Introduction

In order to investigate the relationship between biological response variables (e.g., the abundance of a species) and explanatory environmental variables (e.g., soil characteristics) in ecological field surveys, ecologists collect observations at different spatial locations. They usually depend on systematic or random sampling designs if they have no prior knowledge about the spatial structures of variables. The reference paper seeks to answer whether researchers can use the information, obtained from previous surveys, to revise the sampling design to maximize the ability to identify the relationship between the response and explanatory variables. The authors measure the frequency of type 1 error (the rejection of the null hypothesis when in fact there is no effect of the environment variable on the response variable) and estimate the power for the different combinations of sampling designs and methods of statistical analysis. Power is measured by the rate of rejection of the null hypothesis when an effect of the environment on the response variable is present.

For the statistical analysis, a modified t-test developed by Dutilleul for correlation coefficients is compared to regular regression autocorrelation. They are compared to see if we can eliminate or control for the effect of spatial autocorrelation when it is present.

In this paper, I sought to reproduce the simulation of spatial data that would be sampled and the statistical analyses for the combinations of chosen spatial autocorrelation and sampling designs. All the simulations, illustrations, and analyses are coded in R because of the strength of existing libraries for generating spatial data and conducting statistical analyses.

Spatial autocorrelation

In our simulations, spatial autocorrelation is present both in the response and explanatory variables. Spatial autocorrelation is the term used to describe the presence of systematic spatial variation in a variable and the degree to which values at spatial locations are similar to each other. When generating spatial autocorrelation, the extent of the spatial influence is determined by the range of variogram. The range is the distance beyond which observations are no longer spatially correlated. The nugget effect represents the small-scale spatial variations within the fields usually due to measurement errors from a man-made or sensor measurement. The partial sill is the magnitude of variation of the variable. A higher partial sill means a stronger spatial structure. In the paper, spatial autocorrelation was generated by the exponential variogram model with my choice of the range, nugget, and partial sill.

Models

The implementation of my paper is based on the model description of the original implementation. I attempted to follow the structure and the setting. However, as we do not have access to the code of the original implementation, the simulation codes were developed originally without the dependence on the reference paper. In our simulations, a response variable consists of the sum of 3 separate effects. The effects are the influence of an explanatory environmental variable (), spatial autocorrelation in the response variable (), and a spatially unstructured random error component () taking independent values for each observation i:

The environmental variable consists of a deterministic structure, a spatially autocorrelated error component, and a spatially unstructured random error:

The model is assumed to be a linear function of the ecological to the environmental variables. The linear effect is modelled by multiplying E by an effect-size parameter (transfer parameter). By substituting the environmental variable, the model for the response variable R can be written as follows:

Methods and Modifications

First, I generate an experimental surface in a field containing 100 by 100 units and obtain a distance matrix. Before simulating an explanatory environmental surface and a response surface, I chose the spatial locations that would be sampled. In the original paper, the authors generate an explanatory environmental surface and a response surface, then extract the explanatory variable and response variable. However, the issue with generating an entire variable surface, including unsampled sites, is that a great deal of time would be spent just on 1 simulation. Instead, I select sites that I would sample, then take only the rows and the columns of the distance matrix corresponding to those sites. Then, I simulate the response and explanatory surface on those selected sites. This process saves a lot of time as we simulate 1000 times for each sampling design. Then, the relationship between explanatory and response value is analyzed by conducting a correlation analysis and producing a probability associated with the t-statistics. Pairs of surfaces (E, R) are replicated 1000 times, and results are accumulated over all simulations of a run.

For the final statistic, the proportion of rejections of the null hypothesis was computed. For my simulation, the significance level is set to be 0.05. The null hypothesis for the test is that no effect of the environmental variable on the response variable. The alternative hypothesis for the test is that there is an effect of the environmental variable on the response variable. For correlation analysis, I followed the reference paper and used a regular t-test of the Pearson correlation coefficient. Dutilleul’s modified t-test was also used to examine how well the test can compensate for SA in the environmental and response variables. Modified t-test corrects the variance of the test statistics as well as the degrees of freedom in the presence of spatial autocorrelation. I used modified.ttest function from SpatialPack in R to compute the p-value.

When finding the proportion of type 1 error, I generate surfaces so that the null hypothesis is true. The transfer parameter β was set to 0 to get rid of the effect of environmental variable on response variable. In the simulation runs for studying the power, the transfer parameter β was set to 0.3 so that there is a relationship between the environmental and response variable.

Setup (Sampling Design, Structure, and Range)

The reference paper used thirteen different choices for the sampling designs. Out of 13 choices, I focused on 5 sampling designs, which are random, systematic, vertical stratified, vertical transact, vertical transact with two sampling intervals. For random sampling, 100 spatial points were chosen randomly. For systematic sampling, I select every 70th point on the surface until the sample size reaches 100. The interval for systematic sampling is not specified in the reference paper. The interval of 70 was chosen so that the obtained proportion of type 1 error is close to the result in the paper. For vertically stratified sampling, the surface is stratified vertically into 2 strata. Then, we sample 50 points randomly in each stratum (100 points in total). For vertical transect sampling, I chose the column randomly in the field and selected every 2nd point within the selected column. The transect starts on row 1 and the sample size is 50. Lastly, for the vertical transect with two sampling intervals, I chose the column randomly in the field again and select the points with alternating intervals of 1 and 2 points. For example, I select the first row and second row because of the interval of 1 but skip the third row and grab the fourth row for the interval of 2. The transect started on row 1 and reached down to row 74 in the chosen column. The sample size for this sampling design is 50 as well.