Master Degree Internship :

Development of prediction models for C, N, Fe and Al in volcanic soils in Costa Rica using Infrared spectroscopy

Supervisors : Juan-Carlos MENDEZ (UCR) and Julien DEMENOIS (CIRAD)
Colaborators : Aurélie CAMBOU (IRD), Gilles CHAIX (CIRAD),
Cintya Solano (UCR)

ENS | PSL

UNIVERSIDAD DE COSTA RICA

cirad
AGRICULTURAL RESEARCH FOR DEVELOPMENT

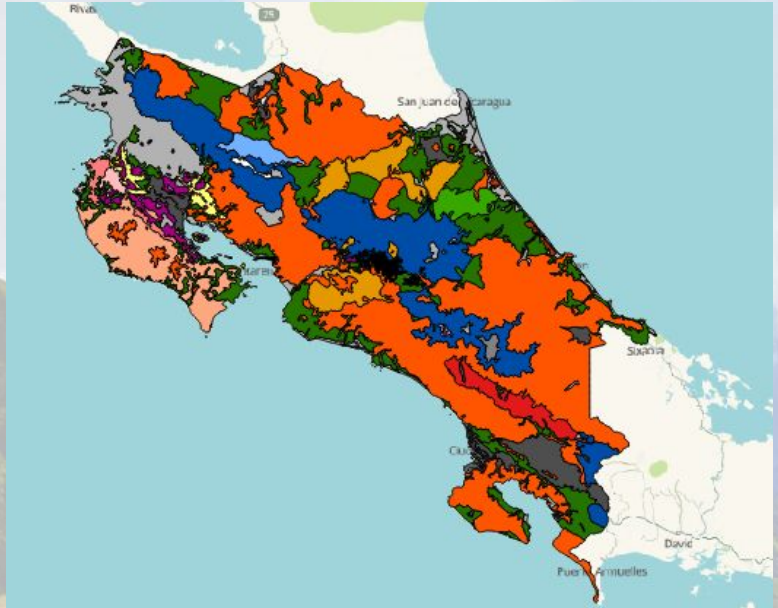
Introduction : do you know Costa Rica ?

Costa Rica : a small country in Central America, well known for its nature and a model of eco-tourism...

Introduction : do you know Costa Rica ?

Costa Rica : a small country in Central America, well known for its nature and a model of eco-tourism...

…But Costa Rica is also a country with intense agriculture (very fertile volcanic **andosols** and **ultisols**), and is the country in the world with the highest use of herbicides per km² *



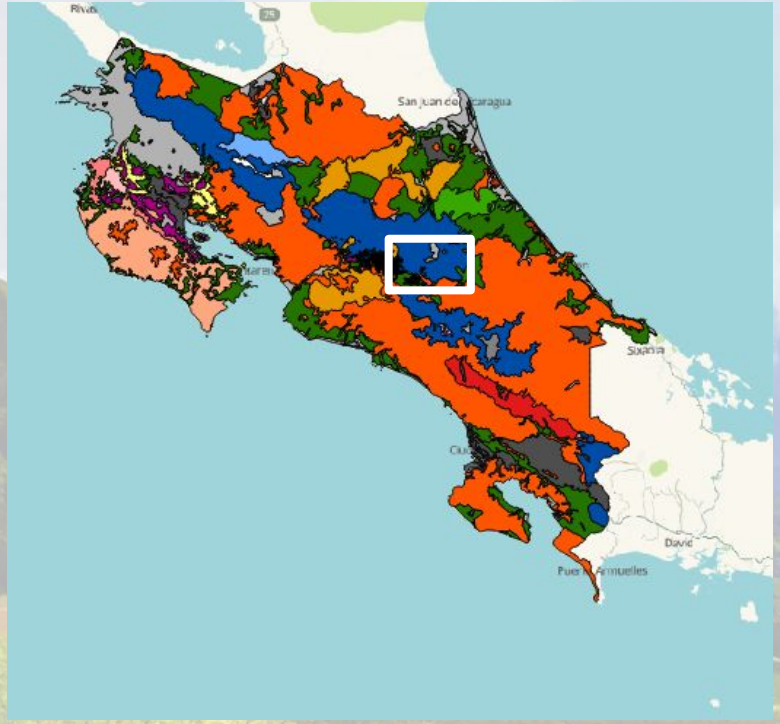
Costa Rica soil suborden maps : SuelosCR built by the UCR

* technically, it is third after the Maldives and Trinity and Tobago, but they both account for less than 0.1% of world global pesticide use (1500 tonnes/year), meanwhile Costa Rica is the 34th country in the world using most pesticides, with 12 811 tonnes/year. source : FAOSTAT

Introduction

The region of the Irazu and Turrialba volcano at the North of Cartago, is the most intensively cultivated, and supplies to the whole country.

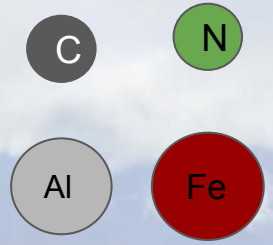
photo credit : Julien Demenois



Introduction

Classic monitoring of agricultural soil implies **laboratory analysis** of C,N, Al and Fe
-> time consuming, expensive and produces waste

photo credit : Julien Demenois



OBJECTIVE : Being able to assess C,N,Al and Fe from soil thanks to **infrared spectroscopy** (cheaper and faster) for a better monitoring of soil characteristics in the region.

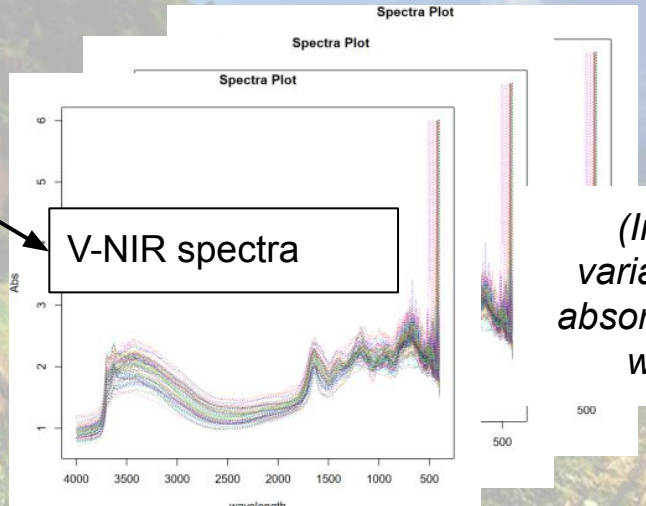
Quick reminder : How to develop a prediction model

Field samples

(Independant variables: altitude, temperature, precipitation, depth)

laboratory analysis

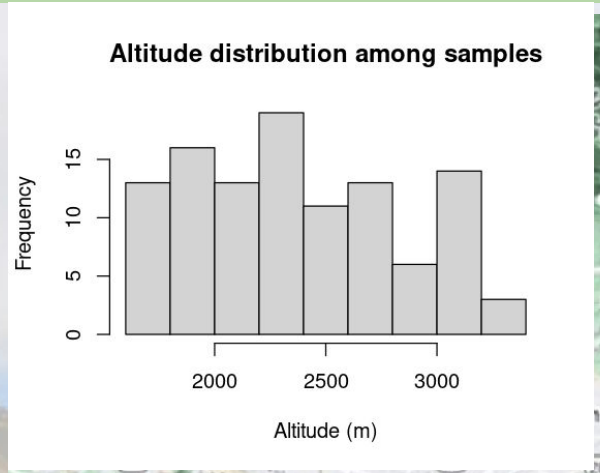
(Target variables : C, N, Fe, Al)



V-NIR spectra

(Independent variables: value of absorbance for each wavelength)

Sampling the Irazu volcano south flank…



At each point, several samples were taken at different depths

Spectroscopy and laboratory measurements

VNIR spectra (500nm - 2500nm) acquired with the FOSS DS2500 provided by CINA

MIR spectra (2000 - 25000 nm) acquired with a PERKIN ELMER provided by CICA (currently analysed)

C I N A

Laboratory analysis : lab provided by CIA (UCR)

SOC : **dry combustion** using C / N analyser (Dumas method)

Al / Fe : **selective dissolution** extraction by **ammonium oxalate**

cia
UNIVERSIDAD DE COSTA RICA
CENTRO DE INVESTIGACIONES AGRONOMICAS
FACULTAD DE CIENCIAS AGROALIMENTARIAS

Final dataset

The dataset is made of:

· A total of **108** samples, from 39 locations, with 2 to 10 horizons sampled at each location

· Environmental data : Soil type, soil subtype, altitude, land use, mean annual temperature, mean annual precipitation

· Laboratory measurement of Al, C, N and Fe for each sample

· V-NIR Spectra measurement for each sample

· MIR spectra measurement for each sample (not analysed yet)

Final dataset

A priori problems :

- 108 samples isn’t that much to make a PLSR model
- some of the data are strongly correlated (samples from the same hole…)

Final dataset

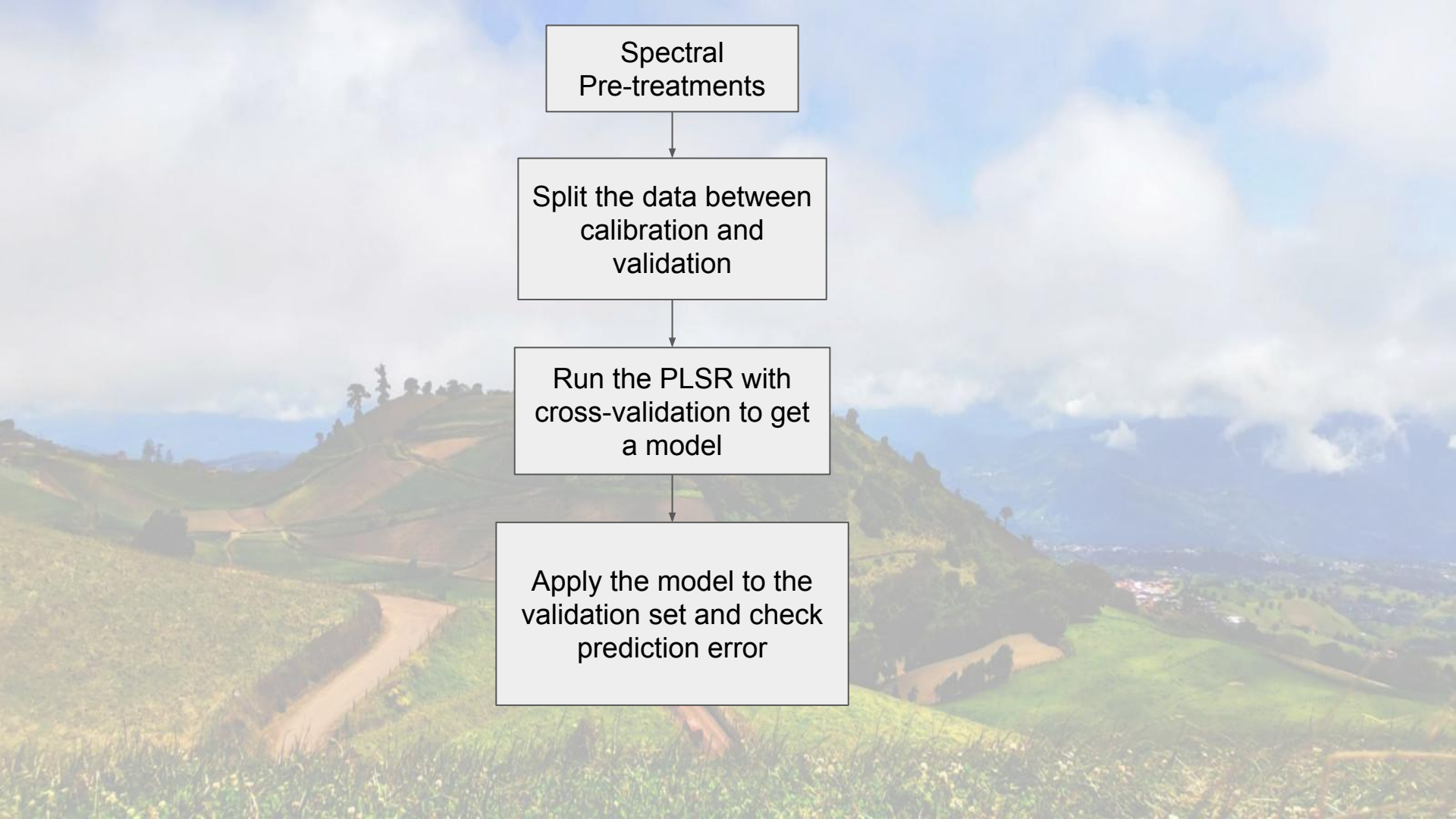
A priori problems of the dataset :

- 108 samples isn’t that much to make a PLSR model
- some of the data are strongly correlated (samples from the same hole…)

Idea to make a better model :

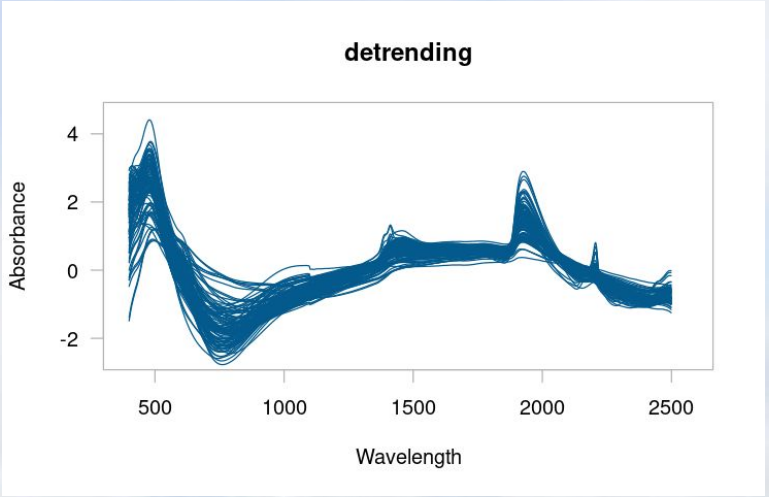
- use VNIR *and* MIR data (separately or together with spiking)
- **add environmental variables** (altitude, depth) as extra covariables

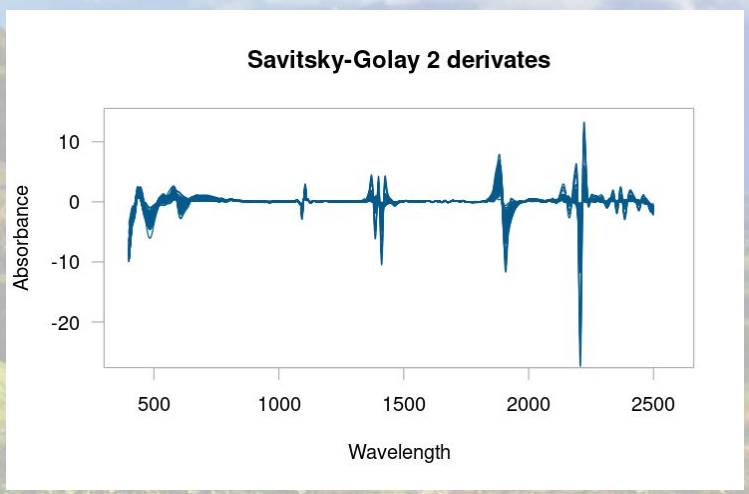
Calibrating the model


Spectral Pre-treatments
Split the data between calibration and validation
Run the PLSR with cross-validation to get a model
Apply the model to the validation set and check prediction error

Spectral Pre-treatments

for each element, 7 different pretreatments were tested (none, detrend, SNV, SavGol1/2, SNV+SavGol1/2)



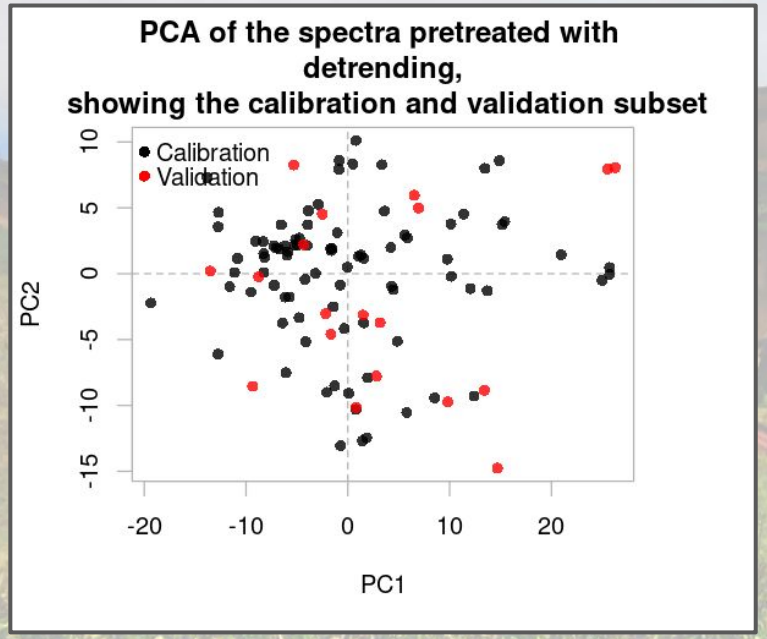


for each element, 7 different pretreatments were tested (none, detrend, SNV, SavGol1/2, SNV+SavGol1/2)

Spectral Pre-treatments

Split the data between calibration and validation

We used a custom Duplex sampling algorithm, enabling us to keep in a same group the samples from the same geographic point -> independence between calibration and validation



for each element, 7 different pretreatments were tested (none, detrend, SNV, SavGol1/2, SNV+SavGol1/2)

Spectral Pre-treatments

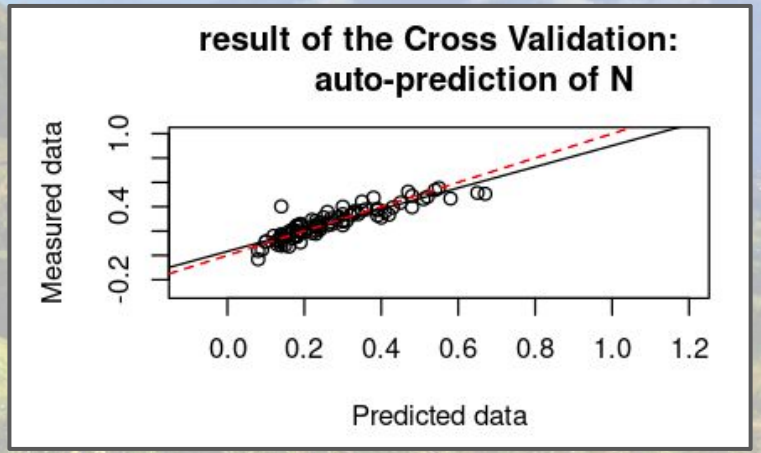
Split the data between calibration and validation

We used a custom Duplex sampling algorithm, enabling us to keep in a same group the samples from the same geographic point -> independence between calibration and validation

Run the PLSR with cross-validation to get a model

The PLSR was run with R package ***rnirs*** and used 3 group of cross-validation sampled with the K-foldings method, with 10 replicates.

We selected at this step the number of Latent variables (LV) for which the RMSECV was the lowest.



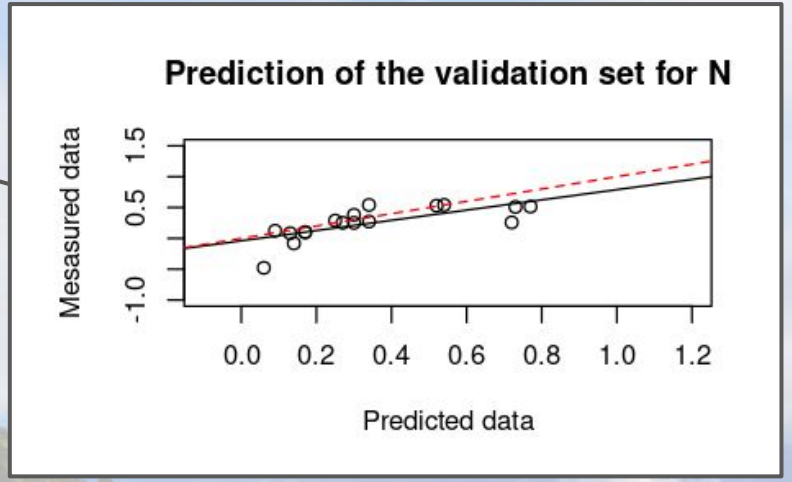
for each element, 7 different pretreatments were tested (none, detrend, SNV, SavGol1/2, SNV+SavGol1/2)

Spectral Pre-treatments

Split the data between calibration and validation

Run the PLSR with cross-validation to get a model

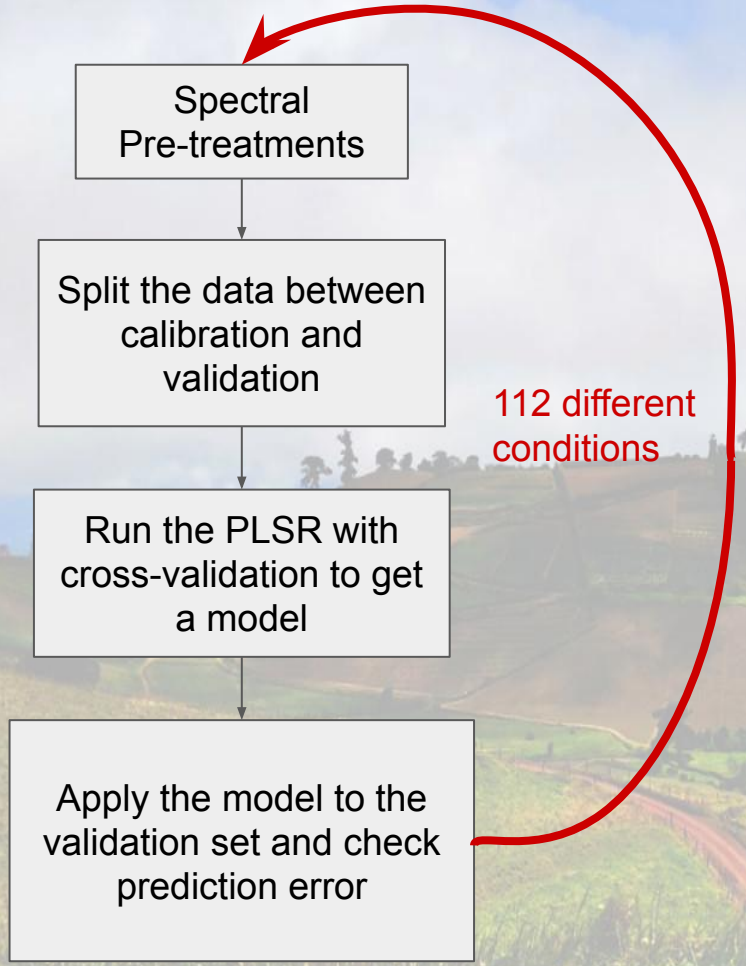
Apply the model to the validation set and check prediction error



The PLSR was run with R package ***rnirs*** and used 3 group of cross-validation sampled with the K-foldings method, with 10 replicates.

We selected at this step the number of Latent variables (LV) for which the RMSECV was the lowest.

For each element and each pretreatment, we looked at the RPD of the prediction of the validation. If **RPD>1.6**, we accept the model.



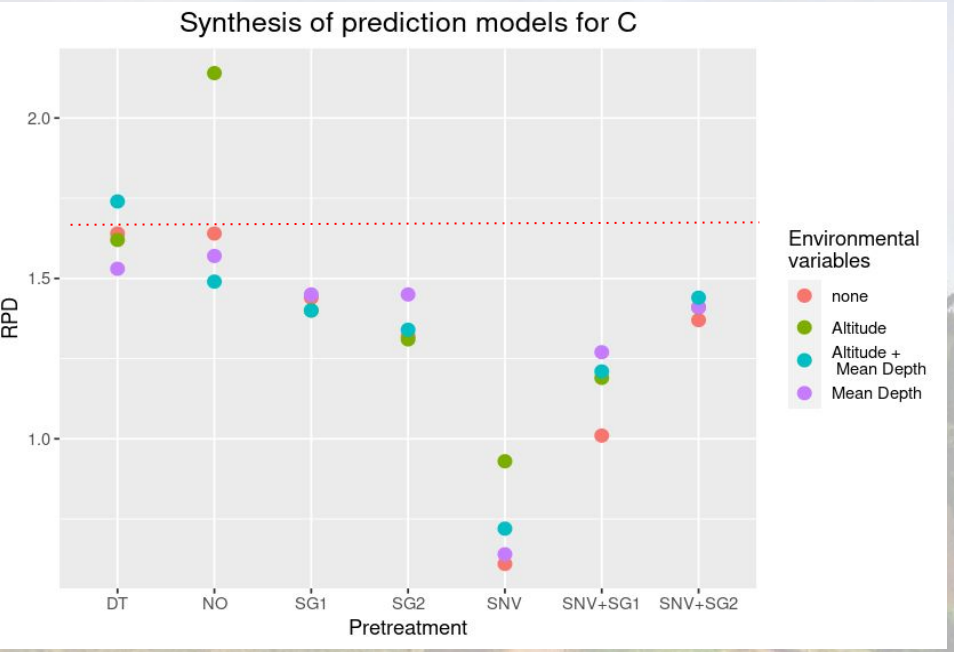
For each element (C, N, Fe, Al) :

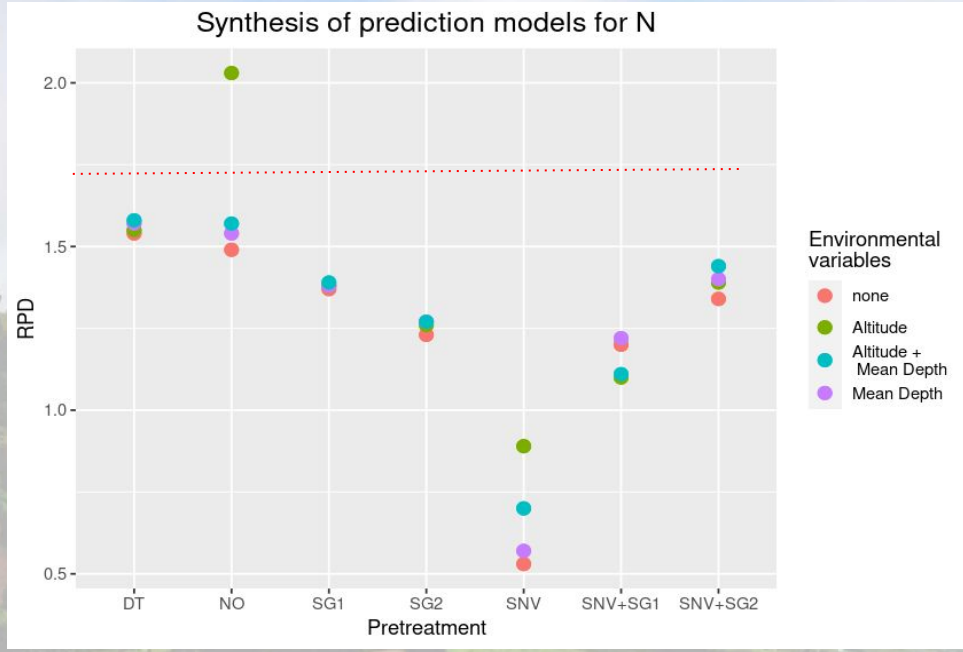
For each of the 7 pretreatments :

For each combination of environmental variables : without, with Altitude, with depth, with altitude+depth

-> 112 PLSR models (28 per element) were run

Synthesis of the results

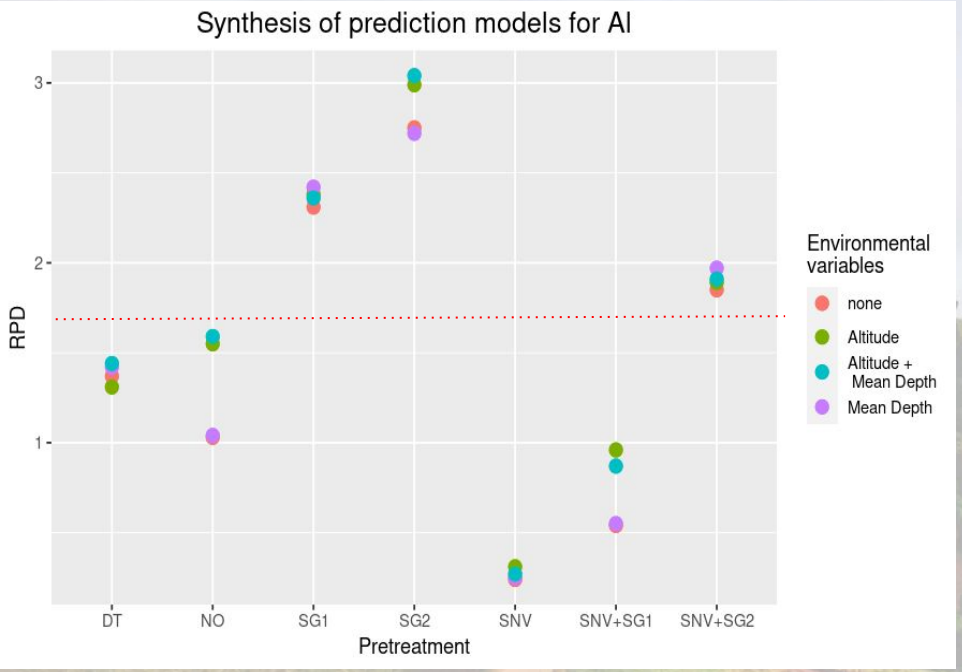


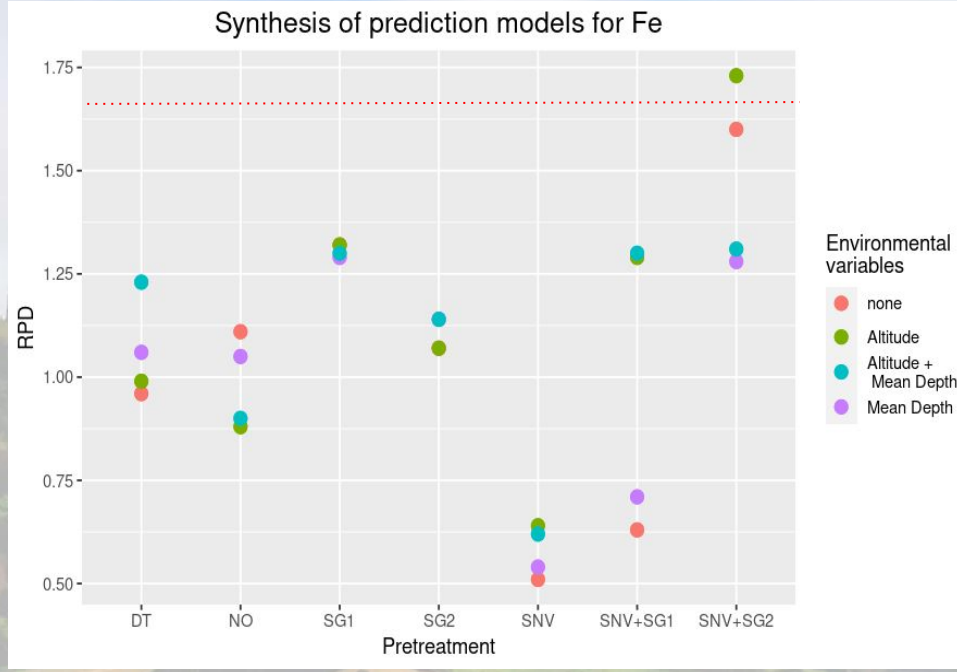


RPD = SDcal / RMSEP

The prediction for C and N is better with lighter/no pretreatments, and improved when we add field covariables

Synthesis of the results : prediction of Al and Fe

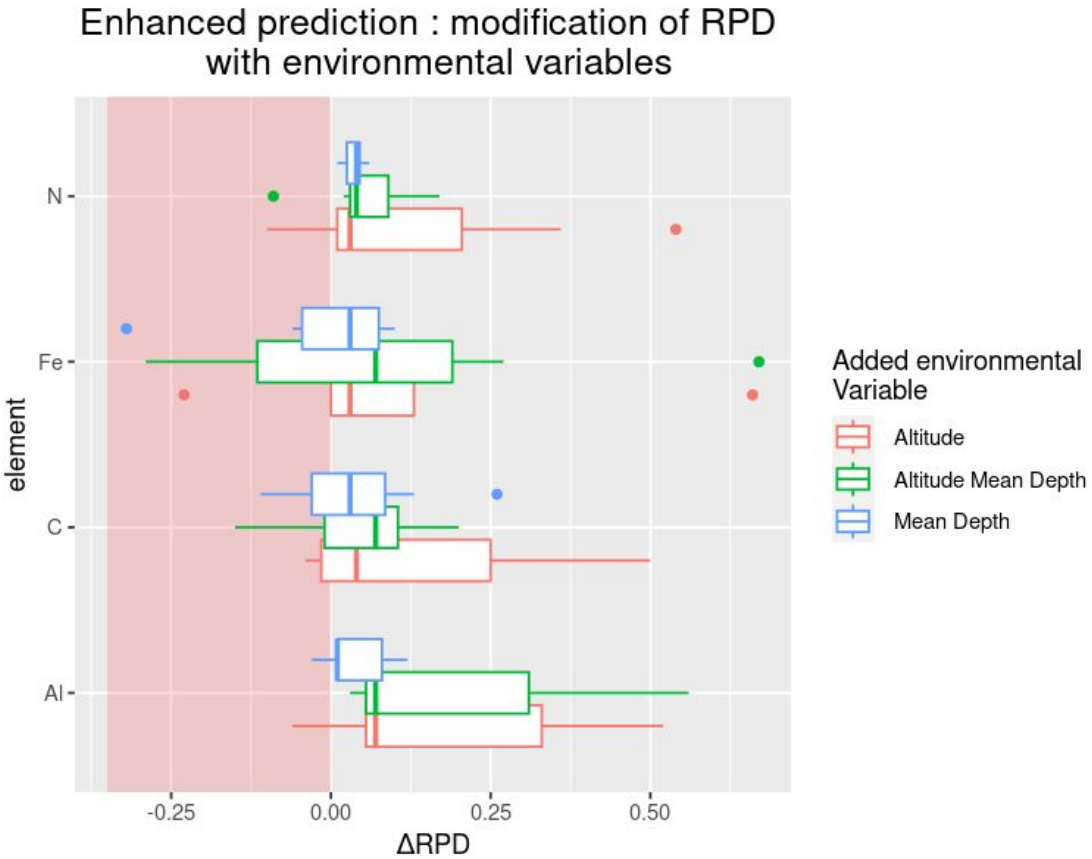




RPD = SD_{cal} / RMSEP

Fe was poorly predicted in almost every situation. Adding environmental variables on heavily-treated spectras seems helping.

Summary



· Adding environmental variables increased the prediction performance of most PLSR models

· For C, N and Al, we encountered some models with a good (RPD>1.6) prediction performance.

· Fe was poorly predicted, but with heavy pre-treatment and environmental variables, we managed to reach the RPD threshold

Limitations and further investigations

· Selection Cal/Val **after** the pretreatments

=> overfitting +

we don’t have the same Cal and Val groups for each model : can we really compare the different RPD with themselves ?

Limitations and further investigations

· Selection Cal/Val **after** the pretreatments
=> overfitting +
we don't have the same Cal and Val groups for each model : can we really compare the different RPD with themselves ?

Solution

· making the cal/val selection **before** the pretreatments to have the same groups

and/or

· making the cal/val selection based on the explanatory variables (y growing) rather than on the spectra

That’s it!
Thanks for your attention!

photo credit for the background : Cintya Solano (UCR)