Statistics Case Study: Cross Fell Grazing 2013

CK

24 August 2017

### Time run

Time run: 2017-08-25 14:29:14

### Packages

library(tidyverse)  
library(lubridate)  
library(foreign)  
library(knitr)

## Data preparation

### Import data - cover part

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 - cover

#### 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)

#### Strip

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

raw[,grep("C", names(raw))] <- NULL  
#names(raw)

#### 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.

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

## [1] 774 2

levels(cov.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.

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

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

head(cov.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(cov.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:

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

we good?

summary(cov.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.**

### Import data - height

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

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

### Wrangle data - height

#### Names

lets have a look at the names:

names(raw)

## [1] "ÿþ.mossht.03." "X.mossht.05." "X.mossht.08." "X.mossht.10."   
## [5] "X.mossht.13." "X.C6." "X.C7.T." "X.Vm.ht.03."   
## [9] "X.Vm.ht.05." "X.Vm.ht.08." "X.Vm.ht.10." "X.Vm.ht.13."   
## [13] "X.C13." "X.C14.T." "X.Fo.ht.03." "X.Fo.ht.05."   
## [17] "X.Fo.ht.08." "X.Fo.ht.10." "X.Fo.ht.13." "X.C20."   
## [21] "X.C21.T." "X.Cx.ht.03." "X.Cx.ht.05." "X.Cx.ht.08."   
## [25] "X.Cx.ht.10." "X.Cx.ht.13." "X.C27." "X.C28.T."   
## [29] "X.grm.ht.03." "X.grm.ht.05." "X.grm.ht.08." "X.grm.ht.10."   
## [33] "X.grm.ht.13." "X.C34." "X.C35.T." "X.C218.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) <- gsub("ht", "", names(raw))  
names(raw)[1] <- "moss03"  
#names(raw)

#### Strip

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

raw[,grep("C", names(raw))] <- NULL  
#names(raw)

#### 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.

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

## tax.surv height  
## 1 moss03 2.5  
## 2 moss03 4.5  
## 3 moss03 3.0  
## 4 moss03 2.0  
## 5 moss03 3.0  
## 6 moss03 4.0

dim(ht.data) #get dimensions

## [1] 1574 2

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

## [1] "moss03" "moss05" "moss08" "moss10" "moss13" "Vm03" "Vm05"   
## [8] "Vm08" "Vm10" "Vm13" "Fo03" "Fo05" "Fo08" "Fo10"   
## [15] "Fo13" "grm03" "grm05" "grm08" "grm10" "grm13"

Now we need to split the tax.surv column.

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

## [1] "Fo" "grm" "moss" "Vm"

head(ht.data)

## taxon year height  
## 1 moss 03 2.5  
## 2 moss 03 4.5  
## 3 moss 03 3.0  
## 4 moss 03 2.0  
## 5 moss 03 3.0  
## 6 moss 03 4.0

summary(ht.data)

## taxon year height   
## Fo :385 Length:1574 Min. : 0.500   
## grm :757 Class :character 1st Qu.: 2.500   
## moss:188 Mode :character Median : 3.500   
## Vm :244 Mean : 4.086   
## 3rd Qu.: 5.000   
## Max. :17.000

Finally sort out the year column:

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

*we good?*

summary(ht.data)

## taxon year height   
## Fo :385 2003:361 Min. : 0.500   
## grm :757 2005:406 1st Qu.: 2.500   
## moss:188 2008:181 Median : 3.500   
## Vm :244 2010:235 Mean : 4.086   
## 2013:391 3rd Qu.: 5.000   
## Max. :17.000

**we good.**

### Import data - all quadrat data

Import minitab file

#raw <- read.table("cross fell cover all.MTP", quote="", comment.char="", skip = 2)

*not working, have asked DM for csv*

### Merge data

*we can't do this at the moment because we don't have the quadrat data*

### Contextual data

#### Quadrats

Number of quadrats surveyed per year (taken from report)

quad.n <- data.frame(year = c("2003", "2005", "2008", "2010", "2013"), n.total = c(42, 44, 20, 31, 56))

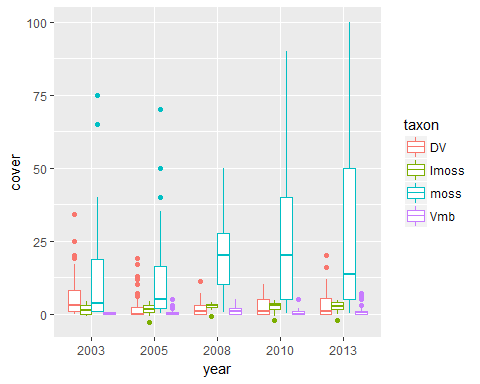
## 

## Data analysis

### Exploratory data anlysis

ok, lets plot cover values:

ggplot(data = cov.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:

cov.means <- cov.data %>%  
 group\_by(taxon, year) %>%   
 summarise(mean = mean(cover), median = median(cover), sd = sd(cover), present = length(cover[which(cover!=0)]), n = length(cover))

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

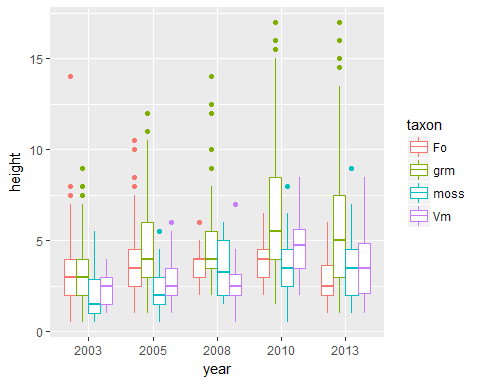
kable(cov.means, digits = 2, caption = "Mean percent cover")

Mean percent cover

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| taxon | year | mean | median | sd | present | n |
| DV | 2003 | 6.39 | 3.00 | 7.97 | 31 | 41 |
| DV | 2005 | 2.59 | 0.00 | 4.72 | 19 | 44 |
| DV | 2008 | 2.10 | 1.00 | 2.79 | 14 | 20 |
| DV | 2010 | 2.27 | 1.00 | 3.02 | 18 | 33 |
| DV | 2013 | 3.62 | 1.00 | 5.10 | 32 | 55 |
| lmoss | 2003 | 1.37 | 1.24 | 1.66 | 36 | 42 |
| lmoss | 2005 | 1.66 | 1.61 | 1.54 | 38 | 44 |
| lmoss | 2008 | 2.75 | 3.00 | 1.10 | 20 | 20 |
| lmoss | 2010 | 2.33 | 3.00 | 1.98 | 32 | 33 |
| lmoss | 2013 | 2.46 | 2.60 | 1.75 | 53 | 56 |
| moss | 2003 | 12.95 | 3.50 | 19.72 | 42 | 42 |
| moss | 2005 | 12.34 | 5.00 | 14.99 | 44 | 44 |
| moss | 2008 | 22.48 | 20.00 | 15.46 | 20 | 20 |
| moss | 2010 | 26.82 | 20.00 | 24.66 | 33 | 33 |
| moss | 2013 | 29.15 | 13.50 | 30.59 | 56 | 56 |
| Vmb | 2003 | 0.14 | 0.00 | 0.23 | 12 | 42 |
| Vmb | 2005 | 0.69 | 0.00 | 1.34 | 19 | 43 |
| Vmb | 2008 | 1.05 | 1.00 | 1.22 | 13 | 20 |
| Vmb | 2010 | 0.80 | 0.00 | 1.33 | 14 | 32 |
| Vmb | 2013 | 1.02 | 0.50 | 1.76 | 28 | 54 |

And lets plot vegetation heights:

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



Height means:

ht.means <- ht.data %>%  
 group\_by(taxon, year) %>%   
 summarise(mean = mean(height), median = median(height), sd = sd(height), present = length(height[which(height!=0)]), n = length(height)) %>%   
 mutate\_if(is.numeric, funs(round(., 2)))  
kable(ht.means, digits = 2, caption = "Mean vegetation height")

Mean vegetation height

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| taxon | year | mean | median | sd | present | n |
| Fo | 2003 | 3.11 | 3.00 | 1.87 | 131 | 131 |
| Fo | 2005 | 3.78 | 3.50 | 1.71 | 134 | 134 |
| Fo | 2008 | 3.88 | 4.00 | 0.92 | 37 | 37 |
| Fo | 2010 | 3.89 | 4.00 | 1.19 | 35 | 35 |
| Fo | 2013 | 2.86 | 2.50 | 1.13 | 48 | 48 |
| grm | 2003 | 3.28 | 3.00 | 1.66 | 167 | 167 |
| grm | 2005 | 4.45 | 4.00 | 2.24 | 176 | 176 |
| grm | 2008 | 5.04 | 4.00 | 2.63 | 80 | 80 |
| grm | 2010 | 6.45 | 5.50 | 3.43 | 130 | 130 |
| grm | 2013 | 5.75 | 5.00 | 3.31 | 204 | 204 |
| moss | 2003 | 1.95 | 1.50 | 1.30 | 42 | 42 |
| moss | 2005 | 2.37 | 2.00 | 1.13 | 43 | 43 |
| moss | 2008 | 3.48 | 3.25 | 1.59 | 20 | 20 |
| moss | 2010 | 3.48 | 3.50 | 1.66 | 30 | 30 |
| moss | 2013 | 3.56 | 3.50 | 1.77 | 53 | 53 |
| Vm | 2003 | 2.40 | 2.50 | 0.94 | 21 | 21 |
| Vm | 2005 | 2.88 | 2.50 | 1.15 | 53 | 53 |
| Vm | 2008 | 2.58 | 2.50 | 1.21 | 44 | 44 |
| Vm | 2010 | 4.66 | 4.75 | 1.73 | 40 | 40 |
| Vm | 2013 | 3.53 | 3.50 | 1.53 | 86 | 86 |

Now lets recreate the table of mean heights from the report (Table 1)

Table1 <- spread(data = select(.data = ht.means, taxon, year, mean),   
 key = year,   
 value = mean)   
knitr::kable(Table1, digits = 2, caption = "Table 1: Mean height of key species groups measured in each survey year")

Table 1: Mean height of key species groups measured in each survey year

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| taxon | 2003 | 2005 | 2008 | 2010 | 2013 |
| Fo | 3.11 | 3.78 | 3.88 | 3.89 | 2.86 |
| grm | 3.28 | 4.45 | 5.04 | 6.45 | 5.75 |
| moss | 1.95 | 2.37 | 3.48 | 3.48 | 3.56 |
| Vm | 2.40 | 2.88 | 2.58 | 4.66 | 3.53 |

Lets try to calculate frequencies

freq <- cov.means[,c("taxon", "year", "present", "n")]  
freq <- full\_join(freq, quad.n, by = "year")  
freq$freq.n.total <- freq$present / freq$n.total  
freq$freq.n <- freq$present / freq$n  
kable(freq, digits = 2, caption = "Frequency: number and proportion of quadrats in which taxon was found")

Frequency: number and proportion of quadrats in which taxon was found

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| taxon | year | present | n | n.total | freq.n.total | freq.n |
| DV | 2003 | 31 | 41 | 42 | 0.74 | 0.76 |
| DV | 2005 | 19 | 44 | 44 | 0.43 | 0.43 |
| DV | 2008 | 14 | 20 | 20 | 0.70 | 0.70 |
| DV | 2010 | 18 | 33 | 31 | 0.58 | 0.55 |
| DV | 2013 | 32 | 55 | 56 | 0.57 | 0.58 |
| lmoss | 2003 | 36 | 42 | 42 | 0.86 | 0.86 |
| lmoss | 2005 | 38 | 44 | 44 | 0.86 | 0.86 |
| lmoss | 2008 | 20 | 20 | 20 | 1.00 | 1.00 |
| lmoss | 2010 | 32 | 33 | 31 | 1.03 | 0.97 |
| lmoss | 2013 | 53 | 56 | 56 | 0.95 | 0.95 |
| moss | 2003 | 42 | 42 | 42 | 1.00 | 1.00 |
| moss | 2005 | 44 | 44 | 44 | 1.00 | 1.00 |
| moss | 2008 | 20 | 20 | 20 | 1.00 | 1.00 |
| moss | 2010 | 33 | 33 | 31 | 1.06 | 1.00 |
| moss | 2013 | 56 | 56 | 56 | 1.00 | 1.00 |
| Vmb | 2003 | 12 | 42 | 42 | 0.29 | 0.29 |
| Vmb | 2005 | 19 | 43 | 44 | 0.43 | 0.44 |
| Vmb | 2008 | 13 | 20 | 20 | 0.65 | 0.65 |
| Vmb | 2010 | 14 | 32 | 31 | 0.45 | 0.44 |
| Vmb | 2013 | 28 | 54 | 56 | 0.50 | 0.52 |

Problems: \* the reported quadrat number for 2010 (31) is lower than the max number of observations we have (33). Will need to check the full data when we get it. For the time being we'll use freq.n the total number of observations calculated from the number of records (including 0 values), rather than n the reported number.

* the data doesn't quite tally with the report, e.g. bilberry in 2013 was found in 28 quadrats, but report says 30.

TableFreq <- spread(data = select(.data = freq, taxon, year, freq.n),   
 key = year,   
 value = freq.n)  
knitr::kable(TableFreq, digits = 2, caption = "Frequency: proportion of quadrats in which taxon was found")

Frequency: proportion of quadrats in which taxon was found

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| taxon | 2003 | 2005 | 2008 | 2010 | 2013 |
| DV | 0.76 | 0.43 | 0.70 | 0.55 | 0.58 |
| lmoss | 0.86 | 0.86 | 1.00 | 0.97 | 0.95 |
| moss | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Vmb | 0.29 | 0.44 | 0.65 | 0.44 | 0.52 |