FES 524: Natural resources data analysis

Reading 3.1: Random effects

2024-01-18

Contents

1 Random effects

In reading 2.1 you were introduced to sources of variation that we designated as *fixed* or *random*. This week we will go through a more formal introduction to random effects.

1.1 Fixed versus random (the classic conundrum)

When we are deciding on fixed versus random effects in this class, we are discussing categorical variables. A continuous variable cannot be a random effect as we are discussing here. Random effects are based on categorical variables or factors. If you have a continuous variable that is a source of variation for your response variable but is not of specific interest, you will put that variable in the fixed effects as a covariate.

In some fields you may see these random effects called "random intercepts", but note that we can also have random slopes where the slope is determined by what group of the categorical variable used as a random effect we are in. In other words, this is in addition to having the continuous variable as a fixed covariate. We will not get to random slopes in this class.

Deciding if a factor should be considered a fixed effect or a random effect is harder than it might look compared to rules you will see in some textbooks. Some variables might be both; a variable may be fixed in one analysis but random in another.

Schabenberger and Pierce (2001), p. 627 have a nice quote about this:

> "Before proceeding further with random field linear models we need to remind the reader of the adage that one modeler's random effect is another modeler's fixed effect."

If you are struggling with whether your factor should be considered fixed or random, here are a couple of resources that may help below.

- Good overview: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#should-i-treat-factor-xxx-as-fixed-or-random
- Good discussion thread: https://stat.ethz.ch/pipermail/r-sig-mixed-models/2010q2/003710.html

Fixed effects

Here are a few indications that a factor should be considered fixed.

- We are interested in making inference about differences in means among groups (levels) within the categorical variable in question.
- We want to make inference only to the existing levels of a factor (the levels of a factor are the different categories/groups that factor has; i.e., the unique values the factor takes).
- If we repeated the study, we would use the same levels of the factor. For example, if we applied three thinning densities to stands in one study, we would choose the same three thinning densities if designing a new study to address a related research question about those thinning densities.

When working with fixed effects, we are interested in means of the groups and differences among those means.

Random effects

If a factor should be considered random we might:

- Be uninterested in talking about differences among group means and instead need to account for the variability caused by the factor.
- Want to make inference to not only the levels of the factor that were measured but to a population of levels that were *not* measured. For example, we might be interested in the population of stands that fall within our scope of inference and not the specific stands selected for the study.

3. Choose different levels of the factor if the study was done again. For example, if we took a new random sample of stands in a follow-up thinning study, the levels of the "stand" factor would be different than the levels from the original study since we used different stands.

When working with random effects, we are interested in *variances*.

Often times, random effects are based on variables we need to deal with because they are sources of variation in our study, but we are not actually interested in their effect on our response variable. Instead, we might be interested in understanding the effects of the other variables given *a typical group* (e.g., for a typical forest stand, what is the effect of the thinning treatment on growth). We understand there is variation among our groups and that we should account for it, but we are not interested in estimating the effect of a particular group. Note that in some fields, though, researchers are specifically interested in these sources of variation. In genetics, for example, investigators are often interested in estimating different variances for different sources of variation (e.g., genetic variation vs environmental variation).

One additional complication in all of this has to do with the number of levels a factor has. If there are very few levels in a factor, it is hard to justify that those levels really represent some larger population of levels. In the extreme case where there are only two levels of a factor, that variable will need to be treated as a fixed effect even if you otherwise would consider it a random effect since we cannot reliably estimate a variance from two data points. You may see some loose rules about the number of levels you should have

stated, such as a factor needs at least five levels to be treated as random, but this is a point of contention among statisticians. There are some random effects "purists" out there who will use a factor with as few as three levels as a random effect (Stroup discusses this in his book on the reading list if you are interested). Others will stick to having 5 levels or more for a random effect. Still others say you need as many as 42 levels before considering something a random effect. I generally use the rule of thumb that there should be at least five levels to treat a factor as random.

1.2 Observation-level random effect

As we touched on last week, we have already been working with random effects in this class. In fact, while it may not have been discussed in this way, you have worked with random effects any time you fit a linear model (e.g., t-test, ANOVA, regression).

Remember our random error term from last week's model? ϵ_i is the random error term for the i^{th} observation, with $\epsilon_1, \epsilon_2, ..., \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. This is an *observation-level* random effect. It is the random error due to the observations. A factor that represents this effect would have the same number of levels as observations in the dataset.

We model observation-level random effects as draws from a specified model distribution. In this case, and most commonly, the distribution is a normal distribution with a mean of 0 and variance σ^2. Remember that the residuals are estimates of the errors. To show something is an estimate, we either put a *hat* on it, or change the notation from a Greek letter to an English alphabet letter. Here is common mathematical notation for the residuals:

$$r_i = (y_i - \hat{y}_i).$$

We have a total n residuals, one for every observation.

The drawing below (Figure 1) shows an example of a normal distribution for the errors. The distribution is centered around 0 (the mean) and has some overall variance σ^2. We expect the residuals should all fall within this distribution. I show the first three residuals (these are theoretical values of the residuals, not the actual values from an analysis of the thinning dataset).

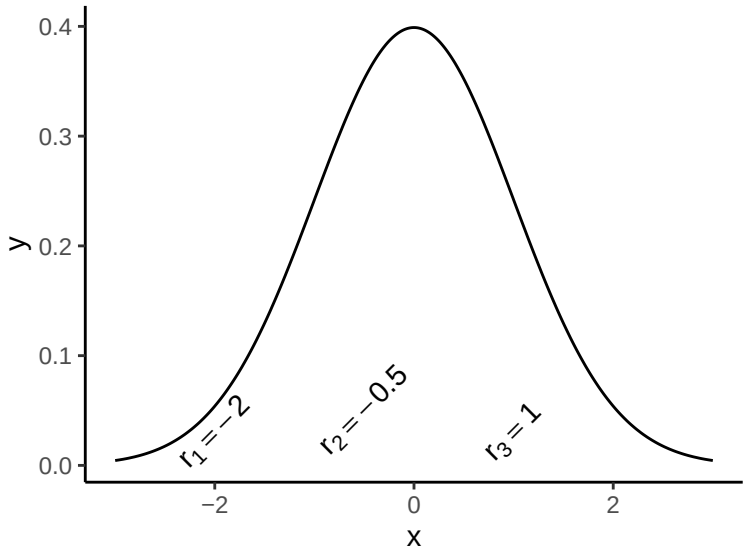

Figure 1: Example of assumed error distribution and hypothetical residuals falling within the distribution.

We use the distribution of the residuals to estimate the variance of the distribution of the errors. This value is used to calculate confidence intervals and test statistics. We are not usually interested in estimating or

reporting the values for the random effect of individual observations (i.e., we aren't explicitly interested in the specific values of the residuals).

2 Blocking

Before starting in on a discussion of blocking, let's review the definition for sources of variation from last week.

> **Source of variation**: A component of a study such that different levels of that component result in different values of a given response variable.

When designing a study, our goal is to minimize unexplained variation. To do this we could, for example:

1. Limit the scope of inference of the study to control variation
2. Collect data on covariates that cause variation to include in the model
3. Use a design to control some sources of variation

Using a study with blocking falls under the third option and is our focus for this week.

2.1 Blocking factors

A blocking factor is a categorical variable. If the blocking characteristic is continuous, then it must be discretized to use it as a blocking factor (e.g., high, medium, low). Blocking is done during the design phase of a study.

We are not directly interested in the effect of the blocking factor. However, we assume the blocking factor is a large source of variation *a priori*. We therefore need to account for it in order to estimate what we are interested in with more precision. The blocking factor is part of the study design and can help us gain resolution on the effect(s) of interest.

We think that the levels of the blocking factor affect the response variable in some systematic way. By "systematic" I mean one level of the blocking factor affects all observations within that level in the same direction. In some levels of the blocking factor the observed response variable values tend to be higher and in others they tend to be lower. This effect of one level of the blocking factor is not identical for every observation but on average the observations are all affected in the same direction. One way to describe this is to say that the mean response of one level of the blocking factor is different from the mean response of another level of the blocking factor.

When we are thinking about blocking factors we know we want to avoid confounding the effect of the blocking factor with the effect of whatever variable or protocol we are actually interested in estimating. Confounding occurs when the investigator can't reasonably eliminate plausible alternative explanations for an observed relationship between a variable of interest and the response variable.

Some examples will help clarify some of these concepts about blocking factors.

2.1.1 Study example - no blocking needed

In a hypothetical study, the goal is to compare means of the variable Y for a protocol with the levels A, B, and C. While soil moisture is not of interest, it is known *a priori* to affect Y. In reality, soil moisture is a continuous variable but in this example the variable has been categorized into low, medium, and high levels.

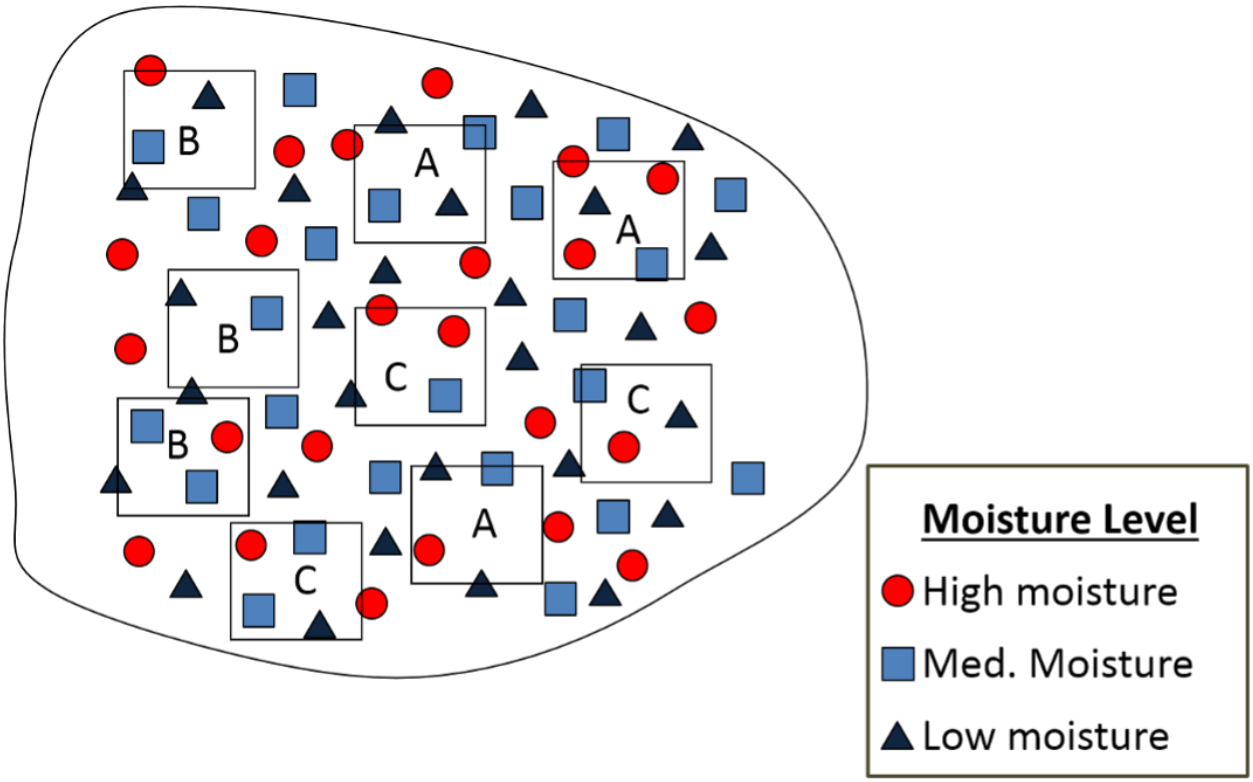

Figure 2: Study design where blocking is not needed to address research question about the protocol.

Figure 2 shows an example of a field study designed to achieve the research goal where blocking is not needed.

You can see the investigators randomly placed 9 plots, represented as rectangles, within the study area. One factor level of interest was randomly assigned to each one.

In this particular example, soil moisture varies across the whole study area. The plots are much larger than the variation in soil moisture, and you can see soil moisture varies within each plot. Because of this, blocking is not needed here. Overall plot soil moisture could be collected as a covariate, since there will be minor variations of average soil moisture within plots, but it would not be useful as a blocking factor.

2.1.2 Study example - Oops! Factor confounded with blocking variable

Figure 3 is another example of a field study designed to address the same research goals. In this scenario blocking is needed.

The investigators again randomly placed 9 plots within the study area and randomly assigned one of the factor levels of interest to each one. However, now soil moisture does not vary among plots like it did in the first example. There is a clear gradient in soil moisture across the study area.

By chance you can see that random assignment led to all of group B being assigned to plots with low soil moisture. Group B and low soil moisture are confounded, and investigators won't know if any patterns they see are due to the protocol of interest (group B) or the difference in soil moisture.

2.1.3 Study example - Complete randomized block design (CRBD)

If investigators knew about the soil moisture gradient in their study area in advance, they would want to block on soil moisture. An example of this is shown in Figure 4.

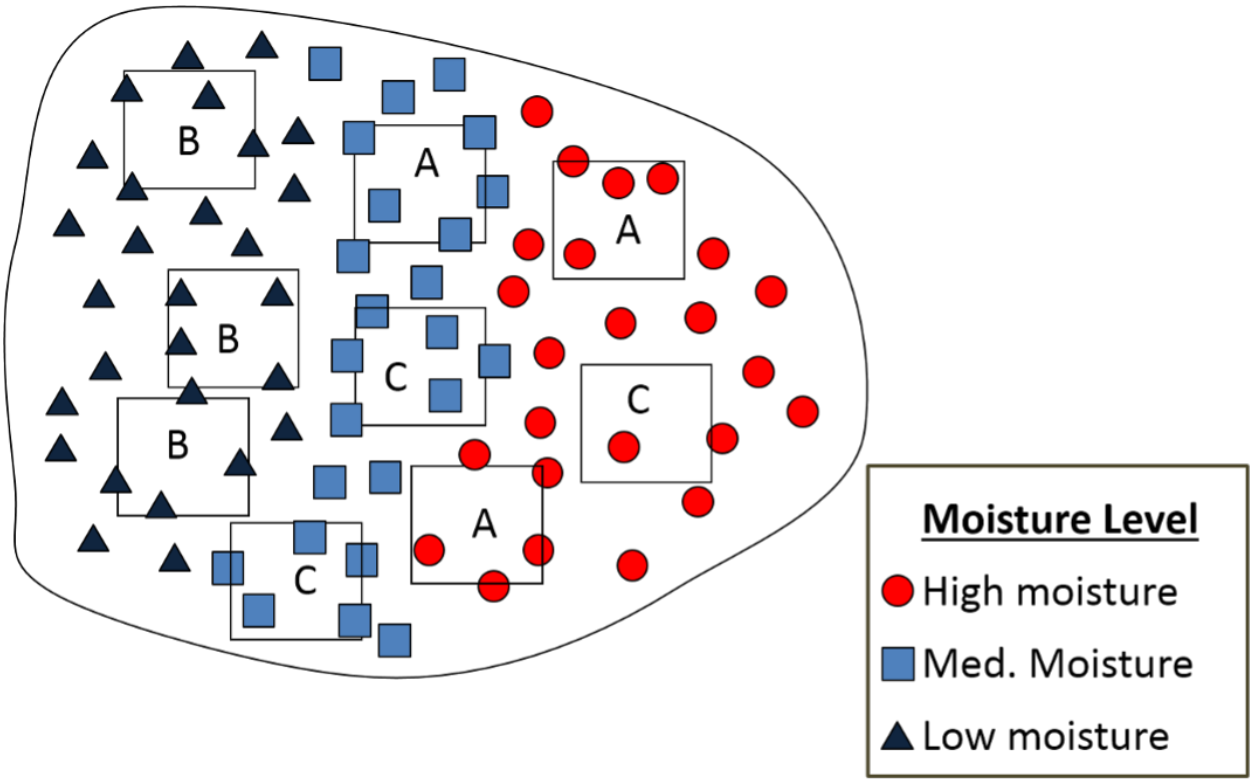

Figure 3: Study design where blocking was needed to address research question about the protocol, but the blocking variable ended up confounded with the treatment variable.

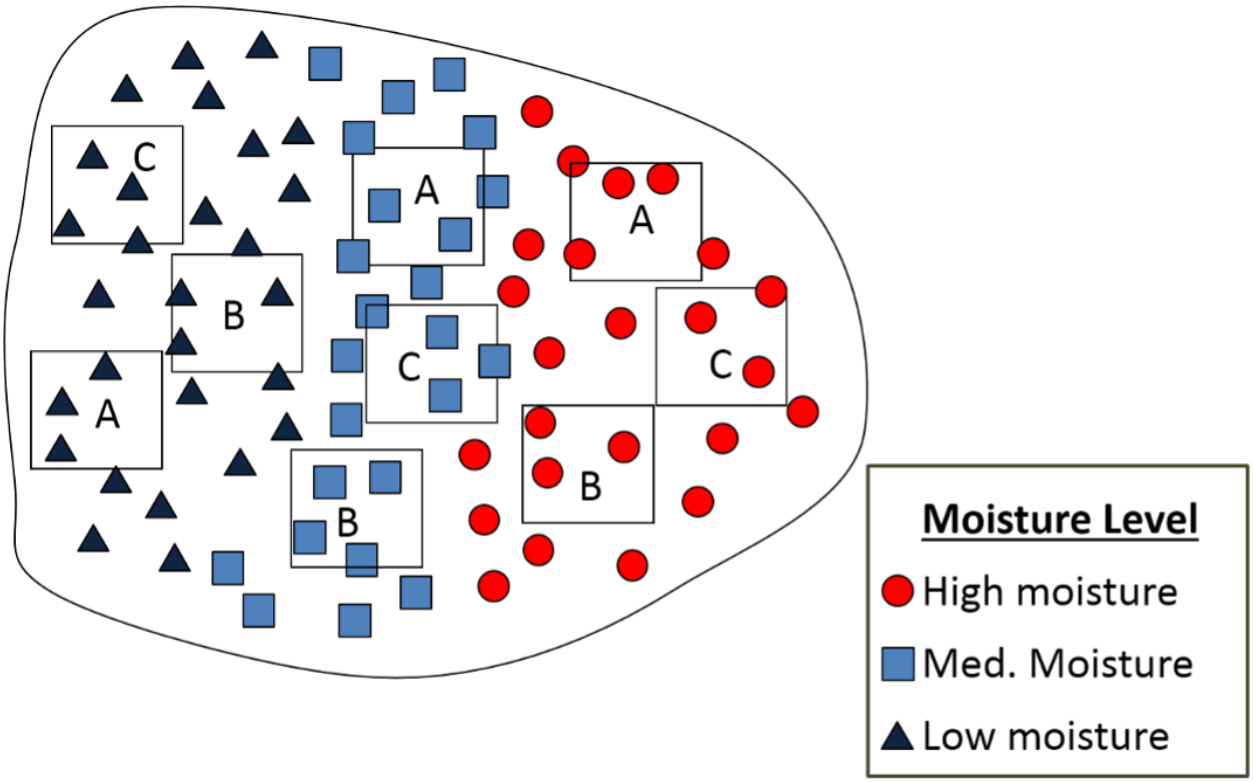

Figure 4: Example of a complete randomized block design.

Instead of randomly placing all 9 plots across the study area and then randomly assigning the factor level to each plot, investigators randomly placed three plots *within* each soil moisture group. They then randomly assign one of each of the levels of the protocol of interest within that soil moisture group. You can see that each of the three soil moisture groups has every level of the protocol of interest in it (A, B, and C).

The design approach described in this last example is what is often referred to as a complete randomized block design. Each soil moisture group could be referred to as a block, for a total of three blocks.

Note that for blocking to be effective in this last example, the investigators needed information about the study area before they started designing the study. Without information on the soil moisture gradient, the investigators could have tried to use blocking by haphazardly cutting up the study area into three pieces based on the geography of the study area and using those pieces as blocks. However, if those pieces didn't correspond to soil moisture or some other gradient, then the blocking would likely not be particularly effective.

2.2 Assumptions about blocking

As discussed earlier, for blocking to be effective we assume the levels of the blocking factor have some systematic effect on the response variable. Closely related to this, we assume that there is homogeneity within blocks. To say this another way, the units that make up what we could refer to as "blocks" should be more alike within each block than between blocks. Investigators need ancillary information to make sure this is true. Blocking on arbitrary sections of land may not work well if the resulting blocks are not homogeneous. This assumption is directly related to the first assumption we defined; if blocks are not relatively homogeneous it isn't very likely the levels of the blocking factor will have a systematic effect on the response variable.

There is one other assumption we make when using a blocking factor. We believe that the levels of the blocking factor do not influence the effect we want to estimate. This means we believe the true effect of a protocol (i.e., the difference in mean response) or any relationships of interest (i.e., estimated slope) is the same across blocks.

This is a big assumption. If we have reason to believe the blocking factor could influence the effect of interest we would need to carefully design our study to address that. This would be an issue of an interaction between the blocking factor and the variable of interest. We will start talking about interactions next week.

As I described in the soil moisture example, a block is how we often refer to a level of the blocking factor, although using this language is definitely not required. If we blocked on stands, say, we would probably refer to a level of the blocking factor as a stand rather than using the term block. Make sure you pick a term for your study units and stick with it; don't start out calling something stands and then later call them blocks.

To sum up everything in this section, here are the three main assumptions about blocking: 1. The levels of the blocking factor affect the response variable in some systematic way.
2. "Blocks" are more alike within each block than between blocks (i.e., homogeneity within blocks). 3. The levels of the blocking factor do not influence the effect we want to estimate.

2.3 Blocking and independence

Based on the assumptions listed above, do you think observations within a block are independent of each other?

Blocking most often leads to clustering observations in space, although blocks do not have to be contiguous. If the blocking factor really has an effect on the response, which is what the investigators believe if they designed a study with blocking, then within-block observations are *not* independent of each other by design.

In reading 2.2 you learned that one way to address issues of correlation is to include the variable that causes the correlation in the model. In order to meet the assumption of the independence of errors when using a blocked design, the blocking factor must be included in the model.

2.4 Statistical model and analysis

The statistical model for a complete randomized block design can be written as

$$y_{ij} = \mu + \alpha_i + \gamma_j + \epsilon_{ij}$$

where

- y_{ij} is the response for protocol i within block j
- μ is the overall mean
- α_i is the effect of protocol i, with $\sum_i \alpha_i = 0$
- $\gamma_1, \gamma_2, ..., \gamma_b \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\gamma)$
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ for all i and j and mutually independent.

This is the "effects parameterization" or "contrast sum" parameterization discussed in lecture. While we will usually not use this parameterization when fitting models in R, it makes writing down the model a little less cumbersome. The effect of the blocks in the statistical model is represented by γ_j and is defined as the (random) effect of the j^{th} block.

A couple things to note from this model:

1. The errors are still assumed to be independent.

2. We have two variance parameters, σ^2 and σ^2_γ, one for the error variance and one for the variance of the random effects (respectively). Both of these will be estimated when we fit models using R, and we will get practice fitting and interpreting results from CRBD models in lab this week.