**Reproduction/Analysis of ecological field surveys**

**Introduction**

In this paper, I sought to reproduce the simulation of data and statistical analyses in the paper, “The consequences of spatial structure for the design and analysis of ecological field surveys” (Legendre et al. 2002). Even though the authors of the paper provide the general setups and methods for analyses, they did not provide any code for their simulations. One of my goals in this paper is to make their simulations reproducible by providing the code in R.

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 variable on the response variable is present.

For the statistical analysis, a modified t-test developed by Dutilleul for correlation coefficients is compared to a t-test on the correlation coefficient assuming independence of observations. 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 ("Spatial Autocorrelation" 2020). 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 ("Variogram And Spatial Autocorrelation – Aspexit" 2020). The nugget effect represents the small-scale spatial variations within the fields usually due to measurement errors from a man-made or sensor measurement (Galton 1973). The sill is the magnitude of variation of the variable while the partial sill is the variation that is spatial in nature. A higher partial sill means a stronger spatial structure. In the paper, spatial autocorrelation for the errors was generated by the following exponential covariance model defined with different values of the range, nugget, and partial sill:

where is the nugget value, is partial sill, ρ is the range, and is obtained from the distance matrix.

**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 there is 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 the 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.

Even though six types of underlying spatial structures were available in the program, I have implemented three deterministic structures (no deterministic, linear gradients, and two zones). No deterministic structure has random normal error only. Hence, all the points on the surface are given 0 for this structure. The second structure is linear gradients from north to south and from west to east. The lowest values are in the upper left-hand corner and the highest values are in the lower righthand corner of the map. The linear gradient structure is defined by the equation D = 0.03x + 0.03y. 0.03 was chosen for the multiplicative value. The third structure is two zones deterministic structure. The surface was separated into north and south potion by a discontinuity. All the values in the north potion are set to be 6 while values in the south portion are 1.

The ranges, which determines the extent of the zone of spatial influence, were {0, 20, 50} points for random, systematic, and vertical stratified sampling designs. They were {0, 4, 16} points for vertical sampling design and vertical sampling design with two intervals. I have summarized all the simulations carried out to study the rate of type 1 error and power in Table 1.

Table 1. List of all the simulations

|  |  |  |
| --- | --- | --- |
| Deterministic Structure in the environmental variable | Sampling Design | Ranges |
| No deterministic structure | Random | {0, 20, 50} |
| No deterministic structure | Systematic | {0, 20, 50} |
| No deterministic structure | Vertical stratified | {0, 20, 50} |
| No deterministic structure | Vertical | {0, 4, 16} |
| No deterministic structure | Two Vertical | {0, 4, 16} |
| Gradients in two directions | Random | {0, 20, 50} |
| Gradients in two directions | Systematic | {0, 20, 50} |
| Gradients in two directions | Vertical stratified | {0, 20, 50} |
| Gradients in two directions | Vertical | {0, 4, 16} |
| Gradients in two directions | Two Vertical | {0, 4, 16} |
| Two Zones | Random | {0, 20, 50} |
| Two Zones | Systematic | {0, 20, 50} |
| Two Zones | Vertical stratified | {0, 20, 50} |
| Two Zones | Vertical | {0, 4, 16} |
| Two Zones | Two Vertical | {0, 4, 16} |

**Results (No deterministic structure in the environmental variable)**

In this section, I present the results of my reproduction of the simulation with no deterministic structure. The results are very similar as shown in Figure 1. The rate of type 1 error is inflated when autocorrelation is present in both environmental and response variables. Dutilleul's modified t-test successfully corrects rate of type 1 error with all sampling designs.

For the power study, the simulation shows that there is not much difference in power for all sampling designs. The power is reduced for vertical and vertical sampling design with two intervals due to smaller sample size of 50. The presence of SA in the response variable reduces the power of tests as shown in Figure 2. In my simulation results, the power is reduced for modified t-test in the presence of SA in the response variable. It was not observed in the reference paper. Overall, the similar results could be reproduced.

Chart

Description automatically generated

Graphical user interface

Description automatically generated

Figure 1. Type 1 error rates and 95 % intervals of the ordinary t-test for increasing values of ranges.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Figure 2. Power study for increasing values of ranges.

**Results (Gradients in two directions in the environmental variable)**

If a broad-scale gradient in the environmental variable is present, Figure 3 shows that the regular t-test is valid only when SA is not present in the response variable. Rate of type 1 error inflated for all other cases. Dutilleul’s modified t-test can corrects the rate of type 1 error by reacting to the deterministic structure.

In power study, power is high for ordinary t-tests when the test is valid. Vertical sampling design has a smaller power due to the smaller size of sampling design. For modified t-test, power is small in the presence of SA in the environmental and response variables. The results are similar to that in the reference paper.

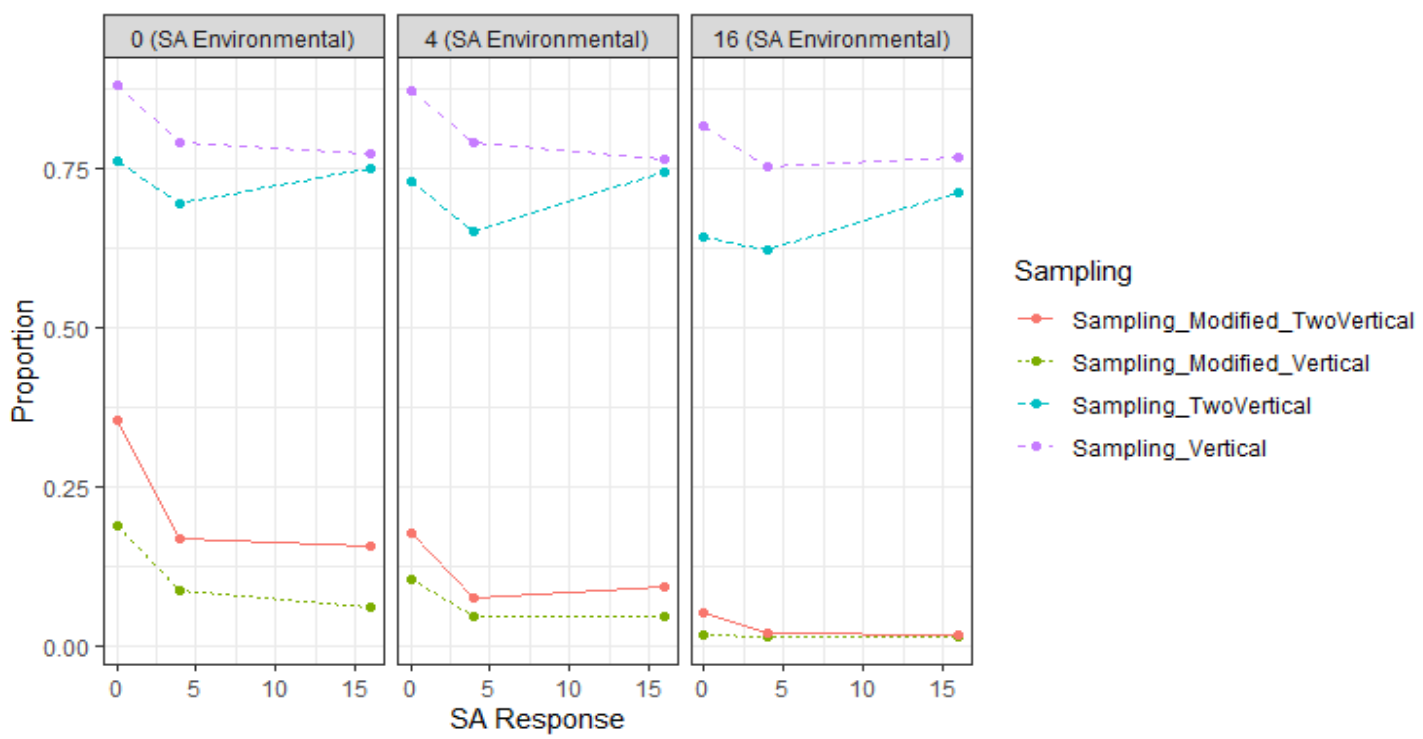
Chart

Description automatically generated

Chart, line chart

Description automatically generated

Figure 3. Type 1 error rates of the ordinary t-test and modified t-test



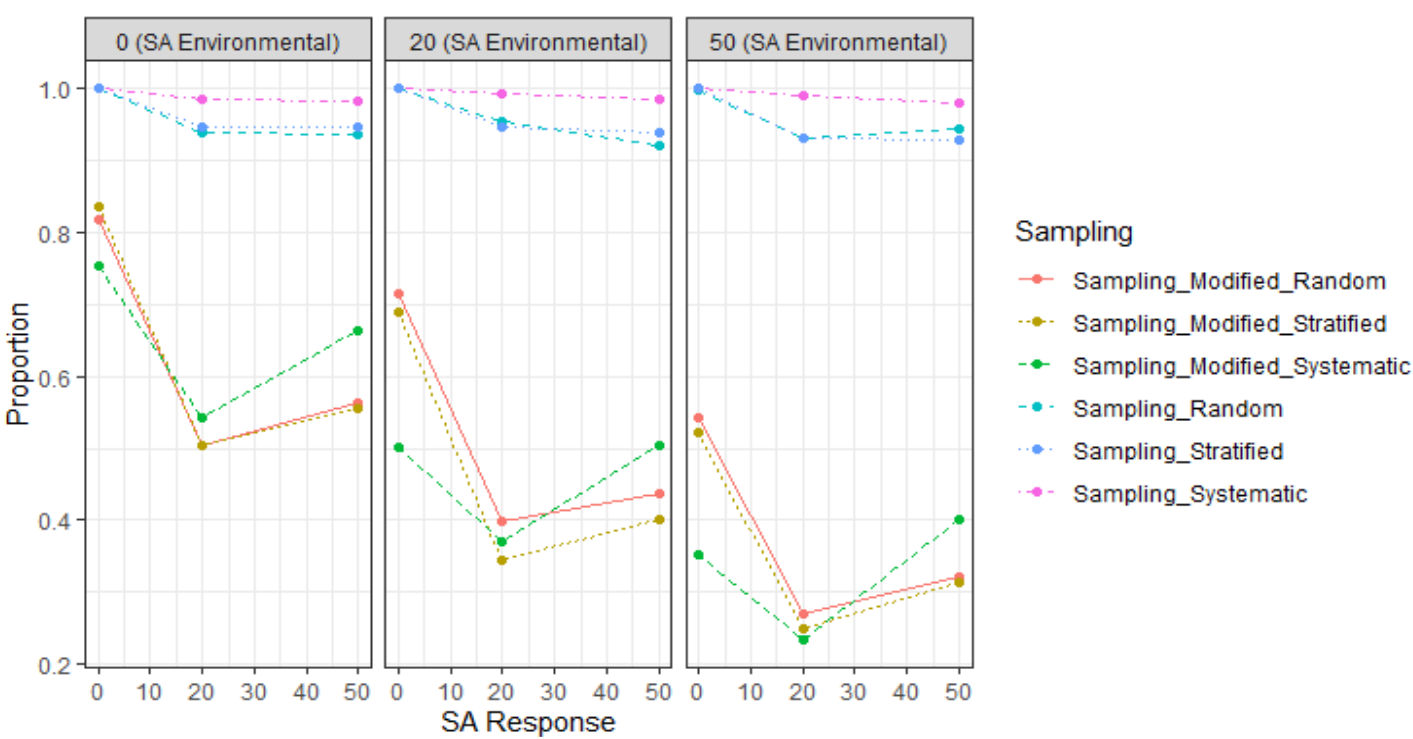


Figure 4. Power study for gradients in two directions

**Results (Two zones)**

Figure 5 shows that rate of type 1 error inflated whenever SA in response variable is present. The observations are the same as in the case of a gradient. Dutilleul’s modified t-test can corrects the rate of type 1 error by reacting to the deterministic structure.

Power study shows that equally strong power was obtained for random, systematic, and stratified sampling designs. For vertical transect with a single sampling interval, power is almost zero as shown in Figure 6. Power is also small for the vertical transect with two sampling intervals.

**Chart, line chart

Description automatically generated**

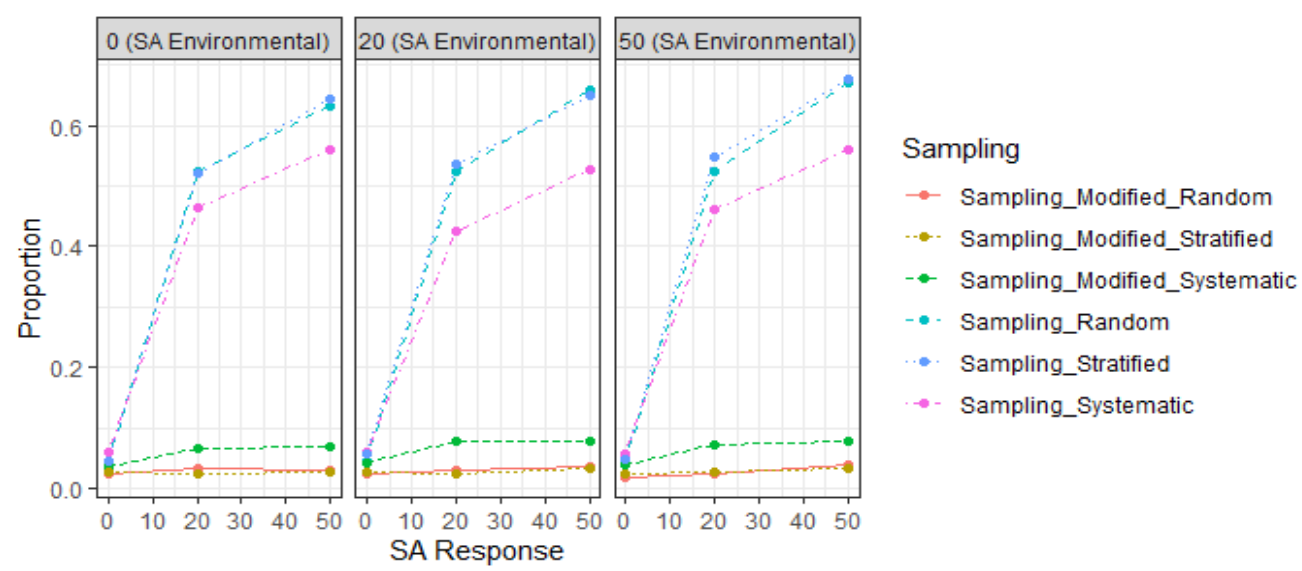


Figure 5. Type 1 error rates of the ordinary t-test and modified t-test for two zones

Chart, line chart

Description automatically generated

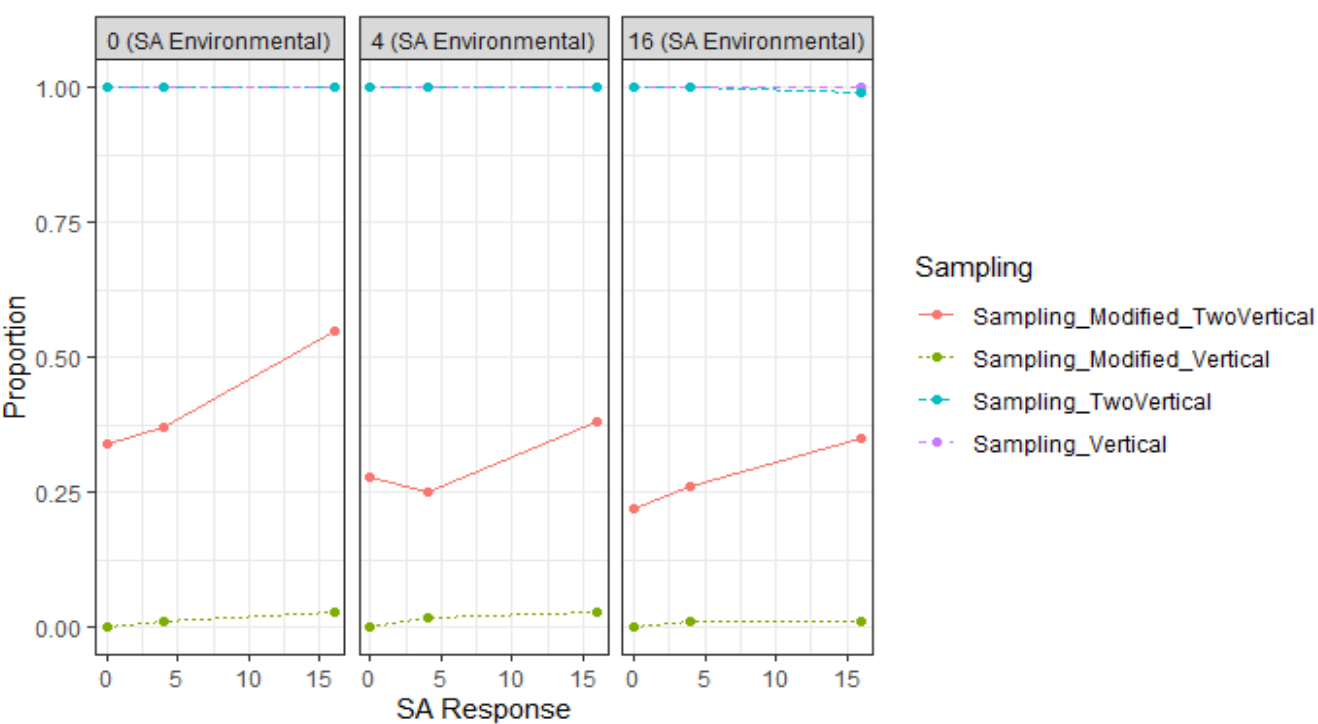


Figure 6. Power study for two zones

**Conclusion**

Overall, the results are comparable to that reported in the reference article. In the presence of spatial autocorrelation in both environmental and response variable, a normal t-test on the correlation coefficient becomes invalid. The test of significance is not disturbed when there is spatial autocorrelation in one of the variables. The tests of significance also inflate rate of type 1 error when there are a broad-scale spatial structure present and spatial autocorrelation in one of the variable. I could confirm that Dutillel’s modified t-test could take the effect of spatial autocorrelation into account and reduce the rate of type 1 error. It could also effectively corrects for the presence of a broad-scale spatial structure in the data.

References

Legendre, Pierre, Mark R. T. Dale, Marie-Josée Fortin, Jessica Gurevitch, Michael Hohn, and Donald Myers. 2002. "The Consequences Of Spatial Structure For The Design And Analysis Of Ecological Field Surveys". Ecography 25 (5): 601-615. doi:10.1034/j.1600-0587.2002.250508.x.

Galton, Francis. 1973. Natural Inheritance. New York: AMS Pr.

"Spatial Autocorrelation — R Spatial". 2020. Rspatial.Org. https://rspatial.org/raster/analysis/3-spauto.html.

"Variogram And Spatial Autocorrelation – Aspexit". 2020. Aspexit.Com. https://www.aspexit.com/en/variogram-and-spatial-autocorrelation/.