Sampling design of ecological experiments

Introduction

Before we start collecting data for our ecological study, we should make informed decisions about the structure of the sampling design, namely where, how often and how many samples will we sample. If the design of the study is flawed, there is often nothing much that can be done with it afterwards; statistics is not a magic wand that can fix everything, and what eventually may happen if the design is incorrect is that we need to delete some data or work with much lower effective sample size than we would wish to.

Ecological experiments are of two types: **manipulative experiments**, and **natural experiments** or **observational studies**. In manipulative experiments, the researcher manipulates levels of predictor variable X (independent variable) and measures the response of variable Y (dependent variable). At the same time, the design of the experiments attempts to control for other confounding factors, e.g. caused by natural environmental heterogeneity between sampling sites. (Confounding factors are those which can also affect our studied object, but we are not interested in them and often even cannot measure them). For practical reasons, manipulative experiments are usually done on small spatial scales in sense of both grain and extent (80% of field experiments are done on plots < 1 m^2), with few replicates, and on fast living organisms. In natural experiments (observational studies), the dependent variable Y is "manipulated by nature", and the researcher observes (monitors, measures) both the independent variable and the output of such "natural" manipulation. Sampling design can partly control for confounding factors (e.g. by selecting only localities which are similar in environmental conditions) and we can also measure additional factors with possible effects on the dependent variable, but this control is often limited, and in many cases, there are confounding factors we are not aware of when we sample the data. Observational studies could be done on both small or large scale in the sense of grain and extent, on short or long-living organisms, and observations can be repeated for decades.

An important difference between a manipulative and natural experiment is the assumption about the direction of the relationship between independent and dependent variables. In manipulative studies, we are directly applying the experimental treatment, while controlling for the confounding effect of other factors. If we observe the change in the dependent variable, we can assume that it is caused by our manipulative treatment, i.e. that there is a causal relationship (*X is causing Y*, e.g. adding fertilizer to the meadow decreases the plant species richness). In natural experiments, the link between independent and dependent variables is much more uncertain, and the real effect could in reality be caused by some other, unmeasured variable correlated with our measured independent one. We can talk about correlation, but not causation, and our conclusions should be carefully formulated in that way (*Y is correlated to X*, i.e. a higher nutrient content is related to higher species richness, but this may be confounded e.g. by nutrient-rich sites being wetter).

In general, experiments are used to answer one of the following questions (adapted from Gotelli & Ellison 2013):

- Are there spatial or temporal differences in variable Y? This is a topic for *pattern description* study of variable Y in space or time, and is one of the most common questions in observational studies.
- What is the effect of predictor X on dependent variable Y? This question is focused on *hypothesis testing*. The most straightforward way is to manipulate X and see the response of Y (manipulative experiments uncovering the causal effect of X on Y) but could be approached also by natural experiments (observing the correlation between X and Y, while acknowledging the possible effect of other confounding factors).
- Are measurements of variable Y consistent with the prediction of hypothesis H? This is a *confrontation between theory and data*, and can be done in both manipulative and natural experiments. The disadvantage is that alternative ecological hypotheses often generate a similar prediction, and also the prediction of some hypotheses is far from simple.
- Which model best represents the relationship between X and Y? This is a task for *mathematical modeling*, i.e. building of the model which best describes the pattern and possibly allows for prediction.

In manipulative experiments, we first sample baseline data (before the treatment is applied), then apply the treatment, and then (after allowing time for treatment to work) we sample (often repeatedly) the response of

the dependent variable. There are two alternative approaches: **press experiment**, in which the treatment is applied at the beginning and then reapplied regularly, and we measure the resistance of the dependent variable to experimental treatment (the extent to which it resists the changes in the environment). On the other side, in **pulse experiment**, we apply the treatment only once (at the beginning) and then observe the resilience of the studied variable, i.e. whether and how fast it will recover from a single perturbation.

In natural experiments (observational studies), we can distinguish between **snapshot experiments**, which are done only once in time and are usually replicated in space, and **trajectory experiments**, which are usually replicated in time (and often also in space). Snapshot experiments are the most common observational studies in community ecology; we collect samples in several (or many) localities in a relatively short time. This includes studies focused on a succession of communities after disturbance, applying the concept of space-for-time substitution, i.e. localities in different successional stages (but hopefully with similar initial environmental conditions) are sampled. On the other hand, trajectory experiments are done by repeated sampling in (usually permanently fixed) sites several times and may include successional studies conducted for several years or monitoring studies focused on the dynamic changes of vegetation (e.g. forest dynamics plots).

Design of manipulative experiments

Here we will focus on the spatial distribution of sampled plots in the field (or lab/greenhouse) manipulative experiments (mainly following Chapter 3 in Šmilauer & Lepš 2014). The important function of sampling design is to control for potential (and unwanted) environmental heterogeneity between sampling plots in the experimental area. Each design has some advantages and some limitations, and design decisions need to be considered when analyzing the data (e.g. when deciding how to restrict the permutation schema when testing the significance, or which covariables to include in the model).

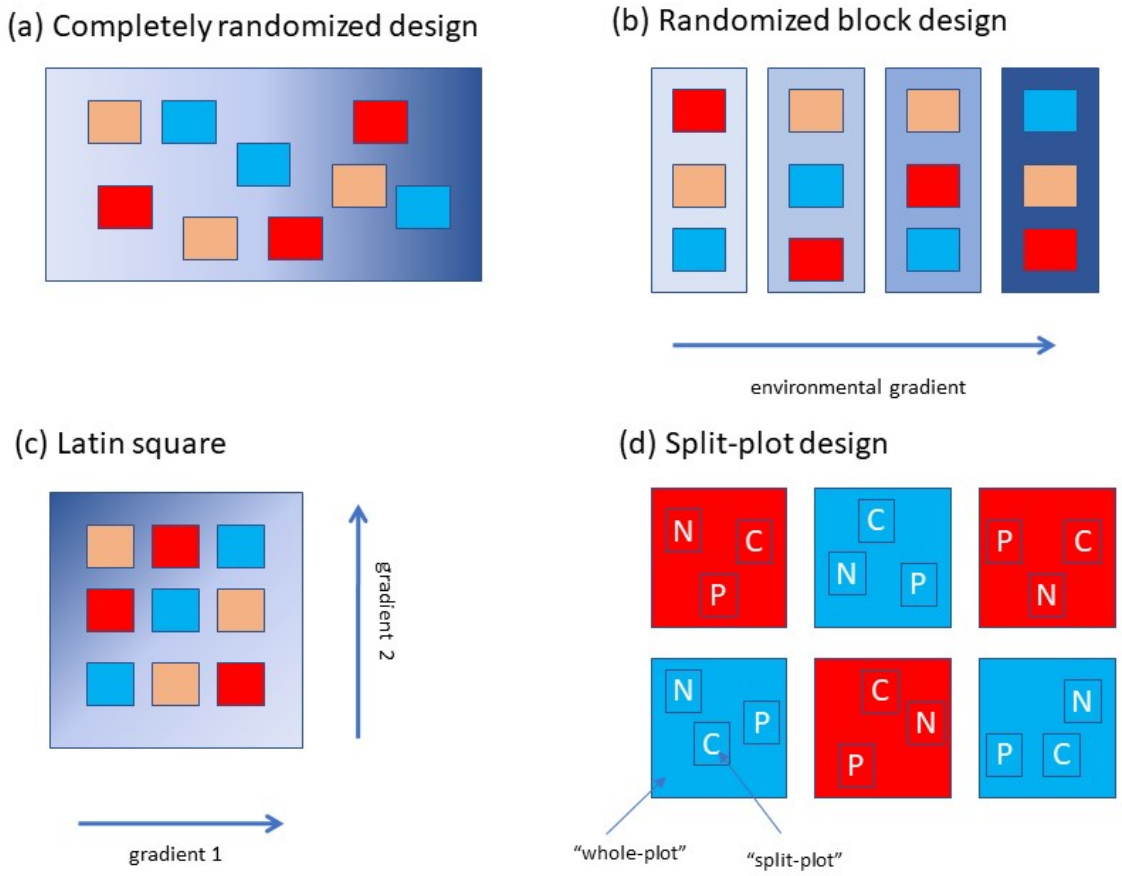

Figure 1

Completely randomized design (Fig. 1, panel a) distributes sample plots randomly across the space. It brings the highest number of degrees of freedom to the analysis (equal to the number of plots), but it does not consider environmental heterogeneity, so if the study area is heterogeneous, it could be unsuitable. It is not so simple to establish (human is bad at generating "random" positions) and may be prone to unintended pseudoreplication problem (see below).

Randomized block design (Fig. 1, panel b) is composed of smaller blocks, each with exactly one replicate of each treatment (number of blocks equals the number of replicates). Block design decreases the number of degrees of freedom, as the assignment of the plot into the block has to be acknowledged as a covariable in the model, and it also limits the number of permutations, as the plots can be permuted only within blocks. On the other side, it offers control for environmental heterogeneity (with minimum heterogeneity within and maximum between blocks), and environmental differences between blocks can then be removed from the analysis by including the block as covariable. Blocks can be set not only in space (spatial) but also in time (temporal).

Latin square (Fig. 1, panel c) is suitable for situations with two environmental gradients in the experimental area. Each column and each row contains exactly one replicate of the treatment, so the number of replicates in the Latin square equals the number of treatments (but there can be more Latin squares per locality).

Factorial design is for experiments with more than one studied factor (e.g., studying the effect of fertilizing and mowing), where each level of each factor is combined with each other (e.g., no fertilizing + no fertilizing, no fertilizing + mowing, fertilizing + no mowing, fertilizing + mowing). Either of the already described designs can be used for them.

Hierarchical design (also "nested design") includes plots with nested subplots. If used for experiments with a single factor, levels of this factor are replicated on the plot level, and subplots represent repeated sampling, which does not increase error degrees of freedom but increases the precision of the estimate of the plot-level value. If used for two factors (also called a split-plot design), then levels of the first factor are replicated on the plot level ("whole-plots"), and the levels of the second factor are nested within plots on the subplot level ("split-plots", see schema on Fig. 1, panel d). The difference between split-plot design and randomized block design is that in the former we are interested in the effect of an environmental variable on the "whole-plot" level, while in the latter, we are not interested in the environmental difference between blocks and consider it a nuisance. There is also a terminological difference: in a split-plot design, we talk about plots with nested subplots, or "whole-plots" with nested "split-plots", while in the randomized block design we talk about "blocks" containing "plots"; see more details here. When analyzing data from a hierarchical design, we need to acknowledge the difference between plots and subplots.

Beware of pseudoreplication issues and mistakes in designing manipulative experiments

Pseudoreplicates are replicates that do not come with a (completely) new independent piece of information, but instead (more or less) replicates information already represented by other samples in the dataset. In statistical terms, pseudoreplicates do not increase error degrees of freedom (while may increase the precision of estimating the value of the response of a given treatment). Pseudoreplication issue arises because plots that are close to each other tend to be more similar in both independent and response variables than would be two plots selected in random (see more details in the section Spatial autocorrelation). The schema below shows two common cases when individual samples are to some degree pseudoreplicates, either because the plots in completely randomized design are spatially clustered with each cluster in a different part of the environmental gradient (Fig. 2, panel a), or because plots in replicated block design are wrongly replicated (plots of the same treatment are in one block, instead of having only one replicate of each treatment per block; Fig. 2, panel b). Pseudoreplicates are not completely useless, and if correctly treated, they can improve the power of the test.

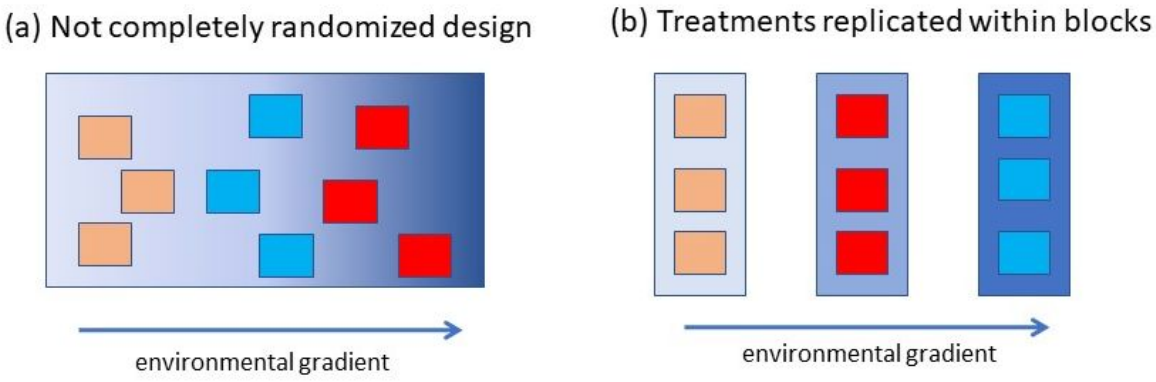

Figure 2

A less obvious mistake may be caused e.g. by the wrong orientation of blocks in the randomized block design (Fig. 3, panel a). The correct structure aims to maximize the environmental heterogeneity between blocks and minimize heterogeneity within blocks (scenario on the left), but this can be violated if blocks are oriented to extend along the gradient instead of perpendicularly to it (scenarios in the middle and right). In Latin square design (Fig. 3, panel b), each row and column should contain exactly one replicate from each treatment (the scenario on the left is correct, but on the right is wrong).

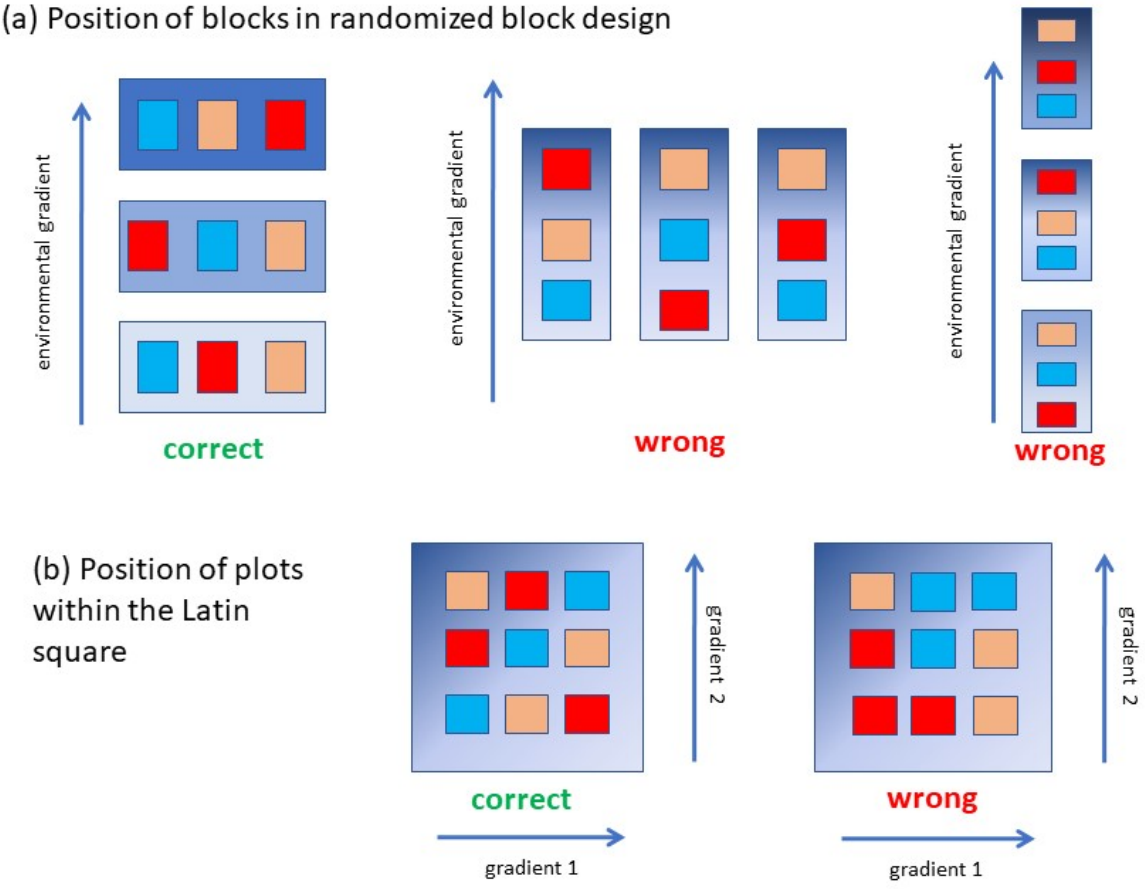

Figure 3

Sampling designs of natural experiments (observational studies)

For observational studies, we usually need to use existing spatial or temporal variation in the environmental variable we will use as a predictor, and also consider the presence of other, confounding variables that can make ecological interpretations of our results more difficult. Several decisions need to be done before going to the field and sampling; we need to know how to select the localities for our plots (preferentially, randomly, or using some sampling schema) and what is the minimum distance between individual plots to decrease the effect of spatial autocorrelation.

Preferential sampling design (Fig. 4, panel a) is a strategy where the locations of the plots are selected by researcher often based on the set of subjective criteria which are difficult to formalize and also difficult to standardize among different researchers. Preferential sampling is useful when one is trying to describe maximum variation of studied object (e.g. vegetation) in the region, while having at least rough idea about its spatial variation. Advantage is that the resulting dataset is likely to include both common and rare types of communities sampled in rather typical and representative sites. Disadvantage can be that the dataset may contain subjective bias caused by researcher's preferences (e.g. for more diverse communities, or communities containing some obvious species). Some studies have demonstrated that preferentially sampled sites tend to be more diverse than randomly selected ones, but the evidence is unclear.

Random sampling design (Fig. 4, panel b) selects locations of plots randomly, usually in advance by computer algorithm or in the field by some randomization procedure (e.g. "random walk" in randomly chosen direction for randomly chosen number of steps). There usually needs to be a set of clearly stated criteria under which randomly selected location is suitable for sampling (clearly, if it falls in the middle of the airport, that may not be a good choice). Advantage is that the subjectivity of preferential sampling is removed, and the dataset realistically represents the real variation of studied object in the landscape (study area). Disadvantage, along to much more time needed for location of sites in the field, is the fact that some rare community types may be completely missed, while common community types will be overrepresented, and also some of the samples may be highly not representative (done on sites where we would otherwise not survey). A solution to even representation of different community types in the dataset is a random stratified sampling design, where the area is first divided into "strata" (polygons with different sets of environmental conditions, e.g. wet + warm, wet + cold, dry + warm, dry + cold), and sampling is then restricted to random locations within these strata, aiming to collect comparable number of replicates from each strata.

Systematic sampling design follows certain spatial structure, e.g. linear transect(s) (located either in a random direction or in the direction of the strongest environmental change, so-called "gradsect", Fig. 4, panel c), regular lattice (with same distances between neighbouring sites in the same row and column, Fig. 4, panel d) or even grids (with plots directly aligning to each other). The lattice design is also called rectangular grid design. Sampling design needs to be considered when analyzing the data (restricted permutation schemas need to be applied in the case of systematic designs).

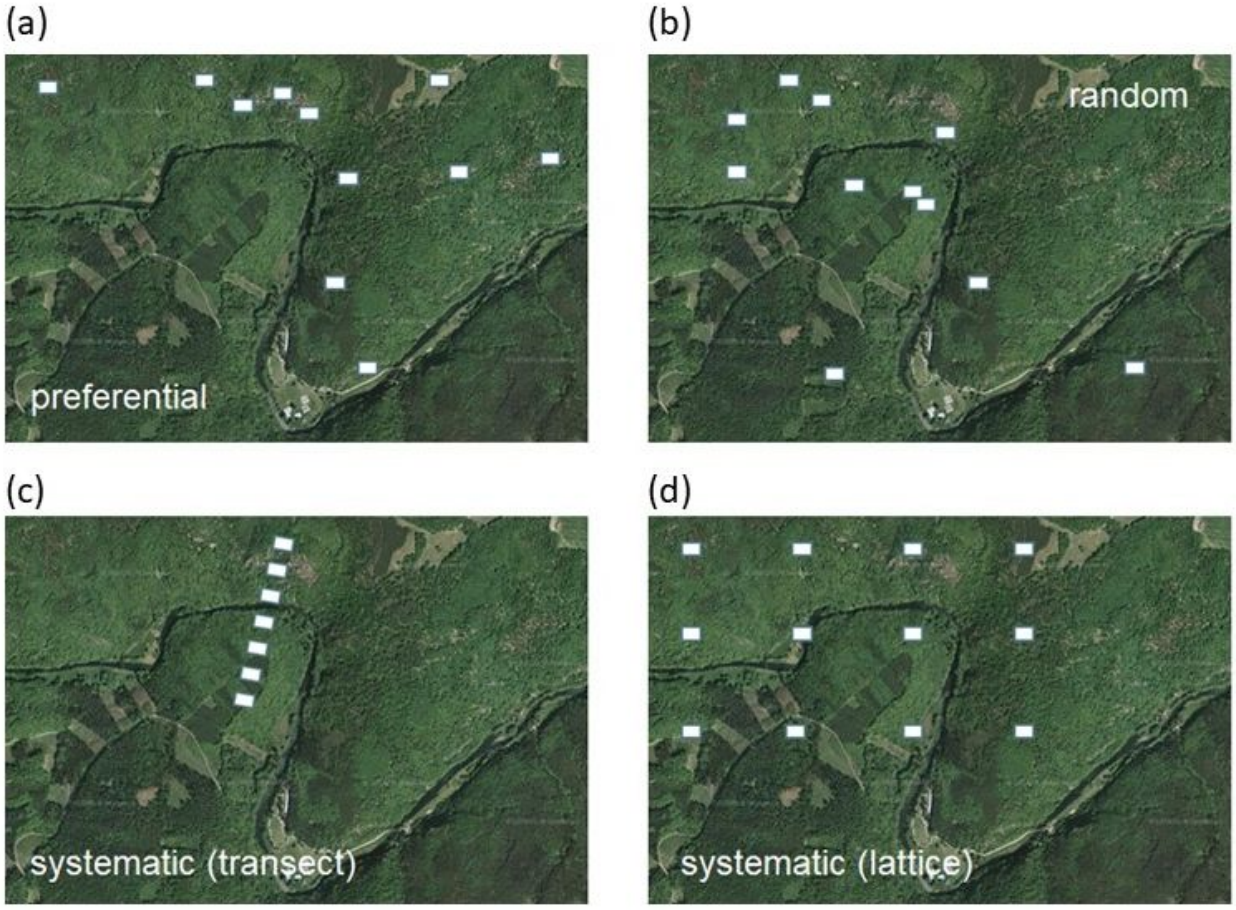

Figure 4

Problem of spatial (or temporal) autocorrelation

The pseudoreplication issue in observational studies is better described as a problem of spatial (or temporal) autocorrelation. For spatially autocorrelated variable, the observations that are closer together tend to be more similar than observations that are further apart (positive spatial autocorrelation). Autocorrelation can be induced by dependence of the observed variable on another one which itself has a spatial structure (e.g. vegetation structure influenced by soil pH which itself has strong spatial autocorrelation), or it can be caused by spatial processes (e.g. neutral processes in the community, including dispersal and random drift, may cause spatial dependence of communities observed nearby, without being caused by environmental differences). The problem with spatial autocorrelation appears when we test the relationship of two (positively) spatially autocorrelated variables (e.g. by regression) and find that the residuals themselves are spatially autocorrelated. If we test the relationship of autocorrelated variables, the test tend to have inflated Type I error rate (returning more optimistic results than would be warranted by data). This is because the effective number of degrees of freedom is lower than the number used in the test (and derived from the sample size), since some of the samples are spatially dependent on others and therefore do not bring entire error degree of freedom to the analysis.

Some designs are more prone to the spatial autocorrelation problem than others. On the schema below are three situations with 12 sites, either randomly distributed in (possibly) large enough distances (Fig. 5, panel a), somewhat clustered into three nearby groups (Fig. 5, panel b) and highly clustered within three compact groups (Fig. 5, panel c). When thinking about these scenarios in terms of effective degrees of freedom, scenario feels the most as not affected by spatial autocorrelation (with 12 independent samples), while scenario c seems as the most affected (with degrees of freedom close to three), and scenario b somewhere in between.

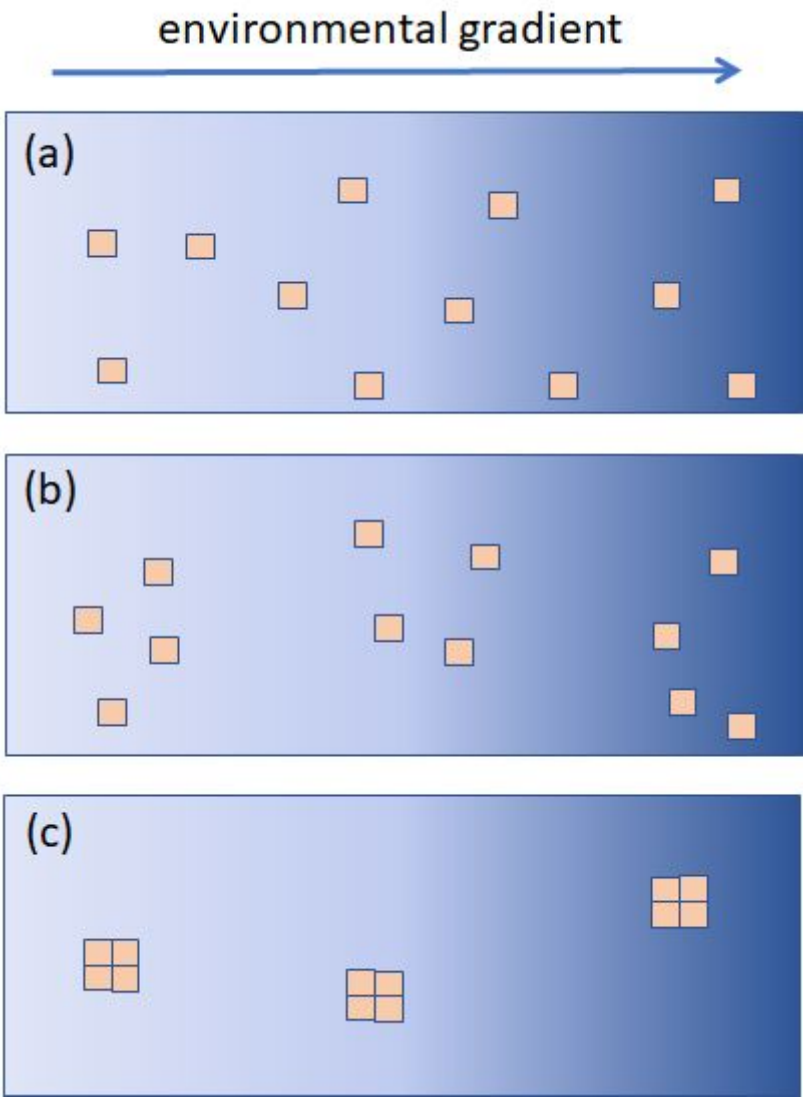

Figure 5

If the design is clearly structured (as panel c in the schema on Fig. 5), this structure is better to be acknowledge in the model or permutation schema (e.g. by testing differences between clusters of sites, not individual sites, and by permuting sets of sites between clusters, not all sites independently). If the design is more fluid (panel a or b of Fig. 5), it may be useful to test the spatial autocorrelation in model residuals (e.g. using Moran's I) to see whether it is present or not. If present, there are few solutions; one is to stratify the dataset by removing sites which are too close to each other and may be spatially dependent (this is possible in case we have large dataset and don't worry about losing data). Alternative solution is to use analytical method which can deal with pseudoreplications caused by spatial autocorrelation. If plots are in line transects or spatial grid, this should be acknowledged in the permutation schema. If the positions of sites is somewhat uneven, methods of spatial filters (PCNM, MEM) may be used to model the spatial structure and remove it as covariable from the model.

(Consider checking my blog post on the issue of pseudoreplications in observational studies, https://davidzeleny.net/blog/2022/03/21/pseudoreplications-more-tricky-than-we-may-think/, where I point out that the problem is much more wide spread than one would expect; I also provide a simple numerical exercise to show how easily can Type I error rate be inflated when analyzing data from nested design without acknowledging the nesting in the model).

From:
https://anadat-r.davidzeleny.net/ - **Analysis of community ecology data in R**

Permanent link:
https://anadat-r.davidzeleny.net/doku.php/en:sampling_design

Last update: **2026/08/16 16:58**