Statistics Case Study: Cross Fell Grazing 2013

CK

24 August 2017

### Packages

library(tidyverse)

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: readr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

### Import data

Import the csv file which I previously opened in Minitab and saved (probably wrongly in Unicode)

raw <- read.csv(file = "CF\_COVER\_2013.txt", sep = ",", quote = "", header = T, skipNul = T)

### Wrangle data

#### Names

lets have a look at the names:

names(raw)

## [1] "ÿþ.Vmb.03.." "X.Vmb.05.." "X.Vmb.08.." "X.Vmb.10.." "X.Vmb13.."   
## [6] "X.C6." "X.C7.T." "X.moss.03.." "X.moss.05.." "X.moss.08.."  
## [11] "X.moss.10.." "X.moss.13.." "X.lmoss03.." "X.lmoss05.." "X.lmoss08.."  
## [16] "X.lmoss10.." "X.lmoss13.." "X.C18." "X.C19.T." "X.DV.03."   
## [21] "X.DV.05." "X.DV.08." "X.DV.10." "X.DV.13." "X.C25."   
## [26] "X.C26.T."

They are a bit of a mess.  
We can strip out a lot automatically, and replace some of the more exotic characters. Here I will remove the X. and the periods, and replace the first column name entirely.

names(raw) <- gsub("X.", "", names(raw))  
names(raw) <- gsub("\\.", "", names(raw))  
names(raw)[1] <- "Vmb03"  
names(raw)

## [1] "Vmb03" "Vmb05" "Vmb08" "Vmb10" "Vmb13" "C6" "C7T"   
## [8] "moss03" "moss05" "moss08" "moss10" "moss13" "lmoss03" "lmoss05"  
## [15] "lmoss08" "lmoss10" "lmoss13" "C18" "C19T" "DV03" "DV05"   
## [22] "DV08" "DV10" "DV13" "C25" "C26T"

#### Strip

We've got some unnecessary summary columns: They all start with C. We'll strip them out:

raw$C6 <- NULL  
raw$C7T <- NULL  
raw$C18 <- NULL  
raw$C25 <- NULL  
raw$C19T <- NULL  
raw$C26T <- NULL  
names(raw)

## [1] "Vmb03" "Vmb05" "Vmb08" "Vmb10" "Vmb13" "moss03" "moss05"   
## [8] "moss08" "moss10" "moss13" "lmoss03" "lmoss05" "lmoss08" "lmoss10"  
## [15] "lmoss13" "DV03" "DV05" "DV08" "DV10" "DV13"

#### Make tidy

Now lets put this into **tidy** format:

each column a variable, each row an observation.

In this case the observations are of percent cover, and they key (currently column names) define a taxon and a year.

data <- gather(data = raw, key = "tax.surv", value = "cover", na.rm = TRUE, factor\_key = TRUE)  
head(data) #have a look

## tax.surv cover  
## 1 Vmb03 0.0  
## 2 Vmb03 0.5  
## 3 Vmb03 0.0  
## 4 Vmb03 0.5  
## 5 Vmb03 0.0  
## 6 Vmb03 0.0

dim(data) #get dimensions

## [1] 774 2

levels(data$tax.surv) #check all the variables are there

## [1] "Vmb03" "Vmb05" "Vmb08" "Vmb10" "Vmb13" "moss03" "moss05"   
## [8] "moss08" "moss10" "moss13" "lmoss03" "lmoss05" "lmoss08" "lmoss10"  
## [15] "lmoss13" "DV03" "DV05" "DV08" "DV10" "DV13"

Now we need to split the tax.surv column.

data <- separate(data = data, col = tax.surv, into = c("taxon", "year"), sep = -3)  
data$taxon <- as.factor(data$taxon)  
levels(data$taxon)

## [1] "DV" "lmoss" "moss" "Vmb"

head(data)

## taxon year cover  
## 1 Vmb 03 0.0  
## 2 Vmb 03 0.5  
## 3 Vmb 03 0.0  
## 4 Vmb 03 0.5  
## 5 Vmb 03 0.0  
## 6 Vmb 03 0.0

summary(data)

## taxon year cover   
## DV :193 Length:774 Min. : -2.996   
## lmoss:195 Class :character 1st Qu.: 0.000   
## moss :195 Mode :character Median : 2.000   
## Vmb :191 Mean : 6.827   
## 3rd Qu.: 5.000   
## Max. :100.000

Finally sort out the year column:

data$year[which(data$year=="03")] <- 2003  
data$year[which(data$year=="05")] <- 2005  
data$year[which(data$year=="08")] <- 2008  
data$year[which(data$year=="10")] <- 2010  
data$year[which(data$year=="13")] <- 2013  
  
data$year <- as.factor(data$year)  
levels(data$year) <- c("2003", "2005", "2008", "2010", "2013")

we good?

summary(data)

## taxon year cover   
## DV :193 2003:167 Min. : -2.996   
## lmoss:195 2005:175 1st Qu.: 0.000   
## moss :195 2008: 80 Median : 2.000   
## Vmb :191 2010:131 Mean : 6.827   
## 2013:221 3rd Qu.: 5.000   
## Max. :100.000

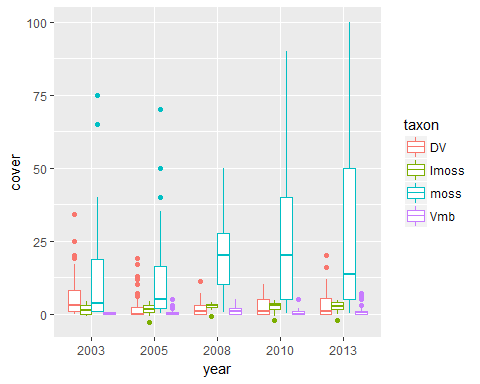
we good.

**outstanding issue: negative cover values. What's that about? Need to look at minitab files again, and maybe take up with DM.**\*

### Exploratory data anlysis

ok, lets plot:

ggplot(data = data, mapping = aes(y = cover, x = year)) +  
 geom\_boxplot(aes(colour = taxon))



Looks good. moss is most common, some sort of pattern maybe.  
Lets look at means:

data.means <- data %>%  
 group\_by(taxon, year) %>%   
 summarise(mean = mean(cover), median = median(cover), sd = sd(cover))

## Warning: package 'bindrcpp' was built under R version 3.3.3

data.means

## # A tibble: 20 x 5  
## # Groups: taxon [?]  
## taxon year mean median sd  
## <fctr> <fctr> <dbl> <dbl> <dbl>  
## 1 DV 2003 6.3902439 3.000000 7.965168  
## 2 DV 2005 2.5909091 0.000000 4.721638  
## 3 DV 2008 2.1000000 1.000000 2.789076  
## 4 DV 2010 2.2727273 1.000000 3.023582  
## 5 DV 2013 3.6181818 1.000000 5.104431  
## 6 lmoss 2003 1.3736807 1.242453 1.657737  
## 7 lmoss 2005 1.6644996 1.609438 1.542746  
## 8 lmoss 2008 2.7519709 2.995732 1.095291  
## 9 lmoss 2010 2.3349150 2.995732 1.979228  
## 10 lmoss 2013 2.4631652 2.596478 1.746859  
## 11 moss 2003 12.9523810 3.500000 19.717149  
## 12 moss 2005 12.3420455 5.000000 14.988100  
## 13 moss 2008 22.4750000 20.000000 15.455773  
## 14 moss 2010 26.8242424 20.000000 24.660951  
## 15 moss 2013 29.1482143 13.500000 30.589778  
## 16 Vmb 2003 0.1428571 0.000000 0.228615  
## 17 Vmb 2005 0.6860465 0.000000 1.336410  
## 18 Vmb 2008 1.0500000 1.000000 1.223670  
## 19 Vmb 2010 0.7968750 0.000000 1.325065  
## 20 Vmb 2013 1.0185185 0.500000 1.764329