Coordinating and integRating state-of-the-art Earth Observation Activities in the regions of North Africa, Middle East and Balkans and Developing Links with GEO related intiatives toward GEOSS

GEOCRADLE SSL and Machine Learning Algorithms

Capacities and Skills: Towards the provision of EO services in the Balkans & MENA region

T4.2 – Improved Food Security and Water Extremes Management

14th of June 2017

Table of Contents

Challenge!

Prerequisite

[Download R-3.4.0 for Windows (32/64 bit)]

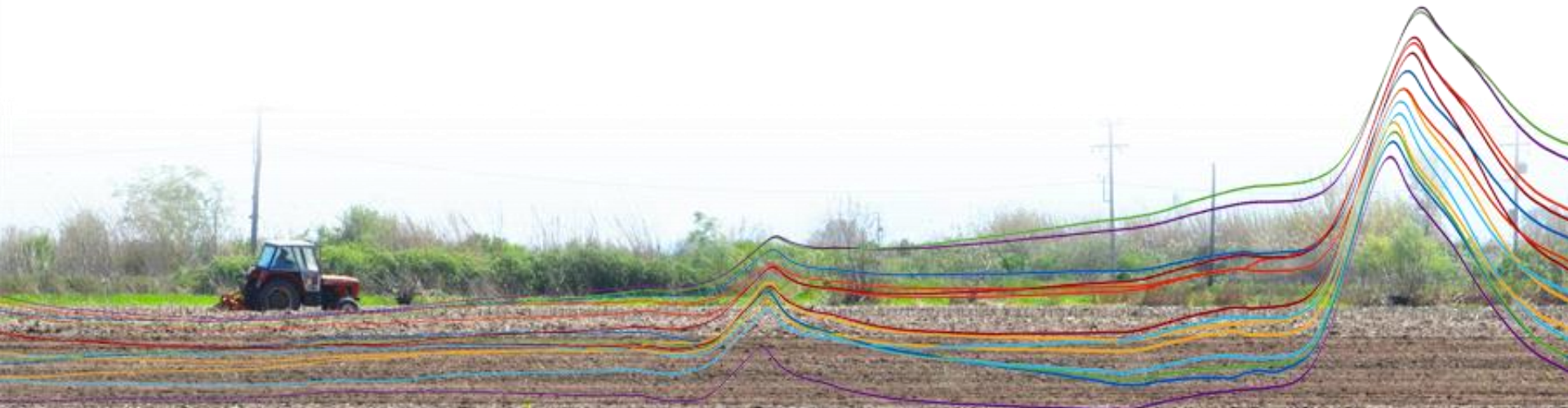
[Download R Studio]

Download the GEOCRADLE Soil Spectral Library

Create the following folders and store the SSL
C:/GEOCRADLE/data

Objective

Prediction of representative agricultural attributes to better manage the agro technical activities, using machine learning algortihms on GEO-CRADLE Soil Spectral Library



Introduction to R programming language

Installing packages

Part 1-Getting the Package onto Your Computer

Type “install.packages(“caret”)” and then press Ctrl+Return.

Part 2-Loading the Package into R Studio

Type “library(caret)” and then press Ctrl+Return

Introduction to R programming language

RStudio

File Edit Code View Plots Session Build Debug Tools Help

GEO-CRADLE_SSL_modelling_v02.R

```
plot( t(spectra[1,])~t(wavelength), type="l", lty=1, ylim=c(0,1), xlab = "Wavelength (nm)", yl
points( t(spectra[10,])~t(wavelength), type="l", lty=2,col=2, ylim=c(0,1))

# Application of data pretreatment techniques to the raw spectral data prior to the regressior
# a)log spectra
spectra_log<-log(1/spectra)
plot( t(spectra_log[1,])~t(wavelength), type="l", lty=1, xlab = "Wavelength (nm)", ylab = "Abs
points( t(spectra_log[10,])~t(wavelength), type="l", lty=2,col=2, ylim=c(0,1))

# b)savitzky golay filter
spectra_sg<-savitzkyGolay (spectra, 1,3,21)
spectra_sg<- as.data.frame(spectra_sg)
wavelength_sg <- as.data.frame(t(c((360:2490))))
plot( t(spectra_sg[1,])~t(wavelength_sg), type="l", lty=1, xlab = "Wavelength (nm)", ylab = "R
points( t(spectra_sg[10,])~t(wavelength_sg), type="l", lty=2,col=2)

# QUESTION: Assess the quality of the soil spectral signatures, explain the form of the curve

# Step 3: Explore the soil properties and apply the Kennard-Stone Algorithm For Calibration Sa
# Prepare variables for soil properties and evaluate the data, #10= organic matter, #15=NO3
var <- 10 # specify the variable

# Create a column with the dependent variable (property)
```

Console C:/GEOCRADLE/data/

```
ance (%)", col = 4, main="Log(1/R) spectra")
> points( t(spectra_log[10,])~t(wavelength), type="l", lty=2,col=2, ylim=c(0,1))
>
> # b)savitzky golay filter
> spectra_sg<-savitzkyGolay (spectra, 1,3,21)
> spectra_sg<- as.data.frame(spectra_sg)
> wavelength_sg <- as.data.frame(t(c((360:2490))))
> plot( t(spectra_sg[1,])~t(wavelength_sg), type="l", lty=1, xlab = "Wavelength (nm)", ylab = "Refl
ectance (%)",ylim=c(-0.005,0.005), col = 4, main="Savintzky Golay filter applied to spectra")
> points( t(spectra_sg[10,])~t(wavelength_sg), type="l", lty=2,col=2)
>
```

Environment History

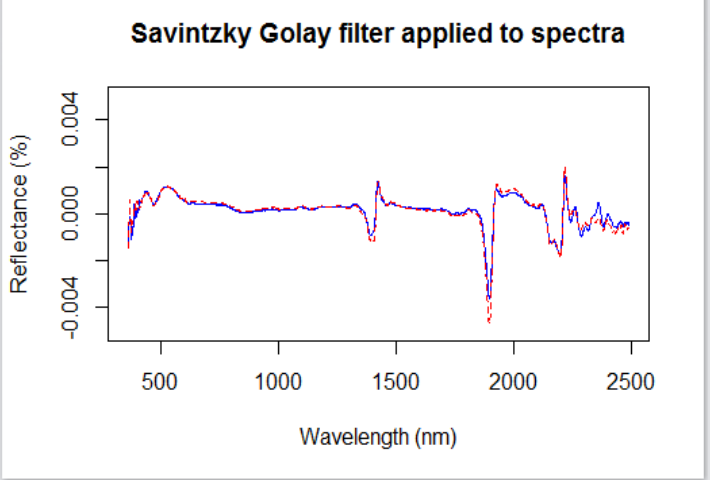
Import Dataset List

Global Environment

Data	
data	928 obs. of 2166 variables
spectra	928 obs. of 2151 variables
spectra_log	928 obs. of 2151 variables
spectra_sg	928 obs. of 2131 variables
wavelength	1 obs. of 2151 variables
wavelength_sg	1 obs. of 2131 variables

Files Plots Packages Help Viewer

Zoom Export Publish



Data exploration

#Import the GEOCRADLE Soil Spectral Library

```
data <- read.table("GEOCRADLE_SSL.csv", header=TRUE, sep=",")
head(data[1:16]) # Check the columns
```

	ID	Date.of.sampling	Latitude	Longitude	Elevation..m.	Depth..cm.	Soil.type..WRB.
1	GR-CM-060-00105	10/07/2015	40.91073	24.75147	8	60	CM
2	GR-CM-090-00105	10/07/2015	40.99353	24.84565	8	90	CM
3	GR-CM-030-00106	10/07/2015	40.98661	24.83491	10	30	CM
4	GR-CM-060-00106	10/07/2015	40.98016	24.83773	10	60	CM
5	GR-CM-090-00106	10/07/2015	40.94469	24.74419	10	90	CM
6	GR-CM-030-00108	10/07/2015	40.97939	24.85433	12	30	CM

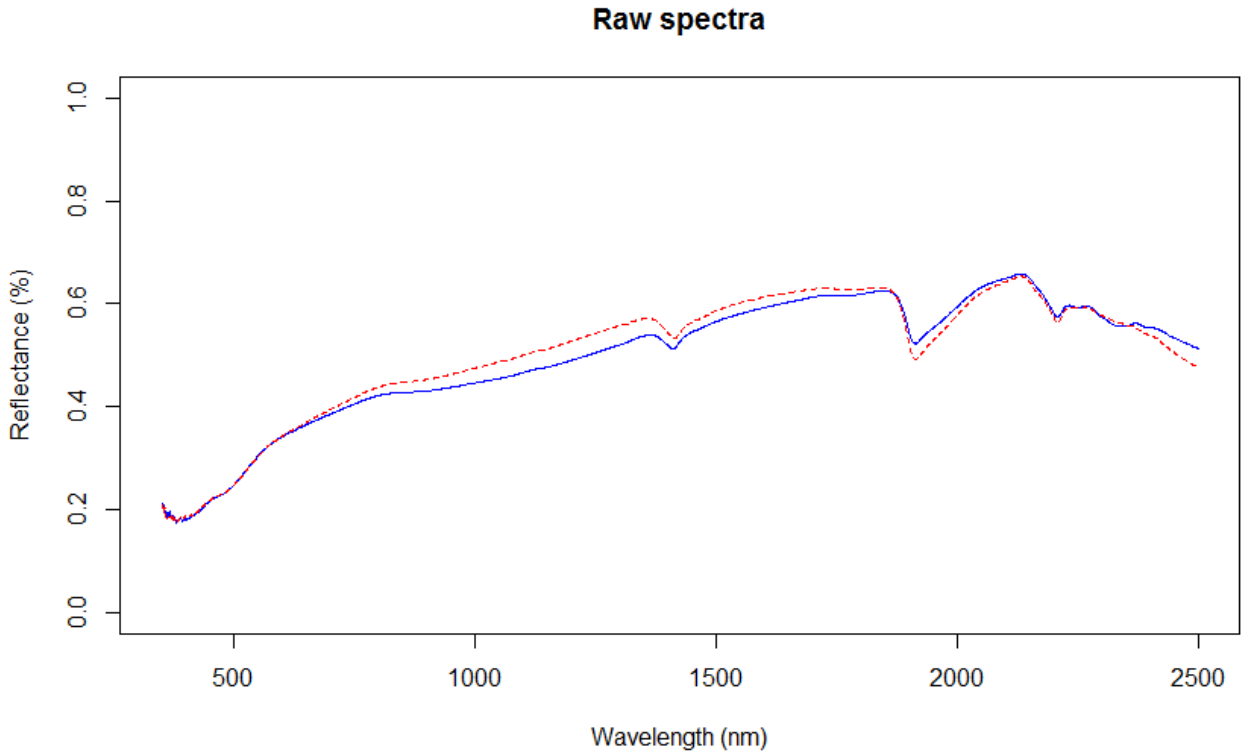
	Soil.type..USDA.	Climate..Koeppen.	OM....	CaCO3....	Sand.Fraction..	Silt.Fraction..
1	INCEPTISOLS	Csa	0.20	0	95	2
2	INCEPTISOLS	Csa	0.16	0	97	2
3	INCEPTISOLS	Csa	1.10	0	71	23
4	INCEPTISOLS	Csa	0.70	0	95	4
5	INCEPTISOLS	Csa	0.56	0	97	1
6	INCEPTISOLS	Csa	0.70	0	72	22

	Clay.Fraction....	NO3..ppm.	X350
1	3	0.5	0.20623
2	1	5.3	0.16328
3	7	4.8	0.17392
4	1	0.7	0.20553
5	2	5.4	0.16861
6	6	0.4	0.18657

What do the columns represent? and where the spectra values are started?

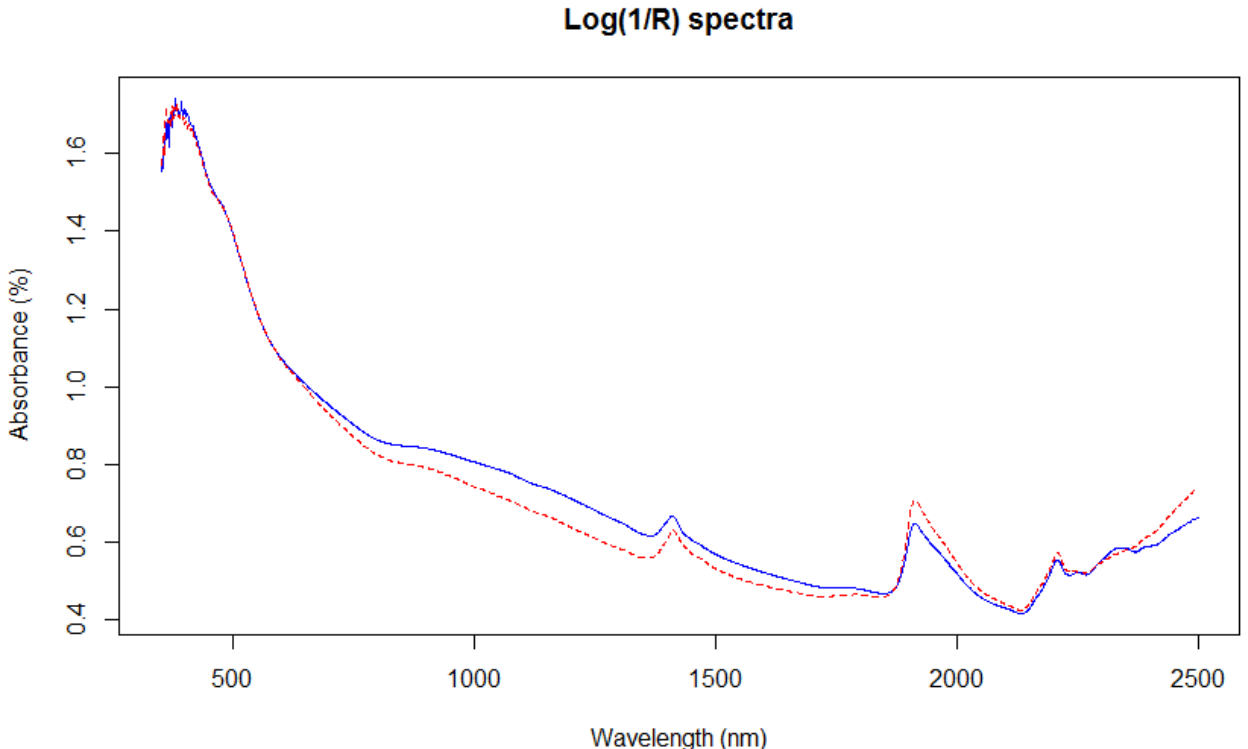
Data exploration

```
# Plotting a random selection of 3 spectra
plot( t(spectra[1,])~t(wavelength), type="l", lty=1, ylim=c(0,1), xlab = "Wavelength
(nm)", ylab = "Reflectance (%)", col = 4, main="Raw spectra")
points( t(spectra[10,])~t(wavelength), type="l", lty=2,col=2, ylim=c(0,1))
```



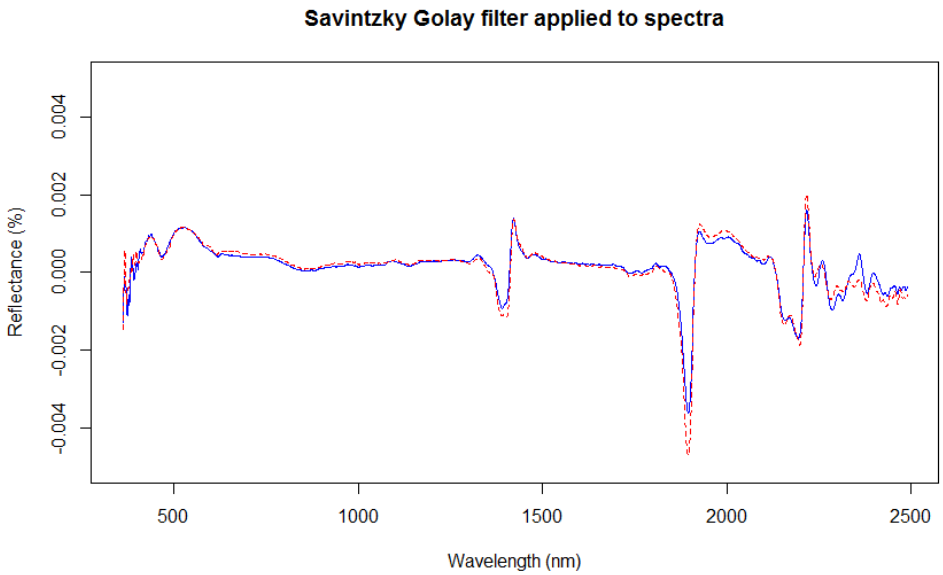
Data exploration

```
#log spectra
spectra_log<-log(1/spectra)
plot( t(spectra_log[1,])~t(wavelength), type="l", lty=1, xlab = "Wavelength (nm)",
ylab = "Absorbance (%)", col = 4, main="Log(1/R) spectra")
points( t(spectra_log[10,])~t(wavelength), type="l", lty=2,col=2, ylim=c(0,1))
```



Data exploration

```
# b)savitzky golay filter
spectra_sg<-savitzkyGolay (spectra, 1,3,21)
spectra_sg<- as.data.frame(spectra_sg)
wavelength_sg <- as.data.frame(t(c((360:2490))))
plot( t(spectra_sg[1,])~t(wavelength_sg), type="l", lty=1, xlab = "Wavelength (nm)",
ylab = "Reflectance (%)",ylim=c(-0.005,0.005), col = 4, main="Savintzky Golay filter
applied to spectra")
points( t(spectra_sg[10,])~t(wavelength_sg), type="l", lty=2,col=2)
```



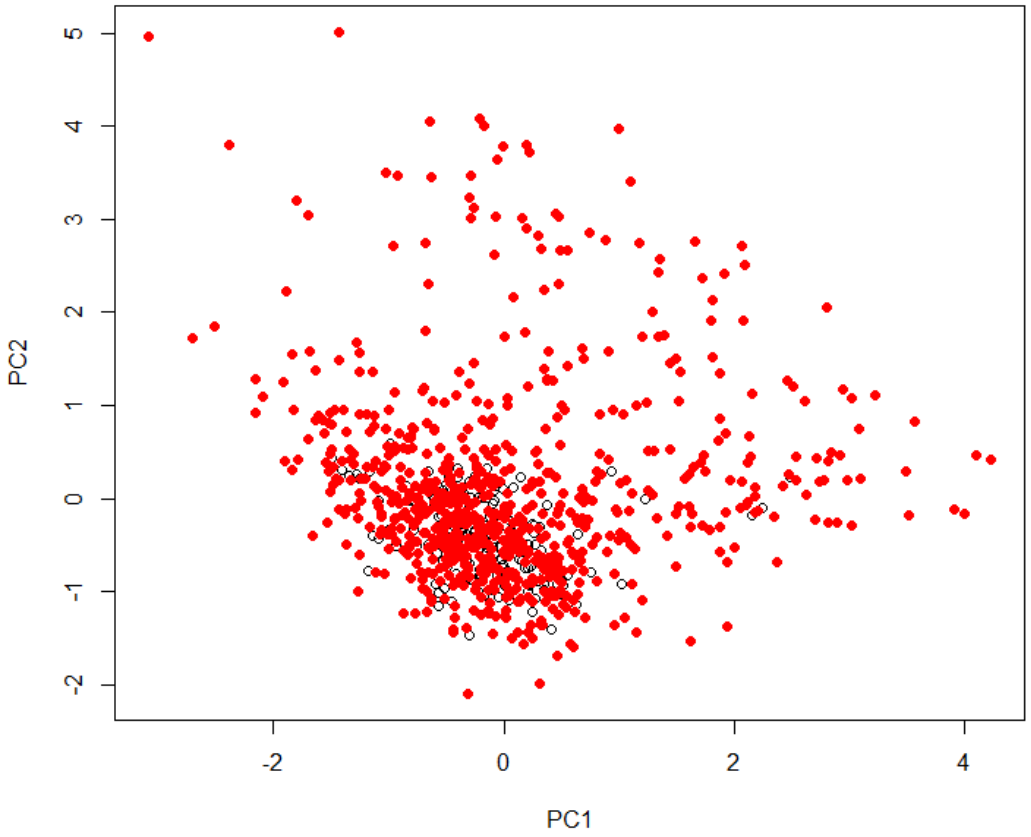
QUESTION: Assess the quality of the soil spectral signatures, explain the form of the curve in 1400 and 1950nm, how do you assess the quality of the data?

Calibration/Validation Dataset

```
# Select calibration samples from a large multivariate data using the Kennard-Stone algorithm
k <- kenStone(X = dataset$spectra, k = as.integer(4/5 * nrow(dataset)), pc = .99)

n_pcs <- 3
pc <- k$pc[,1:n_pcs]
chiMat <- matrix(NA, ncol=2, nrow=nrow(pc))
chiMat[,1] <- mahalanobis(pc[,1:n_pcs], colMeans(pc[,1:n_pcs]), cov(pc[,1:n_pcs]))
chiMat[,2] <- pchisq(c(chiMat[,1]), df = n_pcs)
ordered <- order(chiMat[,2])

# Create the testing (tst) and the calibration (cal) dataset
tst <- ordered[seq(1, length(ordered), 5)]
cal <- ordered[setdiff(c(1:length(ordered)), tst)]
```

Calibration/Validation Dataset

Kennard-Stone algorithm vs. random sample selection: Which one is better for splitting spectroscopy data to create calibration and testing datasets?



Machine learning modelling for predicting soil properties

The "train" function of caret package contains a number of regression models. Possible values can be found using names(getModelInfo()). See more

pls

```
c <- train(property ~ spectra,
           data = dataset[cal,],
           trControl=trainControl(method="cv",number=5),
           method="pls",
           tuneLength = 20)
```

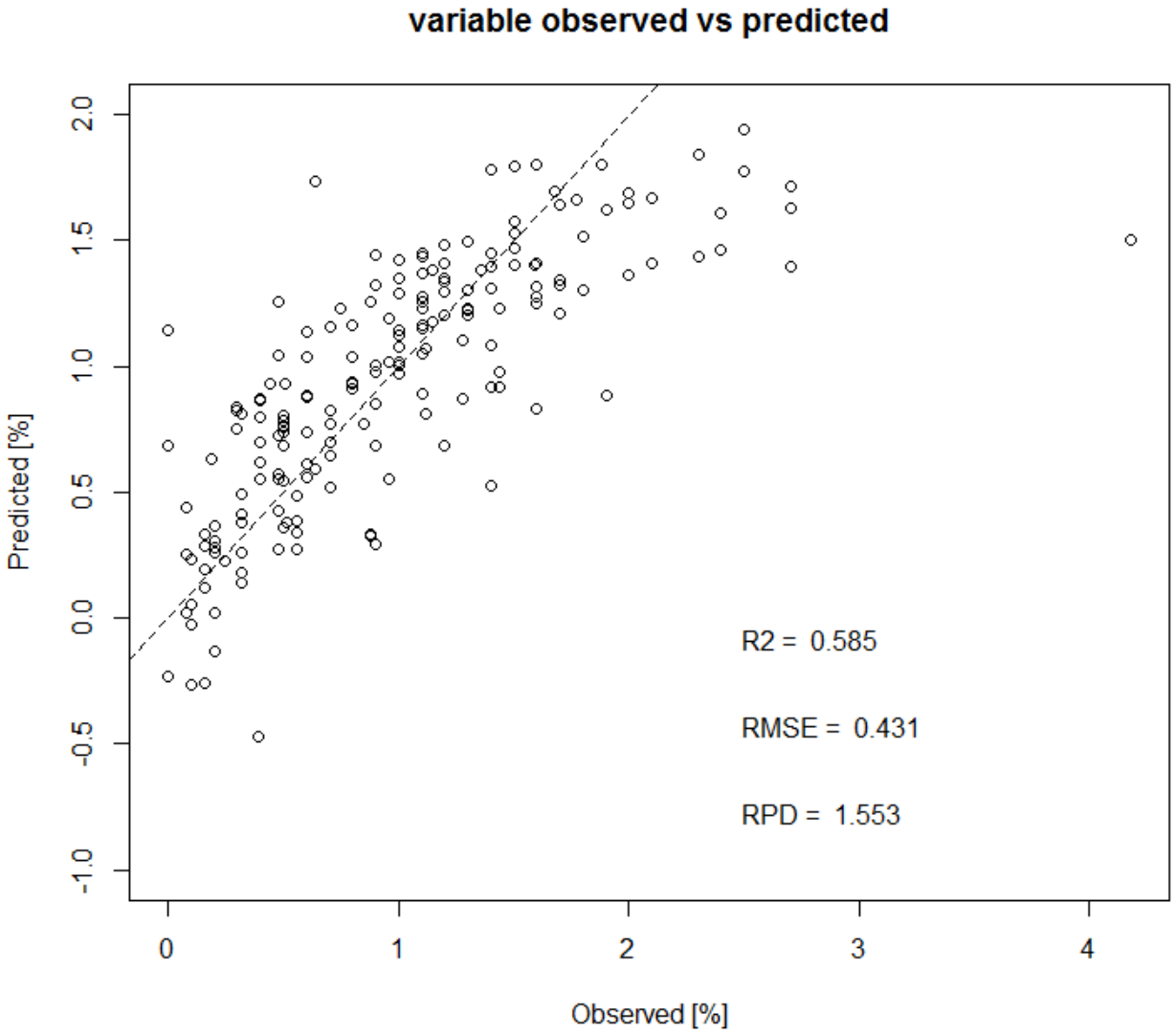
```
> summary(c)
Data:   X dimension: 742 2151
        Y dimension: 742 1
Fit method: oscorespls
Number of components considered: 11
TRAINING: % variance explained
          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
X           84.25    95.49    97.48    99.13    99.54    99.76    99.82    99.87    99.90     99.93     99.94
.outcome    15.13    19.20    29.64    33.90    39.31    44.68    48.76    52.07    55.96     57.44     58.89
```

Assessing the performance of prediction model

```
## Calculate R2, RMSE and RPD for the predicted values of the test set

R2_tst <- round(1 - mean((traindata$property - y_hat)^2) /
var(traindata$property), digits=3)

RMSE <- round(sqrt(mean((traindata$property - y_hat)^2)),digits=3)

RPD <- round((sd(traindata$property)/RMSE), digits=3)
```

Assessing the performance of prediction model

How do you judge the quality of the model based on prediction of the samples of the indepdenent test set?

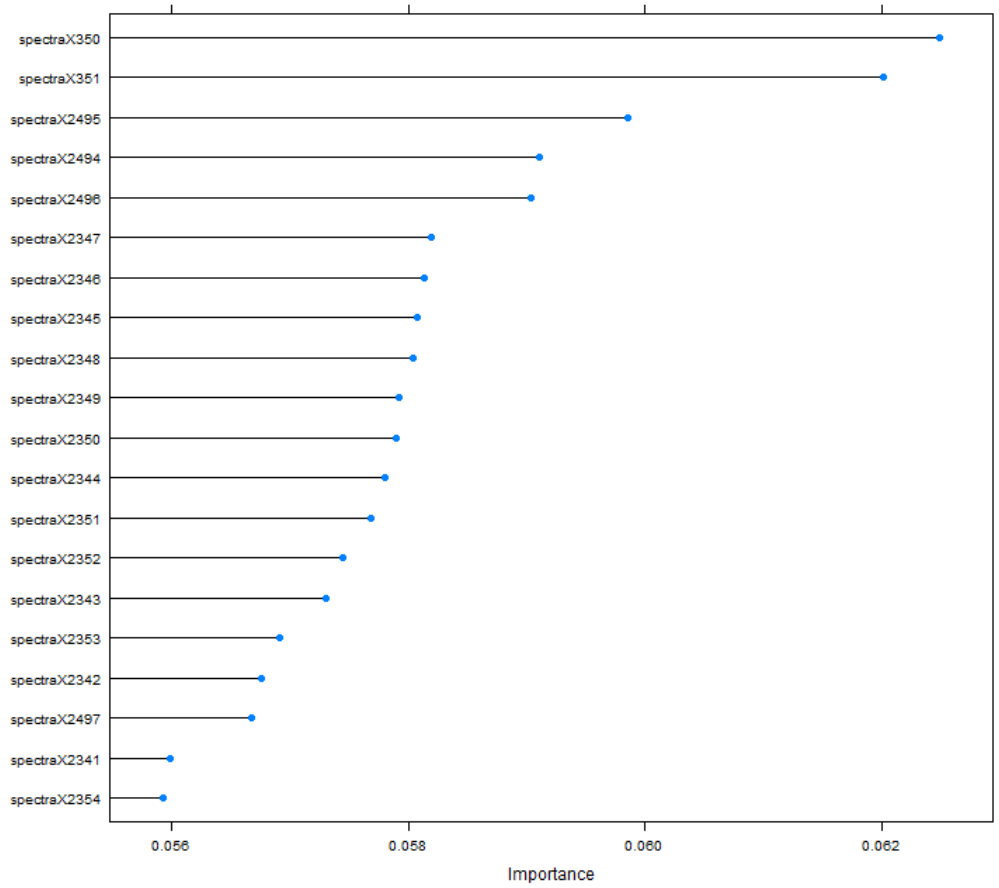


Variable Importance

gbmImp <- varImp(c, scale = FALSE)
plot(gbmImp, top = 20) # present only the 20 most important variables show

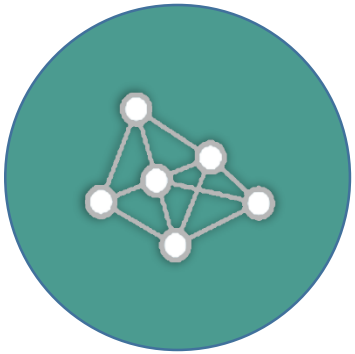
Partial Least Squares: the variable importance measure here is based on weighted sums of the absolute regression coefficients.

[Explore more about model specific metrics]



Challenge

Challenge: Now use the presented script to prepare a **cubist model** for one of the other soil parameters in the dataset like NO3 (column #15).



In addition, you could evaluate if a pre-treatment method (**savintzky golay**) would lead to a better model for the estimation of soil properties.

Questions?