**Uncertainty Workshop Module 2: *Experimental Design for Long-Term Monitoring***

**Hands-on exercise**

In advance of the meeting, please download the R program, which is freely available for Mac, Windows, or Linux machines. You may use the base R program and console (<https://cran.r-project.org/>), or download RStudio Desktop (<https://www.rstudio.com/products/rstudio/download2/>).

You will also need to download several packages installed in order to complete the exercise. To install a package from the R GUI, select Packages | Install Packages. Select a CRAN mirror geographically close to our location. Then scroll down to find the package you are interested in and click ‘OK’. From RStudio, select Tools |Install Packages. Type the package names (ensure that “Install dependencies” is checked) and click on Install. The packages you will need are: lsmeans, car, and nlme.

1. We are interested in monitoring biomass in a certain natural area with four different vegetation types. We have set up a research design whereby we have n=15 replicate plots in each of the k=4 vegetation types. In the initial year of the study, we are simply interested in knowing if the amount of biomass differs by vegetation type. The data for the initial year of study are provided in the file data\_1.csv.
2. Formulate a linear model to test the hypothesis that biomass is approximately equal in each vegetation type.



Hint: use the function lme (in package nlme). No random effects are necessary for this analysis.

1. Is there a significant overall effect of vegetation?
2. What are the model effect estimates?
3. Do your data meet the assumptions necessary to interpret tests of significance?
Hint: produce diagnostic plots.
4. Estimate the marginal means of each vegetation type using the function lsmeans:

> lsmeans(*lmobject*,pairwise~vegtype)
The option ‘pairwise’ performs a post-hoc mean comparison using the Tukey method. Take note of the standard errors for each mean, and which pairs of means are significantly different.

1. We continue our monitoring experiment for four years, so that we have m=4 annual measurements on each of n=15 replicate plots in k=4 vegetation types. The data for the four years are in the file data\_2.csv.
2. Formulate a linear model to test the hypothesis that biomass is approximately equal in each vegetation type over time.

Hint: A random effect is necessary to account for possible correlations among measurements taken on the same replicate over time.

1. Is there a significant effect of vegetation that differs by time (i.e., an interaction)?
2. Do your data meet the assumptions necessary to interpret tests of significance?
3. Estimate the marginal means for each level of each significant effect. Take note of the standard errors for each mean.
4. Your colleague argues that the replicate plots were not set up in a completely randomized design, but instead that they constitute 15 blocks, with each treatment represented once in each block. (The data for this is the same as what was used in question 2, except that the plots are actually grouped in blocks, in the Randomized Complete Block sense. (Data are in the file data\_3.csv.)
5. Formulate a linear model to test the hypothesis that biomass is approximately equal in each vegetation type.
Hint: A random effect is necessary to account for grouping measurements taken in the same vegetation type, nested within block (plot).



1. Check the anova and summary output to ensure the degrees of freedom make sense.
2. Was blocking a good idea?

4. Now consider that each plot within a block has been measured over time. The data are in file data\_4.csv.

1. Formulate a linear model to test the hypothesis that biomass is approximately equal in each vegetation type over time.

Hint: A random effect is necessary to account for times nested within vegetation type, nested within block (plot).

1. Is there a significant effect of vegetation that differs by time (i.e., an interaction)?
2. Do your data meet the assumptions necessary to interpret tests of significance?
3. Estimate the marginal means for each level of each significant effect. Is the uncertainty estimated the same as in #2?
4. How would the study have differed if we had instead made a design whereby we selected plots, and then surveyed sub-samples within those plots? Our design consisted of biomass samples taken in k=4 vegetation types, each with three replicate observations and 5 sub-samples within replicate. The data for the first year of measurement are in the file data\_5.csv.
5. Formulate a linear model to test the hypothesis that biomass is approximately equal in each vegetation type.



Hint: A random effect is necessary to account for grouping subsamples taken in the same replicate plot.

1. Check the anova and summary output to ensure the degrees of freedom make sense.
2. Are your results different from #1 or #3?
3. Now consider that each plot with sub-samples was measured in each of four years. The data for all years of measurement are in the file data\_6.csv.
4. Formulate a linear model to test the hypothesis that biomass is approximately equal in each vegetation type over time.

Hint: A random effect is necessary to account for times nested within subsamples nested in replicate plots.

1. Is there a significant effect of vegetation that differs by time (i.e., an interaction)?
2. Do your data meet the assumptions necessary to interpret tests of significance?
3. Estimate the marginal means for each level of each significant effect. Take note of the standard errors for each mean. Is the uncertainty estimated the same as in #2**?**
4. Consider that different correlation structure may be appropriate for some data. Observations taken on the same experimental unit may be correlated, and that the correlation may depend on how close together the observations are in time. Further, a random slope as well as intercept may be necessary to include all appropriate sources of variation
5. Consider #2 and add an AR(1) structure. Does this model have a better fit?
Hint: use the function AIC
6. Consider #2 and add a random slope associated with time. Does this model have a better fit?
Hint: the random option will have ‘time’ where there was an intercept denoted (‘1’)
7. Consider #4 and add an AR(1) structure. Does this model have a better fit?
8. Consider #6 and add an AR(1) structure. Does this model have a better fit?

R code

## Question 1 ##

lm1 <-lm(biomass ~ vegtype, data=data1)

anova(lm1)

summary(lm1)

plot(lm1)

library(car)

leveneTest(biomass ~ vegtype, data=data1)

library(lsmeans)

lsmeans(lm1, pairwise ~ vegtype)

## Question 2 ##

library(nlme)

data2$time <- as.factor(data2$time)

lme2 <- lme(ann.biomass ~ vegtype\*time, random = ~1|plot, data=data2)

summary(lme2)

anova(lme2)

plot(lme2)

lsmeans(lme2, pairwise ~ vegtype)

lsmeans(lme2, pairwise ~ time)

plot(lsmeans(lme2, ~vegtype:time))

## Question 3 ##

data3$block <- as.factor(data3$block)

library(nlme)

lme3 <-lme(biomass ~ vegtype, random = ~1|block/vegtype, data=data3)

anova(lme3)

summary(lme3)

plot(lme3)

qqnorm(lme3, ~ranef(., level=1))

qqnorm(lme3, ~ranef(., level=2))

lsmeans(lme3, pairwise ~ vegtype)

## Question 4 ##

lme4 <- lme(ann.biomass ~ vegtype\*time, random =~1|block/vegtype, data=data4)

anova(lme4)

summary(lme4)

plot(lme4)

lsmeans(lme3, ~ vegtype:time)

## Question 5 ##

data5$plot5 <- as.factor(data5$plot5)

lme5 <-lme(biomass ~ vegtype, random = ~1|plot5, data=data5)

anova(lme5)

summary(lme5)

plot(lme5)

qqnorm(lme5, ~ranef(., level=1))

lsmeans(lme5, pairwise ~ vegtype)

## Question 6 ##

lme6 <- lme(ann.biomass ~ vegtype\*time, random =~1|plot5/vegtype, data=data6)

anova(lme6)

summary(lme6)

plot(lme6)

## Question 7 ##

lme2.ar <- lme(ann.biomass~ vegtype\*time, random = ~1|plot, corr=corAR1(), data=data2)

anova(lme2.ar)

AIC(lme2.ar)

lme2.rs <- lme(ann.biomass ~ vegtype2\*time, random = ~time|plot2, data=data2)

anova(lme2.rs)

AIC(lme2.rs)