way to get to this is to plot the pairwise geographic distances between farms against the product of their pairwise differences between land cover and roya, as , where and are the landscape and rust Z-score values for a given pair of farms and , over all and for (the lower triangle of a pairwise matrix). This is similar to the semivariogram. Also similar to the semivariogram, you could take the mean over a certain band of separation distances to get something like the covariance for farms separated for that band of distance.

I see a connection between the Pearson correlation and John’s idea for looking at landscape effect. One way to define the Pearson correlation is as the mean of the product of Z scores for the two correlated variables . So by taking the mean of what I call (gamma) above, I think it is a kind of Pearson correlation for farms within groups of separation distances. In the plots below, the black dots are the pairwise values of gamma and the blue dots are the means of the 5000m bins shown by black vertical lines.



**Figure 4** Forest semivariograms. Black points are individual farm pairs with Y axis (gamma) = , or the standardized difference in rust times the standardized difference in landscape. Blue points are the averaged values over the values within 5000m bins, with lines representing standard error



**Figure 5** Pasture semivariograms. Black points are individual farm pairs with Y axis (gamma) = , or the standardized difference in rust times the standardized difference in landscape. Blue points are the averaged values over the values within 5000m bins, with lines representing standard error.