Question 1:

Divide the saana.csv data randomly into two different datasets: ??? model calibration data (70%) ??? model evaluation data (30%) Build the models based on the calibration data and test the predictive performance of the models using the evaluation data. What is the predictive performance of the GLM, GAM and GBM models for Betnan, Cashyp, Empher and Salret based on AUC-values of the model evaluation data? Use mesotopo, soil_moist, soil_temp and soil_ph as predictors. Report the results in one short paragraph (max 5 sentences). Note: you can use sample-function to divide the data, e.g.

```
#load data
data<- read.csv("C:/Users/oyeda/Desktop/MODELLING_PHYSICAL_GEOGRAPHY/assignment3/Data-
20171114 (1)/saana.csv",sep=";")

# Use the caTools package to extract the AUC values and compare them
library(caTools)
library(mgcv)
library(gbm)
```

Betnan

```
#number of times to repeat the models
{rep<-10
bet_auc_glm<-bet_auc_gam<-bet_auc_gbm<-c()
for (i in 1:rep){
  #print(i)
  #sample all the rows, and keep 70%(0.7)
  rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
  cal<- data[rand_sam,]   #get the 70% rows for calibration
  eva<- data[-rand_sam,]  #get the remaining 30% for evaluation

  #create the glm for Betnan occurences
  bet_glm<-glm(Betnan~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,family ="binomial")
  #these could be used to select the Betnan but not necessary anymore. I used eva$Betnan
instead
  #eva_bet<- eva[,grep("Betnan", colnames(data))]
  #which(colnames(data)=="Betnan") or grep("Betnan", colnames(data))
  pred_bet_glm<-predict.glm(bet_glm, newdata = eva, type = "response")
  #check the AUC of the compared prediction and evaluation
  bet_auc_glm_p<-colAUC(pred_bet_glm, eva$Betnan, plotROC=F)
  bet_auc_glm <- c(bet_auc_glm, bet_auc_glm_p[[1]])

  #GAM
  bet_gam<-gam(Betnan~s(mesotopo, k=3) + s(soil_moist, k=3) + s(soil_temp, k=3) +
                 s(soil_ph, k=3), data=cal,family ="binomial")
  pred_bet_gam<-predict.gam(bet_gam, newdata = eva, type = "response")
  bet_auc_gam_p<-colAUC(pred_bet_gam, eva$Betnan, plotROC=F)
  bet_auc_gam <- c(bet_auc_gam, bet_auc_gam_p[[1]])

  #GBM
  bet_gbm<-gbm(formula = Betnan~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,
               distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001,
interaction.depth = 4)
  best.iter<-gbm.perf(bet_gbm, plot.it = F, method = "OOB")
```

```
  pred_bet_gbm<-predict.gbm(bet_gbm,newdata = eva, best.iter, type = "response")
  bet_auc_gbm_p<-colAUC(pred_bet_gbm, eva$Betnan, plotROC = F)
  bet_auc_gbm<- c(bet_auc_gbm, bet_auc_gbm_p[[1]])

}
compared_model_bat=cbind.data.frame(bet_auc_glm, bet_auc_gam, bet_auc_gbm)
}
#print the AUC values of the various models at different replications.
compared_model_bat
```

```
##    bet_auc_glm bet_auc_gam bet_auc_gbm
## 1    0.5344828   0.5737548   0.6264368
## 2    0.6180124   0.6925466   0.7142857
## 3    0.5584291   0.6503831   0.6436782
## 4    0.5710000   0.6990000   0.6990000
## 5    0.5200846   0.6670190   0.6701903
## 6    0.5009653   0.7084942   0.6515444
## 7    0.6322222   0.7533333   0.7833333
## 8    0.5138067   0.6084813   0.5113412
## 9    0.5686499   0.7070938   0.6395881
## 10   0.5771670   0.6310782   0.6522199
```

```
#mean of the the auc values for the various models
colMeans(compared_model_bat)
```

```
## bet_auc_glm bet_auc_gam bet_auc_gbm
##   0.5594820   0.6691184   0.6591618
```

```
#perform wilcoxon test to see if there is a significant improvement between the models
wilcox.test(bet_auc_glm, bet_auc_gam, paired = T)
```

```
##
##  Wilcoxon signed rank test
##
## data:  bet_auc_glm and bet_auc_gam
## V = 0, p-value = 0.001953
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(bet_auc_gam, bet_auc_gbm, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  bet_auc_gam and bet_auc_gbm
## V = 26, p-value = 0.7223
## alternative hypothesis: true location shift is not equal to 0
```

_ as it can be seen, there is a significant improvement from the glm to the gam. but insignificant from gam to gbm. However, with AUC below 0.8, the models are not strong enough._

Cashyp

```
{rep<-7
#create empty lists for the three models about to be built.
  cas_auc_glm<-cas_auc_gam<-cas_auc_gbm<-c()
for (i in 1:rep){
    #print(i)
    #divide the sample into 70:30 training and testing respectively
```

```
    rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
    cal<- data[rand_sam,]     #calibration
    eva<- data[-rand_sam,]    #evaluation

    #create the prediction using glm
    cas_glm<-glm(Cashyp~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,family
="binomial")
    pred_cas_glm<-predict.glm(cas_glm, newdata = eva, type = "response")
    cas_auc_glm_p<-colAUC(pred_cas_glm, eva$Cashyp, plotROC=F)
    cas_auc_glm <- c(cas_auc_glm, cas_auc_glm_p[[1]])

    #GAM
    cas_gam<-gam(Cashyp~s(mesotopo, k=4) + s(soil_moist, k=4) + s(soil_temp, k=4) +
                   s(soil_ph, k=4), data=cal,family ="binomial")
    pred_cas_gam<-predict.gam(cas_gam, newdata = eva, type = "response")
    cas_auc_gam_p<-colAUC(pred_cas_gam, eva$Cashyp, plotROC=F)
    cas_auc_gam <- c(cas_auc_gam, cas_auc_gam_p[[1]])

    #GBM
    cas_gbm<-gbm(formula = Cashyp~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,
                 distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001,
interaction.depth = 4)
    best.iter<-gbm.perf(cas_gbm, plot.it = F, method = "OOB")
    pred_cas_gbm<-predict.gbm(cas_gbm,newdata = eva, best.iter, type = "response")
    cas_auc_gbm_p<-colAUC(pred_cas_gbm, eva$Cashyp, plotROC = F)
    cas_auc_gbm<- c(cas_auc_gbm, cas_auc_gbm_p[[1]])


  }
  compared_model_cas=cbind.data.frame(cas_auc_glm, cas_auc_gam, cas_auc_gbm)
}
#print the AUC values of the various models at different replications.
compared_model_cas

##   cas_auc_glm cas_auc_gam cas_auc_gbm
## 1   0.8933333   0.8933333   0.8933333
## 2   0.9378531   0.9378531   0.8389831
## 3   0.9033333   0.7966667   0.8300000
## 4   0.9566667   0.9533333   0.9466667
## 5   0.9533333   0.9533333   0.9166667
## 6   0.8446328   0.8446328   0.7994350
## 7   0.9549180   0.9549180   0.8975410

#mean of the the auc values for the various models
colMeans(compared_model_cas)

## cas_auc_glm cas_auc_gam cas_auc_gbm
##   0.9205815   0.9048672   0.8746608

#perform wilcoxon test to see if there is a significant improvement between the models
wilcox.test(cas_auc_glm, cas_auc_gam, paired = T)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  cas_auc_glm and cas_auc_gam
```

```
## V = 3, p-value = 0.3711
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(cas_auc_gam, cas_auc_gbm, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  cas_auc_gam and cas_auc_gbm
## V = 19, p-value = 0.09349
## alternative hypothesis: true location shift is not equal to 0
```

Overall, the three models seem to be quite good without much difference, based on the parameters used.

Empher

```
{rep<-10
#create empty lists to imput the models auc values later
emp_auc_glm<-emp_auc_gam<-emp_auc_gbm<-c()
for (i in 1:rep){
  #print(i)
  rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
  cal<- data[rand_sam,]
  eva<- data[-rand_sam,]
  emp_glm<-glm(Empher~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,family ="binomial")
  pred_emp_glm<-predict.glm(emp_glm, newdata = eva, type = "response")
  emp_auc_glm_p<-colAUC(pred_emp_glm, eva$Empher, plotROC=F)
  emp_auc_glm <- c(emp_auc_glm, emp_auc_glm_p[[1]])

  #GAM
  emp_gam<-gam(Empher~s(mesotopo, k=3) + s(soil_moist, k=3) + s(soil_temp, k=3) +
                 s(soil_ph, k=3), data=cal,family ="binomial")
  pred_emp_gam<-predict.gam(emp_gam, newdata = eva, type = "response")
  emp_auc_gam_p<-colAUC(pred_emp_gam, eva$Empher, plotROC=F)
  emp_auc_gam <- c(emp_auc_gam, emp_auc_gam_p[[1]])

  #GBM
  emp_gbm<-gbm(formula = Empher~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,
               distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001,
interaction.depth = 4)
  best.iter<-gbm.perf(emp_gbm, plot.it = F, method = "OOB")
  pred_emp_gbm<-predict.gbm(emp_gbm,newdata = eva, best.iter, type = "response")
  emp_auc_gbm_p<-colAUC(pred_emp_gbm, eva$Empher, plotROC = F)
  emp_auc_gbm<- c(emp_auc_gbm, emp_auc_gbm_p[[1]])
}
#put all the results into a dataframe
compared_model_emp=cbind.data.frame(emp_auc_glm, emp_auc_gam, emp_auc_gbm)
}
#show the results
compared_model_emp
```

```
##    emp_auc_glm emp_auc_gam emp_auc_gbm
## 1    0.7966102   0.8079096   0.7655367
## 2    0.5158730   0.5793651   0.6230159
## 3    0.6106443   0.6344538   0.6960784
## 4    0.7895623   0.6952862   0.7053872
```

```
## 5    0.7828283   0.7508418   0.7474747
## 6    0.6626667   0.6973333   0.6466667
## 7    0.5476190   0.5826331   0.6778711
## 8    0.6290909   0.5818182   0.6509091
## 9    0.6430976   0.6767677   0.6801347
## 10   0.7563636   0.6690909   0.7454545

#the mean of the aucs values of the models
colMeans(compared_model_emp)

## emp_auc_glm emp_auc_gam emp_auc_gbm
##   0.6734356   0.6675500   0.6938529

#compare the models
wilcox.test(emp_auc_glm, emp_auc_gam, paired = T)

##
##  Wilcoxon signed rank test
##
## data:  emp_auc_glm and emp_auc_gam
## V = 29, p-value = 0.9219
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(emp_auc_gam, emp_auc_gbm, paired = T)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  emp_auc_gam and emp_auc_gbm
## V = 11.5, p-value = 0.1139
## alternative hypothesis: true location shift is not equal to 0
```

The gbm seems to be the best for Empher. However, all the models are not strong enough

Salret

```
{rep<-7   #number of replicatons of the sampling for the modelling
#create empoty lists´to imput the auc values later
sal_auc_glm<-sal_auc_gam<-sal_auc_gbm<-c()
for (i in 1:rep){
  #print(i)
  rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
  cal<- data[rand_sam,]
  eva<- data[-rand_sam,]
  sal_glm<-glm(Salret~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,family ="binomial")
  pred_sal_glm<-predict.glm(sal_glm, newdata = eva, type = "response")
  sal_auc_glm_p<-colAUC(pred_sal_glm, eva$Salret, plotROC=F)
  sal_auc_glm <- c(sal_auc_glm, sal_auc_glm_p[[1]])

  #GAM
  sal_gam<-gam(Salret~s(mesotopo, k=3) + s(soil_moist, k=3) + s(soil_temp, k=3) +
                 s(soil_ph, k=3), data=cal,family ="binomial")
  pred_sal_gam<-predict.gam(sal_gam, newdata = eva, type = "response")
  sal_auc_gam_p<-colAUC(pred_sal_gam, eva$Salret, plotROC=F)
  sal_auc_gam <- c(sal_auc_gam, sal_auc_gam_p[[1]])

  #GBM
```

```
  sal_gbm<-gbm(formula = Salret~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,
               distribution = "bernoulli",n.trees = 3000, shrinkage = 0.001,
interaction.depth = 4)
  best.iter<-gbm.perf(sal_gbm, plot.it = F, method = "OOB")
  pred_sal_gbm<-predict.gbm(sal_gbm,newdata = eva, best.iter, type = "response")
  sal_auc_gbm_p<-colAUC(pred_sal_gbm, eva$Salret, plotROC = F)
  sal_auc_gbm<- c(sal_auc_gbm, sal_auc_gbm_p[[1]])
}
compared_model_sal=cbind.data.frame(sal_auc_glm, sal_auc_gam, sal_auc_gbm)
}
#show the results
compared_model_sal
```

```
##   sal_auc_glm sal_auc_gam sal_auc_gbm
## 1   1.0000000   0.9892473   0.9784946
## 2   0.9523810   0.9523810   0.9523810
## 3   0.9126984   0.9285714   0.9761905
## 4   0.8968254   0.8968254   0.8650794
## 5   0.9761905   0.9761905   0.9206349
## 6   0.9682540   0.9682540   0.9404762
## 7   0.8933333   0.9066667   0.8466667
```

```
#the mean of the aucs values of the models
colMeans(compared_model_sal)
```

```
## sal_auc_glm sal_auc_gam sal_auc_gbm
##   0.9428118   0.9454480   0.9257033
```

```
#compare the models
wilcox.test(sal_auc_glm, sal_auc_gam, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  sal_auc_glm and sal_auc_gam
## V = 1, p-value = 0.4227
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(sal_auc_gam, sal_auc_gbm, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  sal_auc_gam and sal_auc_gbm
## V = 17, p-value = 0.2084
## alternative hypothesis: true location shift is not equal to 0
```

all the models for Salret seem strong, with glm performing best

Question 2.

What is the predictive performance of the GLM, GAM and GBM models for veg_height and vasc_spr? Again, build the models using calibration data and test the models using evaluation data. Use Spearman correlation -values as the evaluation metrics. Use the same set of predictors that you used in question 1). Report the results in one short paragraph (max 5 sentences).

```
data<- read.csv("C:/Users/oyeda/Desktop/MODELLING_PHYSICAL_GEOGRAPHY/assignment3/Data-
20171114 (1)/saana.csv" ,sep=";")
# Use the caTools package to extract the AUC values and compare them
#library(caTools)
#library(mgcv)
#library(gbm)

#number of times to repeat the models
{rep<-10
  h_auc_glm<-auc_gam<-h_auc_gbm<-c()
for (i in 1:rep){
    #print(i)
    #sample all the rows, and keep 70%(0.7)
    rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
    cal<- data[rand_sam,]   #get the 70% rows for calibration
    eva<- data[-rand_sam,]  #get the remaining 30% for evaluation

    #create the glm for veg_height occurences
    h_glm<-glm(veg_height~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,family
="gaussian")
    pred_h_glm<-predict.glm(h_glm, newdata = eva, type = "response")
    h_cor_glm<-cor(pred_h_glm, eva$veg_height, method = "spearman")


    #GAM
    h_gam<-gam(veg_height~s(mesotopo, k=3) + s(soil_moist, k=3) + s(soil_temp, k=3) +
                 s(soil_ph, k=3), data=cal,family ="gaussian")
    pred_h_gam<-predict.gam(h_gam, newdata = eva, type = "response")
    h_cor_gam<-cor(pred_h_gam, eva$veg_height, method = "spearman")

    #GBM
    h_gbm<-gbm(formula = veg_height~mesotopo+soil_moist+soil_temp+soil_ph, data=data,
          distribution = "gaussian",n.trees = 3000, shrinkage = 0.001, interaction.depth
= 4)
    best.iter<-gbm.perf(h_gbm, plot.it = F, method = "OOB")
    pred_h_gbm<-predict.gbm(h_gbm,newdata = eva, best.iter, type = "response")
    h_cor_gbm<-cor(pred_h_gbm, eva$veg_height, method = "spearman")
  }
  compared_model_h=cbind.data.frame(h_cor_glm, h_cor_gam, h_cor_gbm)
}
#comparison between the corelation between predicted and observed vegetation height of
the models
compared_model_h

##   h_cor_glm h_cor_gam h_cor_gbm
## 1  0.397975 0.4106639 0.4999944
```

Using the three models, they show low correlation with the observed vegetation height from the evaluation/testing data

vasc_spr

```
{rep<-7
vspr_auc_glm<-vspr_auc_gam<-vspr_auc_gbm<-c()
for (i in 1:rep){
  #print(i)
```

```
  #sample all the rows, and keep 70%(0.7)
  rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
  cal<- data[rand_sam,]   #get the 70% rows for calibration
  eva<- data[-rand_sam,]  #get the remaining 30% for evaluation

  #create the glm for veg_height occurences
  vspr_glm<-glm(vasc_spr~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,family
="poisson")
  pred_vspr_glm<-predict.glm(vspr_glm, newdata = eva, type = "response")
  vspr_cor_glm<-cor(pred_vspr_glm, eva$vasc_spr, method = "spearman")

  #GAM
  vspr_gam<-gam(vasc_spr~s(mesotopo, k=3) + s(soil_moist, k=3) + s(soil_temp, k=3) +
                  s(soil_ph, k=3), data=cal,family ="poisson")
  pred_vspr_gam<-predict.gam(vspr_gam, newdata = eva, type = "response")
  vspr_cor_gam<-cor(pred_vspr_gam, eva$vasc_spr, method = "spearman")

  #GBM
  vspr_gbm<-gbm(formula = vasc_spr~mesotopo+soil_moist+soil_temp+soil_ph, data=data,
            distribution = "poisson",n.trees = 3000, shrinkage = 0.001, interaction.depth
= 4)
  best.iter<-gbm.perf(vspr_gbm, plot.it = F, method = "OOB")
  pred_vspr_gbm<-predict.gbm(vspr_gbm,newdata = eva, best.iter, type = "response")
  vspr_cor_gbm<-cor(pred_vspr_gbm, eva$vasc_spr, method = "spearman")
}
compared_model_vspr=cbind.data.frame(vspr_cor_glm, vspr_cor_gam, vspr_cor_gbm)
}
#compare the models
compared_model_vspr
```

```
##   vspr_cor_glm vspr_cor_gam vspr_cor_gbm
## 1    0.7181524    0.7355887    0.8446751
```

Here, the GBM seems to be the best and others are fairly good too but less reliable

Question 3.

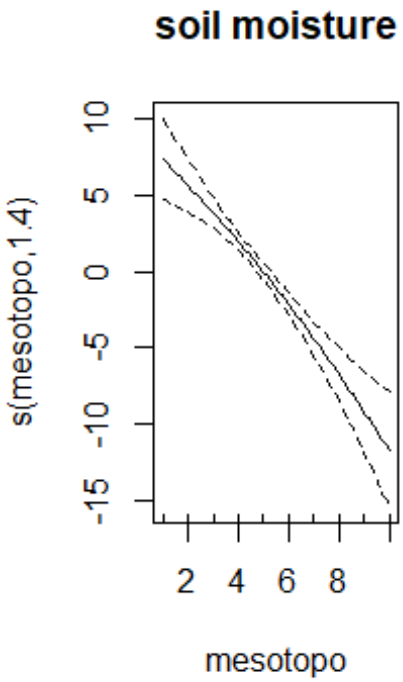
Characterize soil_moist, soil_temp, soil_ph, veg_height and vasc_spr conditions along the mesotopographic gradient using GAM. Model the values of these five responses at the valley bottom (mesotopo 1), mid-slope (mesotopo 5) and ridge-top (mesotopo 10). Present the results as an informative figure. Report the results in one short paragraph (max 5 sentences).

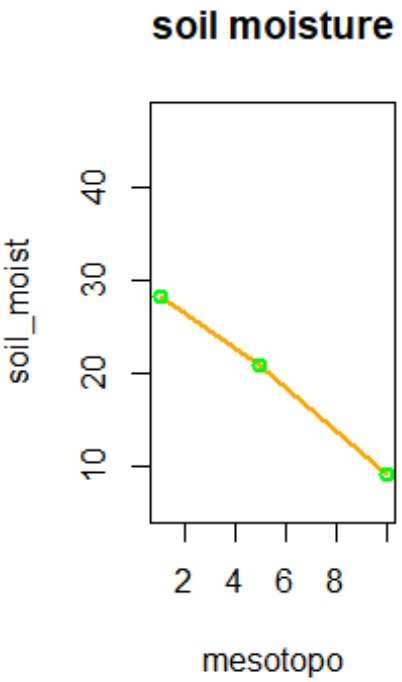
```
#library(mgcv)
data<- read.csv("C:/Users/oyeda/Desktop/MODELLING_PHYSICAL_GEOGRAPHY/assignment3/Data-
20171114 (1)/saana.csv"
                ,sep=";")
attach(data)
#This is the first method makes the prediction without classifying the mesotopography
#GAM
```

Soil moisture

```
gam_moist <- gam(soil_moist~s(mesotopo, k=3), data = data, family = "gaussian")
summary(gam_moist)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## soil_moist ~ s(mesotopo, k = 3)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.7514     0.4879   42.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df     F p-value
## s(mesotopo) 1.396  1.635 36.71 3.9e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.228   Deviance explained = 23.3%
## GCV = 51.999  Scale est. = 51.423    n = 216

par(mfrow=c(1,2))
plot(gam_moist, main = "soil moisture")

#the values at the valley bottom, mid-slope and ridge-top
mesotopo2 <- c(1,5,10)
newdata <- data.frame(mesotopo=mesotopo2)
pred.gam_moist <- predict.gam(gam_moist, newdata, type="response")
plot(mesotopo, soil_moist, pch=19, cex=0.2, col="grey",type="n", main="soil moisture")
lines(mesotopo2, pred.gam_moist, lty=1,lwd= 2,col="orange")
points(mesotopo2, pred.gam_moist, lty=1,lwd= 2,col="green")
```



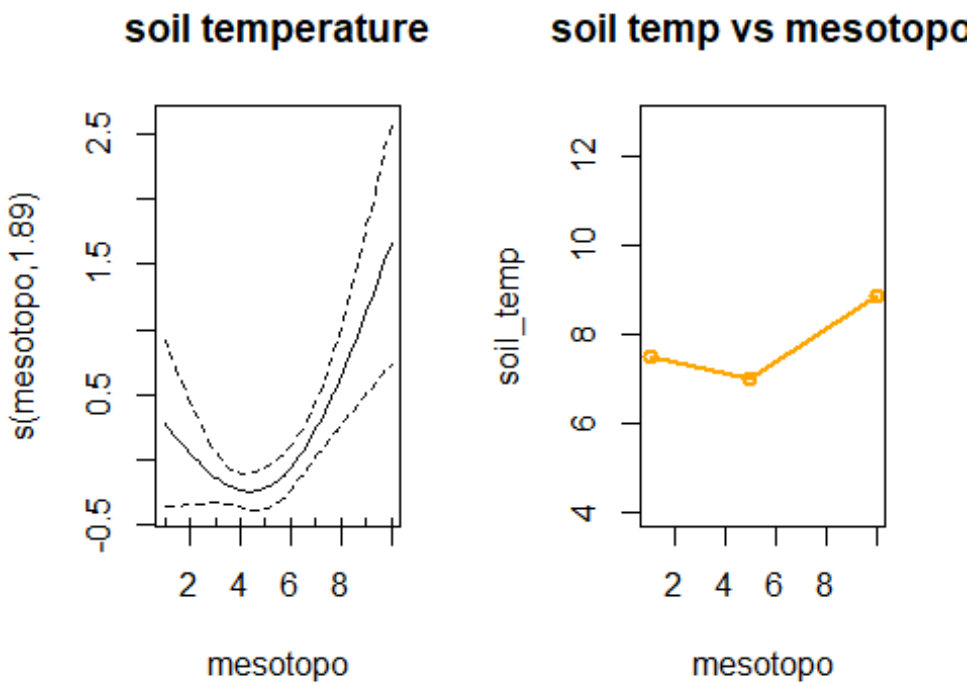


Overall, soil moisture seems to be reducing with increase in mesotopographical gradient. The response curve shows high confidence at the mid-slope which shos that there are more samples from tha area.

soil temperature

```
gam_temp<- gam(soil_temp~s(mesotopo, k=3), data = data, family = "gaussian")
summary(gam_temp)   #summary soil temperature

##
## Family: gaussian
## Link function: identity
##
## Formula:
## soil_temp ~ s(mesotopo, k = 3)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.21135    0.09504   75.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df     F p-value
## s(mesotopo) 1.889  1.988 5.965 0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0526   Deviance explained = 6.09%
## GCV = 1.9773  Scale est. = 1.9509    n = 216

par(mfrow=c(1,2))
plot(gam_temp, main="soil temperature") #response curve
pred.gam_temp <- predict.gam(gam_temp, newdata, type="response")
plot(mesotopo, soil_temp, pch=19, cex=0.2, col="grey",type="n", main = "soil temp vs
```

```
mesotopo")
lines(mesotopo2, pred.gam_temp, lty=1,lwd= 2,col="orange")
points(mesotopo2, pred.gam_temp, lty=1,lwd= 2,col="orange")
```

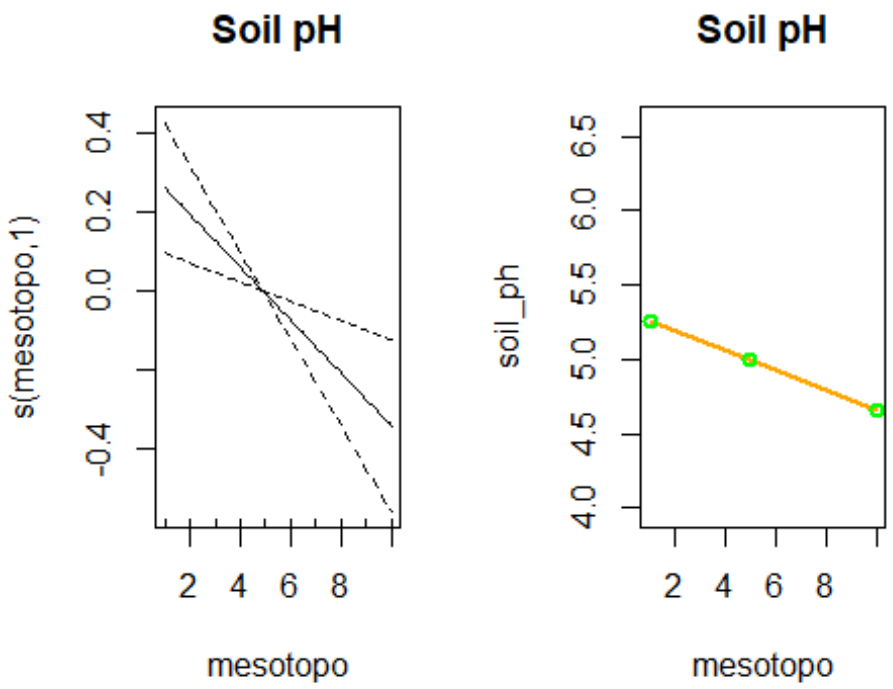


soil temperature seems to be slightly reducing as one approaches mid-slope, then increases towards the rigde-top. This could perhaps,be because there are lesser/sparse vegeation at the ridge-top, thereby exposing the soil to solar radiation. However, the confidence are slower at the valley-bottom and ridge-top, ostensibly because, there are less data from those areas.

soil pH

```
gam_ph<- gam(soil_ph~s(mesotopo, k=3), data = data, family = "gaussian")
summary(gam_ph)   #summary soil_pH

##
## Family: gaussian
## Link function: identity
##
## Formula:
## soil_ph ~ s(mesotopo, k = 3)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.99978    0.04018   124.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(mesotopo)   1      1 9.893 0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## R-sq.(adj) =  0.0397   Deviance explained = 4.42%
## GCV = 0.35205  Scale est. = 0.34879   n = 216

par(mfrow=c(1,2))
plot(gam_ph, main = "Soil pH")  #response curve

pred.gam_ph <- predict.gam(gam_ph, newdata, type="response")
plot(mesotopo, soil_ph, pch=19, cex=0.2, col="grey",type="n", main = "Soil pH")
lines(mesotopo2, pred.gam_ph, lty=1,lwd= 2,col="orange")
points(mesotopo2, pred.gam_ph, lty=1,lwd= 2,col="green")
```



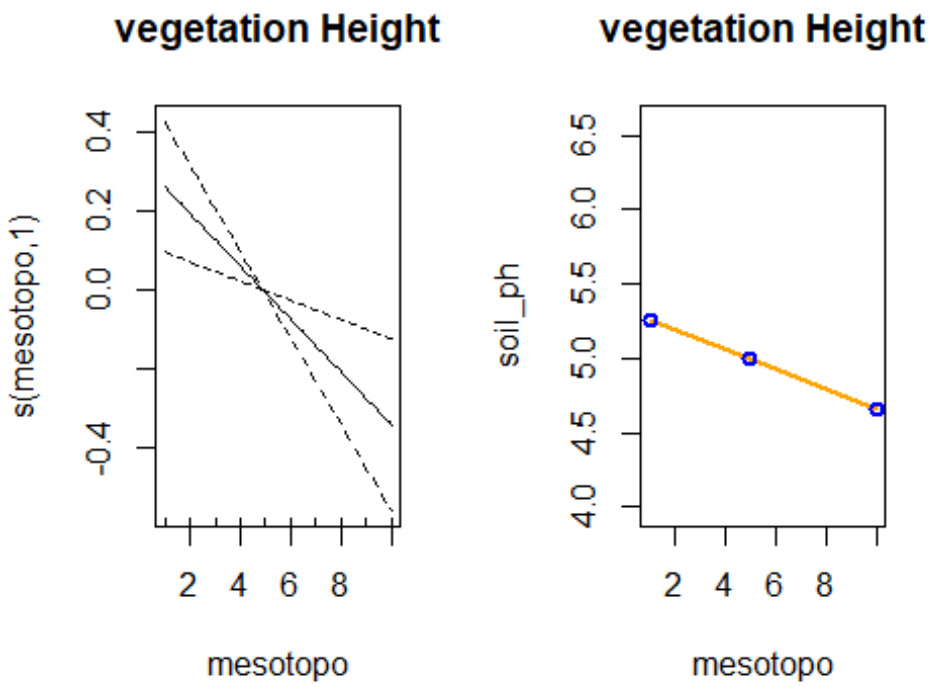
The soil ph also appears to be reducing upslope. This means that the soil upslope are more acidic. However, the confidence levels are low at the valley bottom and ridge-top

Vegetation Height

```
par(mfrow=c(1,2))
gam_vh<- gam(veg_height~s(mesotopo, k=3), data = data, family = "poisson")
summary(gam_vh)

##
## Family: poisson
## Link function: log
##
## Formula:
## veg_height ~ s(mesotopo, k = 3)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.70664    0.02929   58.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
```

```
##              edf Ref.df Chi.sq  p-value
## s(mesotopo) 1.87  1.983  34.16 2.29e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0724   Deviance explained = 10.4%
## UBRE = 0.6694  Scale est. = 1         n = 216
```

```
plot(gam_ph, main = "vegetation Height")
pred.gam_ph <- predict.gam(gam_ph, newdata, type="response")
plot(mesotopo, soil_ph, pch=19, cex=0.2, col="grey",type="n", main = "vegetation Height")
lines(mesotopo2, pred.gam_ph, lty=1,lwd= 2,col="orange")
points(mesotopo2, pred.gam_ph, lty=1,lwd= 2,col="blue")
```



This is also similar to the soil pH. This is expected, as height would be affected by the acidity and soil moisture which is lower upslope. The vegetation height seems to be reducing, as one approaches upslope

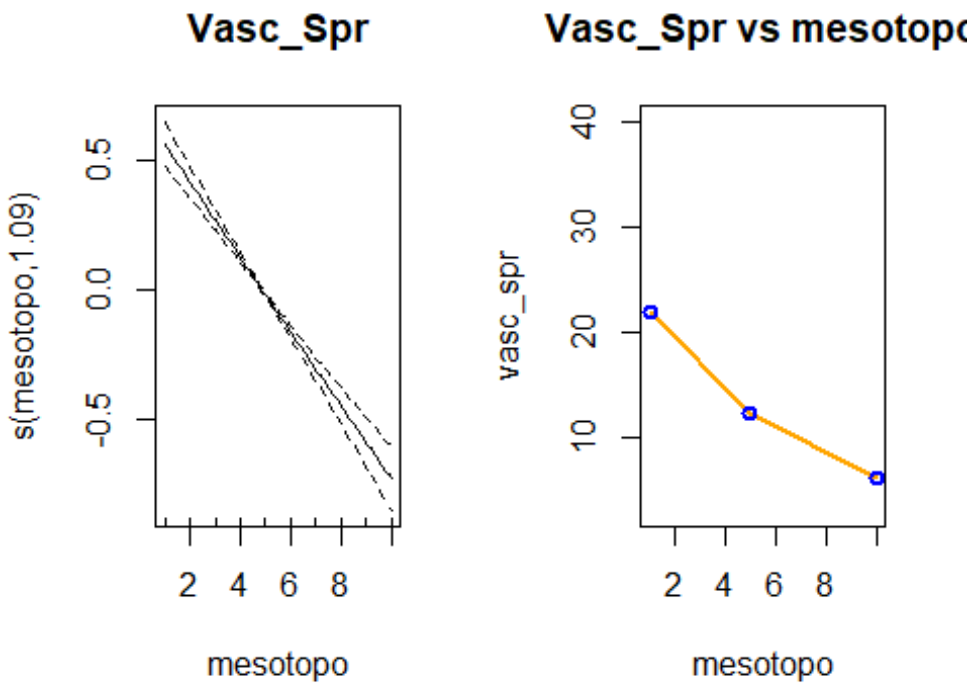
Vascular species richness

```
gam_vaspr<- gam(vasc_spr~s( mesotopo, k=3), data = data, family = "poisson")
summary(gam_vaspr)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## vasc_spr ~ s(mesotopo, k = 3)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.52237    0.01962   128.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Approximate significance of smooth terms:
##                edf Ref.df Chi.sq p-value
## s(mesotopo) 1.085  1.163    189  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.248   Deviance explained =   27%
## UBRE = 1.4259  Scale est. = 1         n = 216
```

```
par(mfrow=c(1,2))
plot(gam_vaspr, main = "Vasc_Spr")
pred.gam_vaspr <- predict.gam(gam_vaspr, newdata, type="response")
plot(mesotopo, vasc_spr, pch=19, cex=0.2, col="grey",type="n", main = "Vasc_Spr vs
mesotopo")
lines(mesotopo2, pred.gam_vaspr, lty=1,lwd= 2,col="orange")
points(mesotopo2, pred.gam_vaspr, lty=1,lwd= 2,col="blue")
```



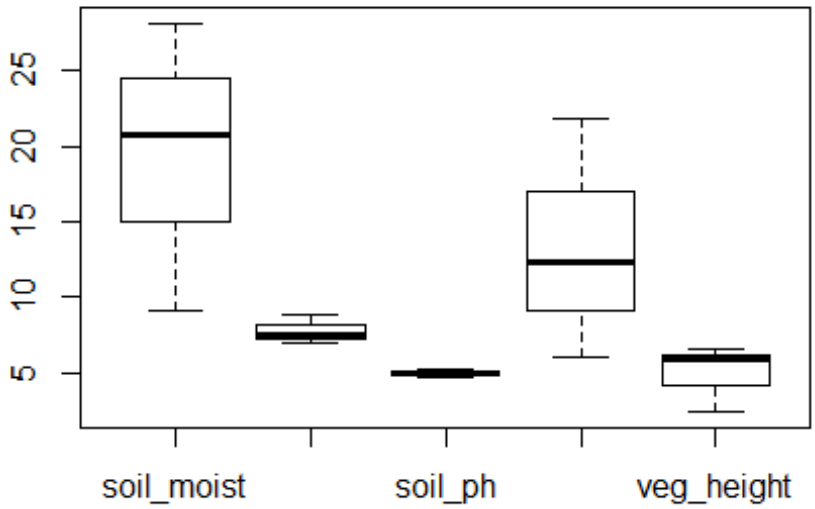
vascular species richness also seems tobe reducing uplspe. Same conditions that affect vegeatin heights are expected to affect the vascular species richness.

```
################################
#data$topo_level<-cut(mesotopo, breaks = c(0,4,7,10))
#levels(data$topo_level)<-c("valley-bottom","mid-slope","ridge-top")

#Subsetting the mesotop into the various parts
val_bot<-data[mesotopo==1,]    #valley bottom
mid_sl<-data[mesotopo==5,]      #mid-slope
r_top<-data[mesotopo==10,]      #ridge-top


#create data frame to impute the modlled values at variouis topo gradients.
topo<-matrix(ncol = 5, nrow = 3)
```

```r
topo<- data.frame(topo)
row.names(topo)<- c("valley_bottom", "mid-slope", "ridge-top")
colnames(topo)<-c("soil_moist", "soil_temp", "soil_ph", "vasc_spr", "veg_height")


#predicting the values at the valley bottom for the responses
vb1<- topo[1,1]<- mean(predict.gam(gam_moist, val_bot, type="response"))
vb2<- topo[1,2]<- mean(predict.gam(gam_temp, val_bot, type="response"))
vb3<- topo[1,3]<- mean(predict.gam(gam_ph, val_bot, type="response"))
vb4<- topo[1,4]<- mean(predict.gam(gam_vaspr, val_bot, type="response"))
vb5<- topo[1,5]<- mean(predict.gam(gam_vh, val_bot, type="response"))

#predicting the values at the mid-slope for the responses
ms1<- topo[2,1]<- mean(predict.gam(gam_moist, mid_sl, type="response"))
ms2<- topo[2,2]<- mean(predict.gam(gam_temp, mid_sl, type="response"))
ms3<- topo[2,3]<- mean(predict.gam(gam_ph, mid_sl, type="response"))
ms4<- topo[2,4]<- mean(predict.gam(gam_vaspr, mid_sl, type="response"))
ms5<- topo[2,5]<- mean(predict.gam(gam_vh, mid_sl, type="response"))

#predicting the values at the ridge-top for the responses
rt1<- topo[3,1]<- mean(predict.gam(gam_moist, r_top, type="response"))
rt2<- topo[3,2]<- mean(predict.gam(gam_temp, r_top, type="response"))
rt3<- topo[3,3]<- mean(predict.gam(gam_ph, r_top, type="response"))
rt4<- topo[3,4]<- mean(predict.gam(gam_vaspr, r_top, type="response"))
rt5<- topo[3,5]<- mean(predict.gam(gam_vh, r_top, type="response"))

#view the dataframe
topo
```

```
##               soil_moist soil_temp  soil_ph  vasc_spr veg_height
## valley_bottom  28.124142  7.487852 5.261009 21.851141   6.556635
## mid-slope      20.752299  6.996628 4.992635 12.250299   5.830233
## ridge-top       9.100299  8.875371 4.657167  6.047608   2.411296
```

```r
#see the boxplot
boxplot(topo)
```



There seems to be trends in variables changes across slope

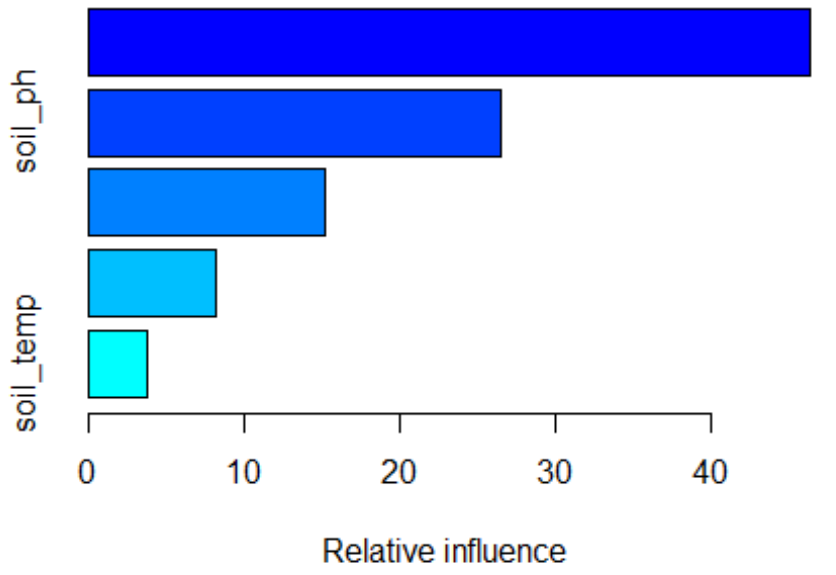
Question 4.

Does the cover of Empetrum hermaphroditum (Empher_cover) have an effect on the vasc_spr when all other predictors are controlled for? Use the same set of predictors as used in question 1). Use all three modelling frameworks to test the hypothesis. Report the results in one short paragraph (max 5 sentences).The main idea behind this question: as a dominant species Empher_cover might have a strong influence on the vegetation properties - can we see the effect? Please, test it!

```
data<- read.csv("C:/Users/oyeda/Desktop/MODELLING_PHYSICAL_GEOGRAPHY/assignment3/Data-
20171114 (1)/saana.csv",sep=";")

# Use the caTools package to extract the AUC values and compare them
#library(caTools)
#library(mgcv)
#library(gbm)

attach(data)

#create the glm for vasc_spr occurences'
#GLM
vaspr_glm<-glm(vasc_spr~Empher_cover+mesotopo+soil_moist+soil_temp+soil_ph,
data=data,family ="poisson")
summary(vaspr_glm)

##
## Call:
## glm(formula = vasc_spr ~ Empher_cover + mesotopo + soil_moist +
##     soil_temp + soil_ph, family = "poisson", data = data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9517  -0.8621  -0.1215   0.7040   2.8930
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.638905   0.267599   2.388    0.017 *
## Empher_cover -0.005616   0.001012  -5.552 2.82e-08 ***
## mesotopo     -0.053376   0.012330  -4.329 1.50e-05 ***
## soil_moist    0.025377   0.002470  10.275  < 2e-16 ***
## soil_temp     0.016219   0.015725   1.031    0.302
## soil_ph       0.310197   0.034137   9.087  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 712.23  on 215  degrees of freedom
## Residual deviance: 241.45  on 210  degrees of freedom
## AIC: 1178.1
##
## Number of Fisher Scoring iterations: 4

#GAM
vaspr_gam<-gam(vasc_spr~s(Empher_cover)+s(mesotopo)+s(soil_moist)+
                  s(soil_temp)+s(soil_ph), data=data,family ="poisson")
summary(vaspr_gam)

##
## Family: poisson
## Link function: log
##
## Formula:
## vasc_spr ~ s(Empher_cover) + s(mesotopo) + s(soil_moist) + s(soil_temp) +
##     s(soil_ph)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.47080    0.02048   120.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df Chi.sq  p-value
## s(Empher_cover) 5.270  6.387 55.081 1.11e-09 ***
## s(mesotopo)     2.669  3.387 10.699   0.0172 *
## s(soil_moist)   1.744  2.185 91.887  < 2e-16 ***
## s(soil_temp)    1.045  1.089  1.415   0.2707
## s(soil_ph)      2.687  3.363 97.433  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.735   Deviance explained = 73.3%
## UBRE = 0.013044  Scale est. = 1         n = 216
```

```
#GBM
vaspr_gbm<-gbm(formula = vasc_spr~Empher_cover+mesotopo+soil_moist+soil_temp+soil_ph,
data=data,
                  distribution = "poisson",n.trees = 3000, shrinkage = 0.001,
interaction.depth = 4)
summary(vaspr_gbm)
```




```
##                         var   rel.inf
## soil_moist     soil_moist 46.312510
## soil_ph           soil_ph 26.451865
## Empher_cover Empher_cover 15.225292
## mesotopo         mesotopo  8.239242
## soil_temp       soil_temp  3.771091
```

From the above, Empher_cover appears to be a significant predictor and has the third relative importance as shown in GBM. Therefore, we can say Empher_Cover has a strong influence on vegetation properties

testing the model Without Empher_Cover

```
{rep<-10
  vaspr_auc_glm<-vaspr_auc_gam<-vaspr_auc_gbm<-c()
for (i in 1:rep){
    #print(i)
    rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
    cal<- data[rand_sam,]
    eva<- data[-rand_sam,]
    vaspr_glm<-glm(vasc_spr~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,family
="poisson")
    pred_vaspr_glm<-predict.glm(vaspr_glm, newdata = eva, type = "response")
    vaspr_auc_glm_p<-colAUC(pred_vaspr_glm, eva$vasc_spr, plotROC=F)
    vaspr_auc_glm <- c(vaspr_auc_glm, vaspr_auc_glm_p[[1]])
```

```
    #GAM
    vaspr_gam<-gam(vasc_spr~s(mesotopo, k=3) + s(soil_moist, k=3) + s(soil_temp, k=3) +
                   s(soil_ph, k=3), data=cal,family ="poisson")
    pred_vaspr_gam<-predict.gam(vaspr_gam, newdata = eva, type = "response")
    vaspr_auc_gam_p<-colAUC(pred_vaspr_gam, eva$vasc_spr, plotROC=F)
    vaspr_auc_gam <- c(vaspr_auc_gam, vaspr_auc_gam_p[[1]])

    #GBM
    vaspr_gbm<-gbm(formula = vasc_spr~mesotopo+soil_moist+soil_temp+soil_ph, data=cal,
                   distribution = "poisson",n.trees = 3000, shrinkage = 0.001,
interaction.depth = 4)
    best.iter<-gbm.perf(vaspr_gbm, plot.it = F, method = "OOB")
    pred_vaspr_gbm<-predict.gbm(vaspr_gbm,newdata = eva, best.iter, type = "response")
    vaspr_auc_gbm_p<-colAUC(pred_vaspr_gbm, eva$vasc_spr, plotROC = F)
    vaspr_auc_gbm<- c(vaspr_auc_gbm, vaspr_auc_gbm_p[[1]])
  }
  compared_model_vaspr1=cbind.data.frame(vaspr_auc_glm, vaspr_auc_gam, vaspr_auc_gbm)
}
```

Tesing the model when Empher_cover is included in the prediction

```
{rep<-10
  vaspr_auc_glm<-vaspr_auc_gam<-vaspr_auc_gbm<-c()
for (i in 1:rep){
    #print(i)
    rand_sam<-sample(1:nrow(data), size = 0.7*nrow(data) )
    cal<- data[rand_sam,]
    eva<- data[-rand_sam,]
    vaspr_glm<-glm(vasc_spr~Empher_cover+mesotopo+soil_moist+soil_temp+soil_ph,
data=cal,family ="poisson")
    pred_vaspr_glm<-predict.glm(vaspr_glm, newdata = eva, type = "response")
    vaspr_auc_glm_p<-colAUC(pred_vaspr_glm, eva$vasc_spr, plotROC=F)
    vaspr_auc_glm <- c(vaspr_auc_glm, vaspr_auc_glm_p[[1]])

    #GAM
    vaspr_gam<-gam(vasc_spr~s(Empher_cover, k=3)+  s(mesotopo, k=3) + s(soil_moist, k=3)
+ s(soil_temp, k=3) +
                     s(soil_ph, k=3), data=cal,family ="poisson")
    pred_vaspr_gam<-predict.gam(vaspr_gam, newdata = eva, type = "response")
    vaspr_auc_gam_p<-colAUC(pred_vaspr_gam, eva$vasc_spr, plotROC=F)
    vaspr_auc_gam <- c(vaspr_auc_gam, vaspr_auc_gam_p[[1]])

    #GBM
    vaspr_gbm<-gbm(formula = vasc_spr~Empher_cover+mesotopo+soil_moist+soil_temp+soil_ph,
data=cal,
                   distribution = "poisson",n.trees = 3000, shrinkage = 0.001,
interaction.depth = 4)
    best.iter<-gbm.perf(vaspr_gbm, plot.it = F, method = "OOB")
    pred_vaspr_gbm<-predict.gbm(vaspr_gbm,newdata = eva, best.iter, type = "response")
    vaspr_auc_gbm_p<-colAUC(pred_vaspr_gbm, eva$vasc_spr, plotROC = F)
    vaspr_auc_gbm<- c(vaspr_auc_gbm, vaspr_auc_gbm_p[[1]])
  }
  compared_model_vaspr2=cbind.data.frame(vaspr_auc_glm, vaspr_auc_gam, vaspr_auc_gbm)
}

compared_model_vaspr1
```

```
##    vaspr_auc_glm vaspr_auc_gam vaspr_auc_gbm
## 1      1.0000000     1.0000000     1.0000000
## 2      0.6000000     0.7000000     0.6000000
## 3      0.8333333     0.8333333     0.6666667
## 4      0.8000000     0.8000000     1.0000000
## 5      0.7500000     1.0000000     0.5000000
## 6      1.0000000     1.0000000     0.5000000
## 7      1.0000000     0.5000000     1.0000000
## 8      0.5000000     1.0000000     1.0000000
## 9      0.7500000     0.5833333     0.7500000
## 10     0.5833333     0.6666667     0.6666667
```

```
compared_model_vaspr2
```

```
##    vaspr_auc_glm vaspr_auc_gam vaspr_auc_gbm
## 1      1.0000000     0.6666667     1.0000000
## 2      0.7500000     0.7500000     0.7500000
## 3      0.5000000     1.0000000     1.0000000
## 4      0.7500000     0.6250000     0.8750000
## 5      0.5000000     0.5000000     1.0000000
## 6      0.5000000     0.7500000     1.0000000
## 7      0.6666667     0.6666667     0.6666667
## 8      1.0000000     0.5000000     1.0000000
## 9      1.0000000     1.0000000     1.0000000
## 10     0.5000000     0.5000000     0.6666667
```

```
wilcox.test(mean(compared_model_vaspr1[,1]), mean(compared_model_vaspr2[,1]))
```

```
##
##  Wilcoxon rank sum test
##
## data:  mean(compared_model_vaspr1[, 1]) and mean(compared_model_vaspr2[, 1])
## W = 1, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(mean(compared_model_vaspr1[,2]), mean(compared_model_vaspr2[,2]))
```

```
##
##  Wilcoxon rank sum test
##
## data:  mean(compared_model_vaspr1[, 2]) and mean(compared_model_vaspr2[, 2])
## W = 1, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(mean(compared_model_vaspr1[,3]), mean(compared_model_vaspr2[,3]))
```

```
##
##  Wilcoxon rank sum test
##
## data:  mean(compared_model_vaspr1[, 3]) and mean(compared_model_vaspr2[, 3])
## W = 0, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

I decided to try my hands on building the model with and without Empher_cover and comparing the auc values. There seems to be a slight improvement, as shown in the tables, However, the predictions did not improve significantly