Lecture 4:
Linear mixed effects models: random model types and diagnostics

Notes are from Robert Bagchi,
University of Connecticut

Mixed model type problems: single random effects

Very common to have MEEs that employ single random effects

These may differ in their complexity,

from

– random intercept models (each random group has its own intercept estimate)

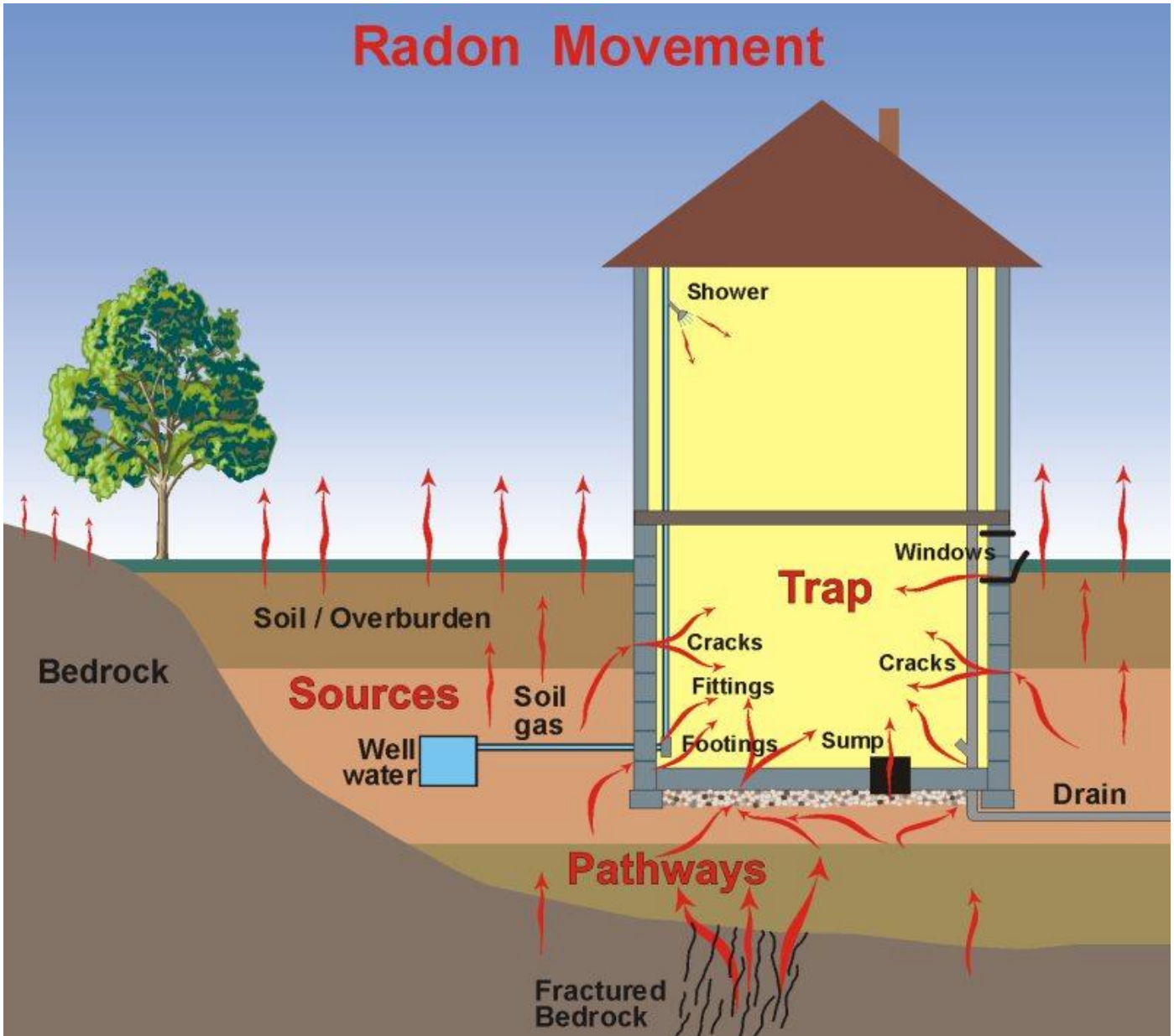
to

– random slope models (each random group has its own slope estimate)

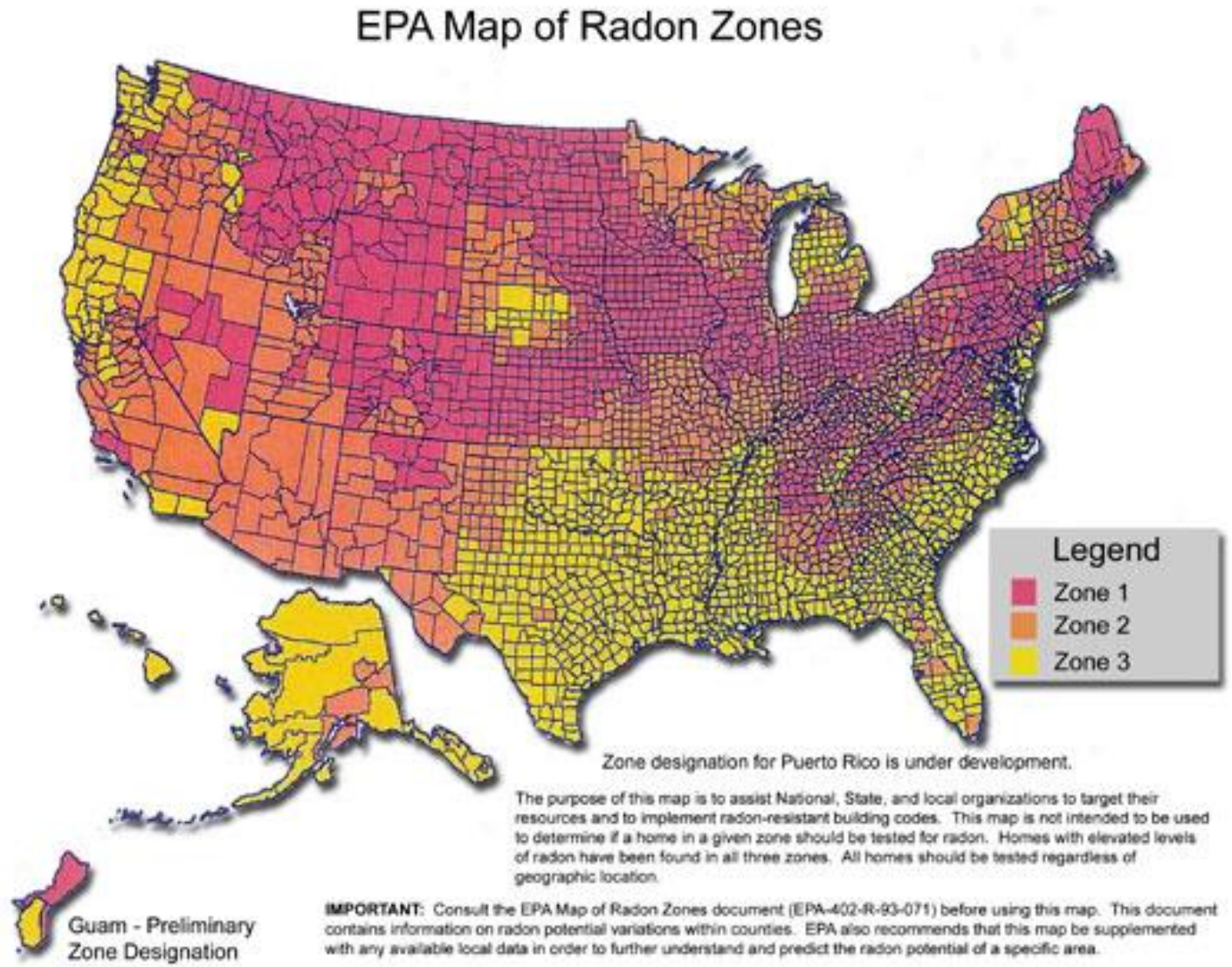
Radon concentrations in US houses

radon is a naturally occurring radioactive chemical element that causes lung cancer

It seeps out of the ground and into our homes



the dataset


EPA Map of Radon Zones
Legend
Zone 1
Zone 2
Zone 3
Zone designation for Puerto Rico is under development.
The purpose of this map is to assist National, State, and local organizations to target their resources and to implement radon-resistant building codes. This map is not intended to be used to determine if a home in a given zone should be tested for radon. Homes with elevated levels of radon have been found in all three zones. All homes should be tested regardless of geographic location.
IMPORTANT: Consult the EPA Map of Radon Zones document (EPA-402-R-93-071) before using this map. This document contains information on radon potential variations within counties. EPA also recommends that this map be supplemented with any available local data in order to further understand and predict the radon potential of a specific area.
Guam - Preliminary Zone Designation

Exercise 4.1

- Read in the radon data set.

- Take a look at the data:
  - What is the range of radon levels?
  - How many counties are there?
  - What does the relationship look like if you plot it? How does it look for different counties?

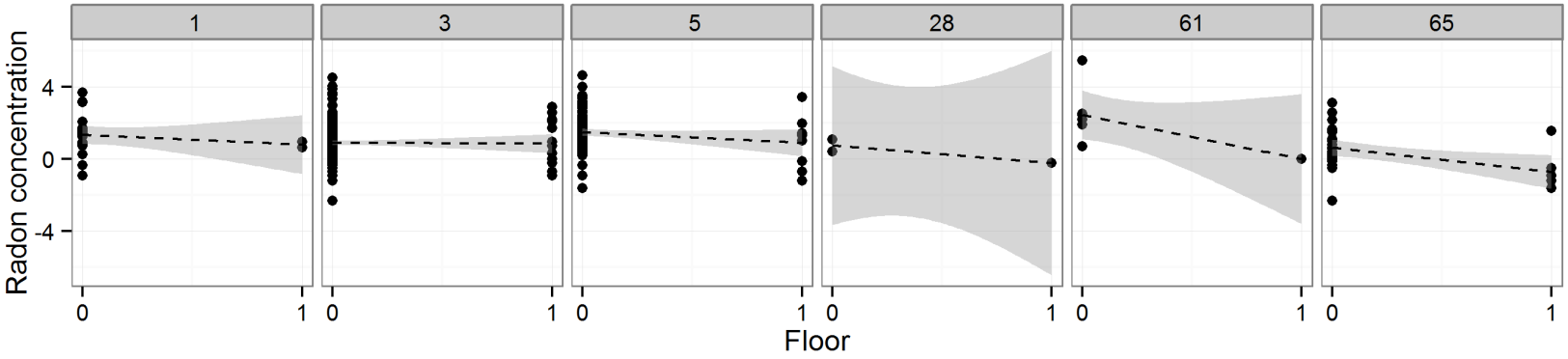
- Fit a linear model to describe the relationship between floor and radon levels.
  - What is the average radon concentration on the ground floor?
  - What is the difference between the ground and first floors?
  - What is the average radon concentration on the first floor?

Fixed vs. random effects

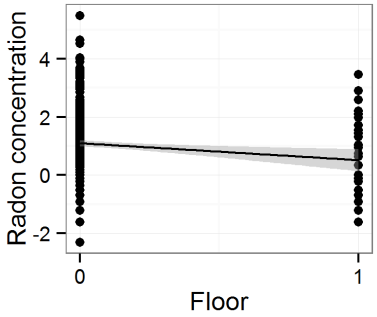
	Fixed effects	Random effects
Design	Treatments specifically chosen.	Treatments randomly chosen as a representative sample of the "population"
Repetition	If you repeated the experiment you'd use the same levels	You'd use different levels drawn from the same "population"
Desired inference	SPECIFICALLY the levels used	The "population" from which the levels are drawn
Parameters	Estimates	Predictions (Best Linear Unbiased Predictors – BLUPS) Include "shrinkage"
Primary Inference	Means	Variances and Covariances

Mixed effect models include both fixed and random effects

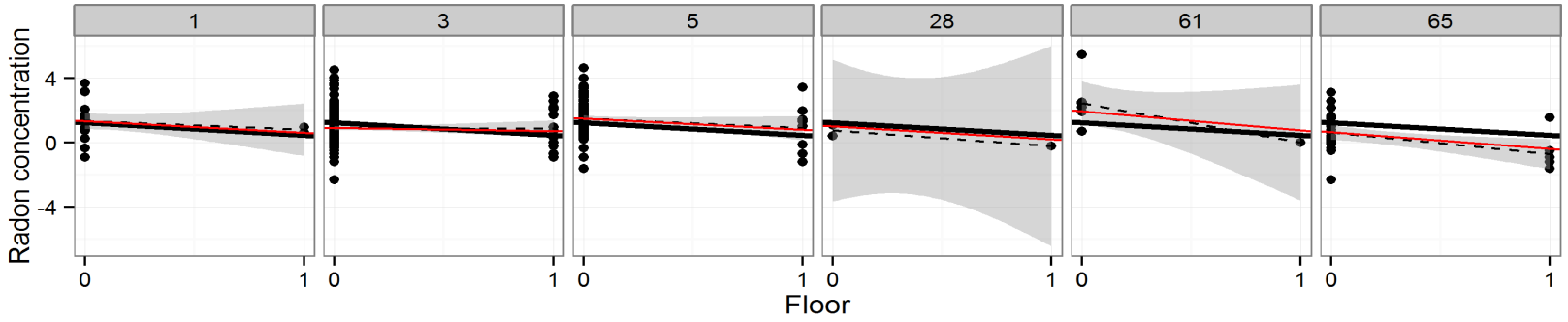
Shrinkage



Complete pooling – all data together – one linear model



Partial pooling –sites with less data to "borrow strength" from the rest of the dataset



Code 4.1

Fitting a mixed effects model with lmer

```
Linear mixed model fit by REML ['lmerMod']
Formula: radon ~ floor + (1 | county)
   Data: radon

Random effects:
 Groups   Name        Variance Std.Dev.
 county   (Intercept) 0.257    0.507
 Residual             1.200    1.095

Number of obs: 2369, groups: county, 68
Fixed effects:
            Estimate    Std. Error t value
(Intercept) 1.3369      0.0698       19.1
floor       -0.7930     0.0751      -10.6

Correlation of Fixed Effects:
      (Intr)
floor -0.125
```

What if the effect of floors varies between counties?

- Perhaps the difference between floors is greater in some counties than others
- To model this, fit a random-slope model
- Modify the previous code like this

lmer(radon ~ floor + (1 + floor|county), …)

Grouping factor is still **county**
The intercept and floor effect are allowed to vary between counties.

Code 4.2

Random slope models with lmer

```
Linear mixed model fit by REML ['lmerMod']
Formula: radon ~ floor + (1 + floor | county)
   Data: radon

REML criterion at convergence: 7282

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.598 -0.656 -0.053  0.587  4.108

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 county   (Intercept) 0.2636   0.513
          floor       0.0366   0.191    -0.35
 Residual             1.1963   1.094

Number of obs: 2369, groups: county, 68

Fixed effects:
            Estimate Std. Error t value
(Intercept)   1.3413     0.0706    19.0
floor        -0.8115     0.0808   -10.1

Correlation of Fixed Effects:
      (Intr)
floor -0.217
```

Assumptions of LMMs

1. The model makes biological sense
2. Additivity and linearity
3. Independence of errors
4. Homoscedasticity
5. Normality of variances– at all levels! (random effects and residual errors)

Can test these with diagnostic plots

Code 4.3
Assessing mixed-effects model fits

Mixed model type problems: Nested random effects

- When you have one grouping factor which is subdivided into smaller ones.
  - Plots within blocks
  - Pots within a shade-house
  - Populations within forests
- Nested means that multiple cases of one group factor are only found in one case of a larger grouping factor
  - E.g. If you move pots between shade-houses, then pot is not nested in shade-house anymore!

Are my data nested?

- Simple check:
  ```
  > table(dat$group1, dat$group2)
  ```
- Levels of group 2 only show up in one level of group 1.

Nested random effects models in lmer

lmer(radon ~ floor + (1|cgroup/county), …)

county is still a grouping factor.
BUT
It is nested in **cgroup**

Alternative formulation:

lmer(radon ~ floor + (1|cgroup) + (1|cgroup:county), …)

Code 4.4

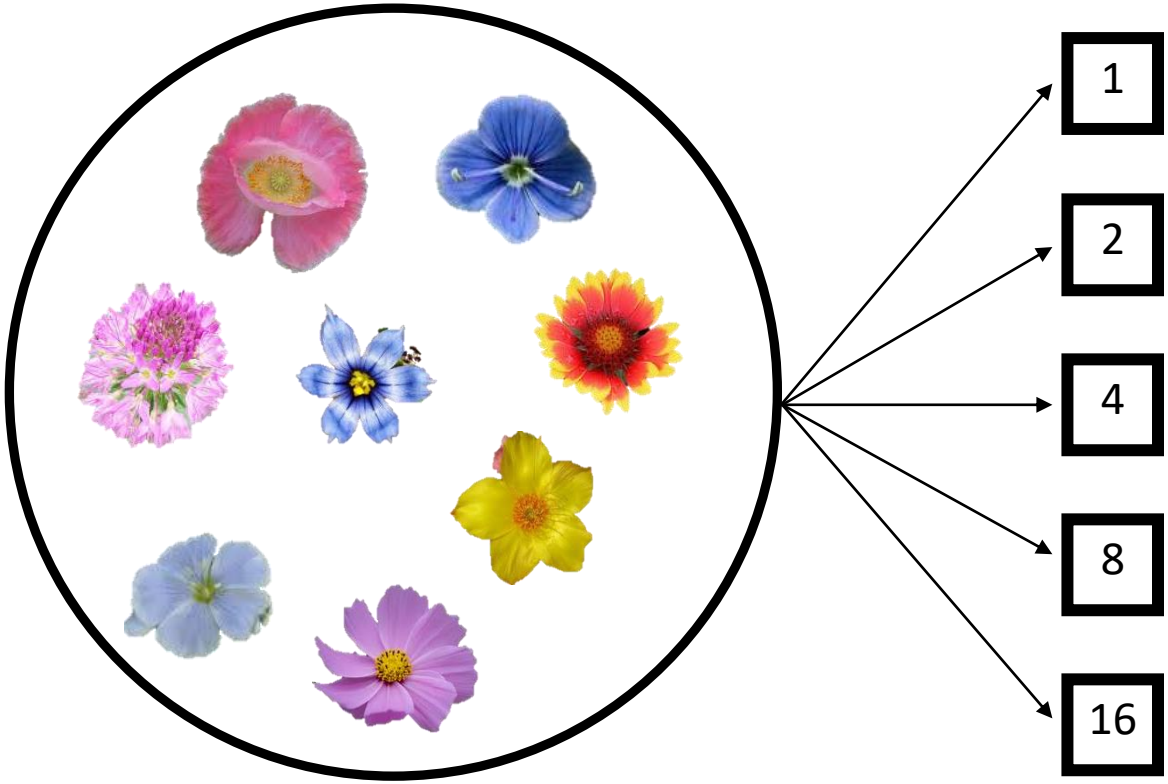
Analysing nested data with lmer

Looking at the output

```
Random effects:
 Groups           Name        Variance Std.Dev.
county:cgroup     (Intercept) 2.57e-01 5.07e-01
cgroup            (Intercept) 8.40e-15 9.16e-08
Residual                      1.20e+00 1.10e+00

Number of obs: 2369, groups: county:cgroup, 68;
cgroup, 10
```

Exercise 4.2:
The Biodepth experiment



http://www.esapubs.org/archive/mono/M075/001/suppl-1.htm

Exercise 4.2

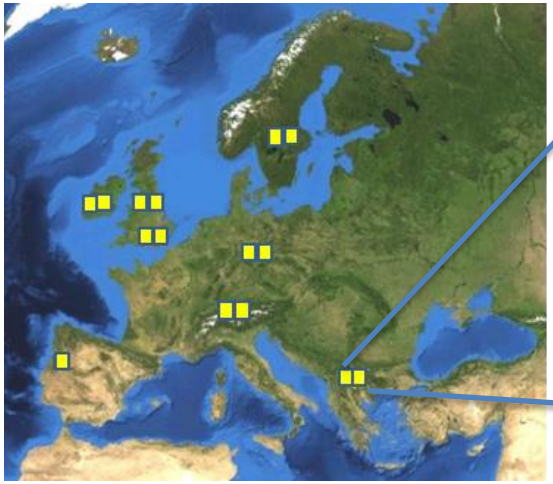
- The biodepth experiment attempted to understand how species richness affected ecosystem functioning in grassland communities, using a repeated sampling design in blocks at several locations in Europe. Here we only consider the effect of species richness on biomass production.

- Load the Biodepth data (Biodepth.csv) and fit a mixed-effects model to answer the following questions:
  1. What is the relationship between biomass and log diversity?
  2. Does this relationship vary among sites?
  3. Check the diagnostics of the model – are there ways in which you might want to improve it

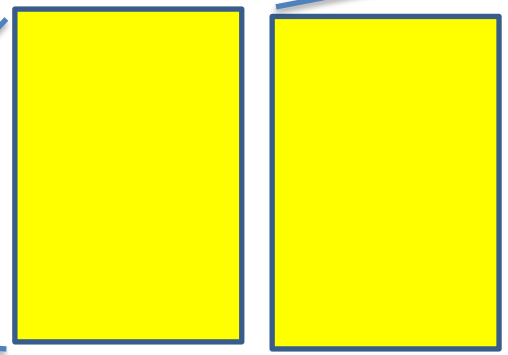
  Ps – there is a problem with the data that will need fixing before you can finish this correctly (clue – we are looking at nested designs!!)

Nested random effects

- Case: $y \sim x_1 + (1|z_1/z_2) + \varepsilon$
- Property: levels of inner random effect (z_2) only occur in one level of outer random effect (z1)



Country (Z_1)



Blocks (Z_2)

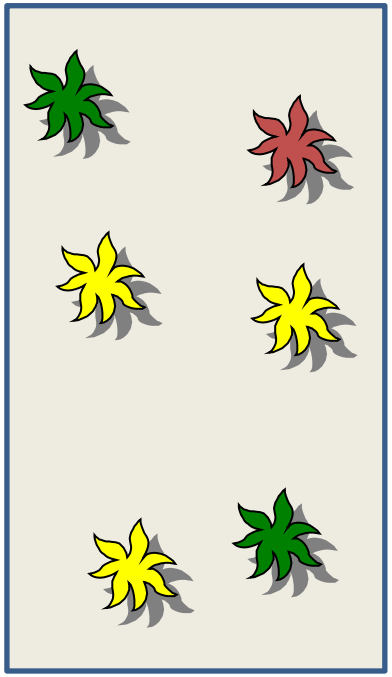
! Every block must have a unique ID

Treatments (x_1)

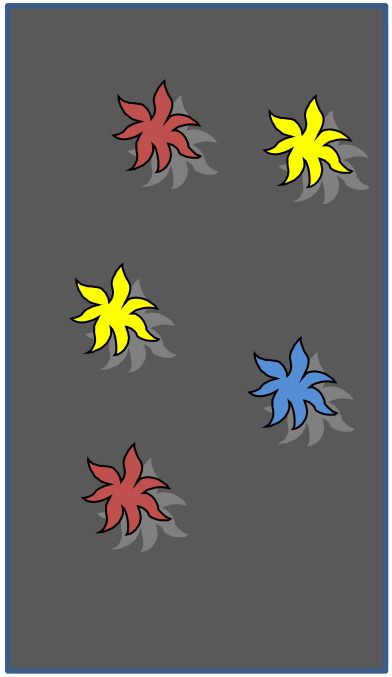
- Design: Growth ~ Treatment + (1|Country/Block)

A crossed random effects problem
Effects of soil type on *Shorea* growth

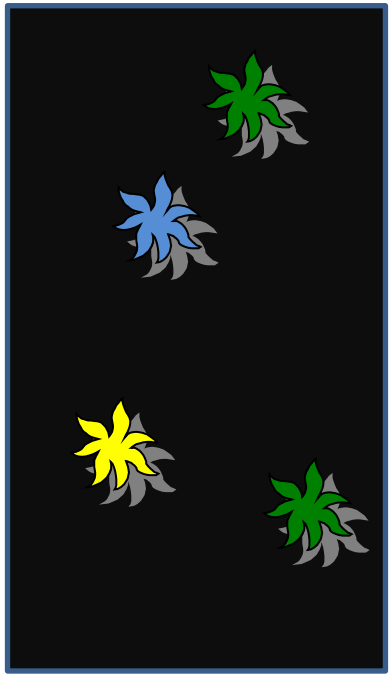
x10 plots on each soil type!



Alluvial



Mudstone



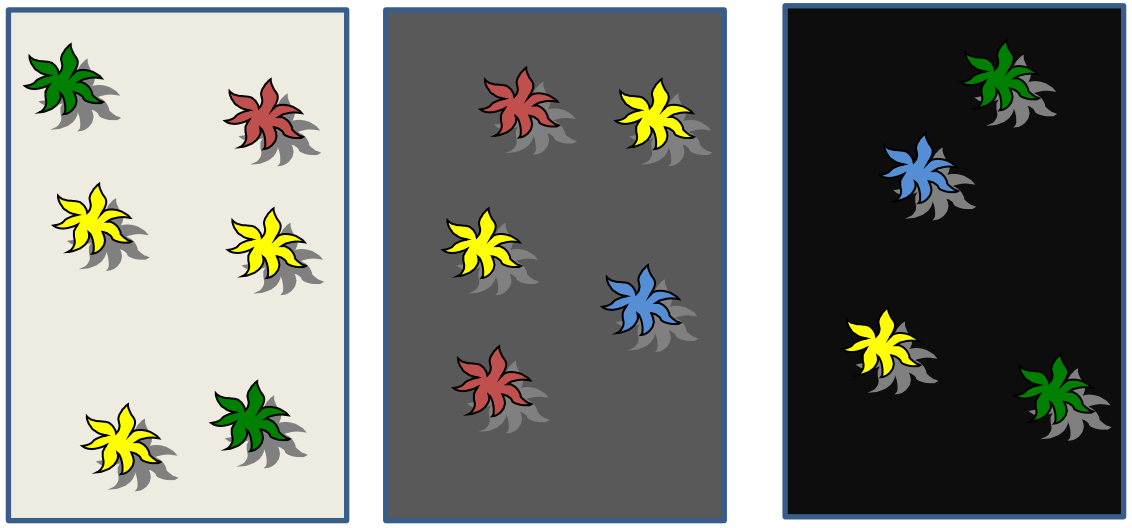
Sandstone

- 14 species of *Shorea* seedlings used
- Species occur unevenly across plots

Crossed random effects

- Case: $y \sim x_1 + (1|z_1) + (1|z_2) + \varepsilon$
- Property: Random effects are not strictly nested

x10 plots on each soil type



Alluvial Mudstone Sandstone

- 14 species of seedlings
- Species found across plots
- species and plots are random effects

- Design: Growth ~ Soil + (1|Species) + (1|Plot)

Exercise 4.3

- Read in and analyse the data in the file plantsoil.csv.
- Check the diagnostics.
- Is there an effect of soil type on plant growth?
- Which soil supports the highest growth rate?
- Which has the lowest growth rate?
- Do differences in growth between species across soil types explain a significant amount of variation in the data?

End of Lecture 4