Memory based learning methods and tools: towards efficient modelling, predicting and managing tasks in large scale soil spectral libraries

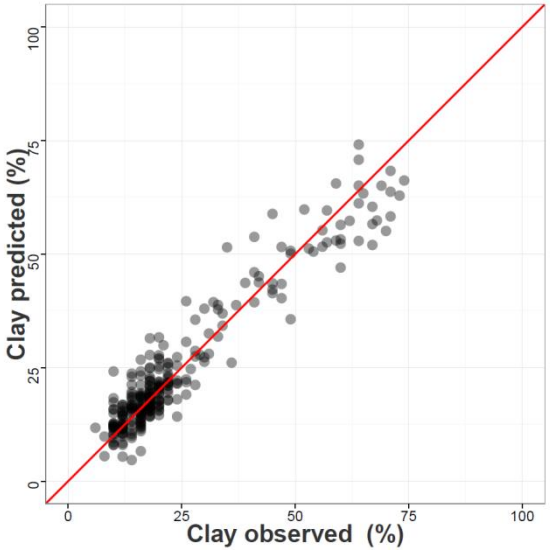
Leonardo Ramirez-Lopez [1,2]*
Antoine Stevens [3]

[1] Swiss Federal Institute of Technology, Zürich (ETHz), Switzerland
[2] Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Switzerland
[3] Universite Catholique de Louvain, Belgium

Soil vis–NIR libraries

Field scale

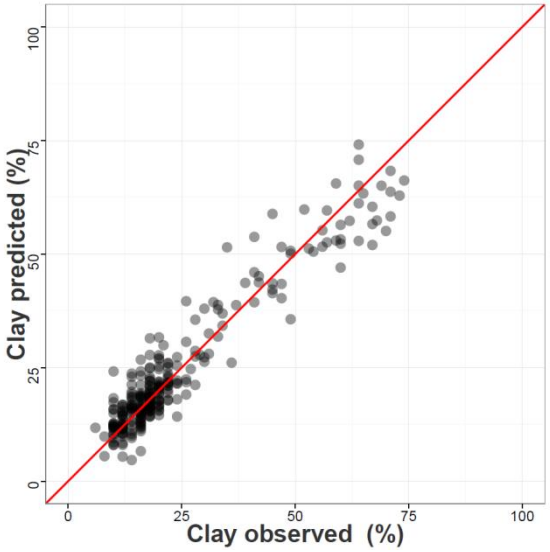
R^2 = 0.91; RMSE=5.02%



Soil vis–NIR libraries

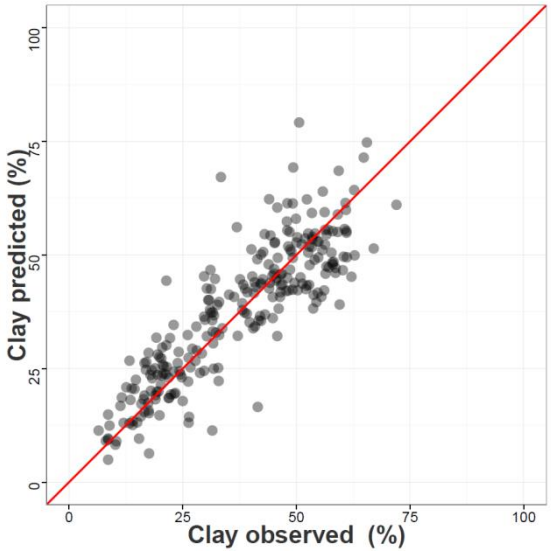
Field scale

R^2 = 0.91; RMSE=5.02%



Regional scale

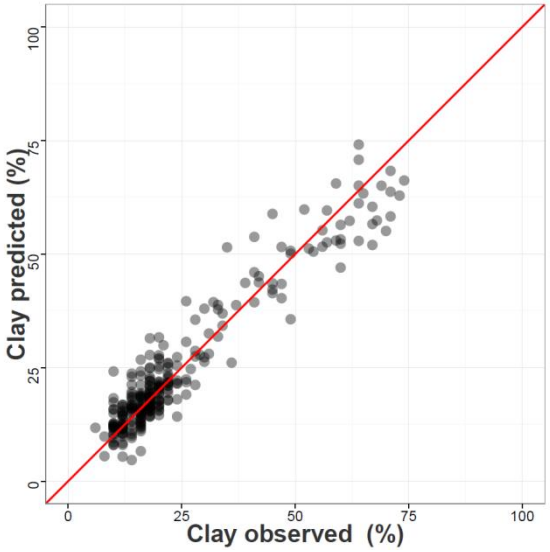
R^2 = 0.75; RMSE=8.03%



Soil vis–NIR libraries

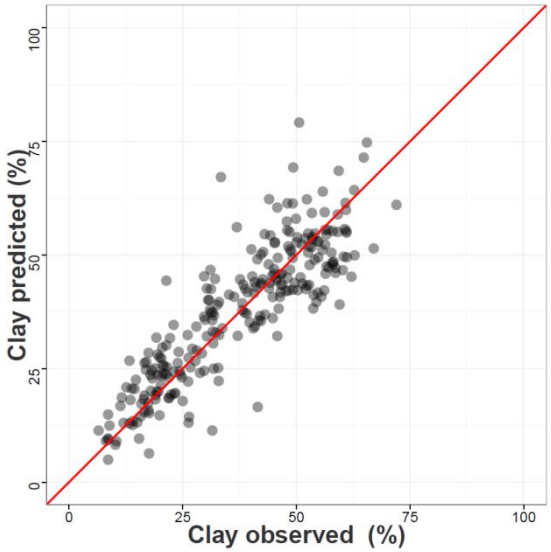
Field scale

R^2 = 0.91; RMSE=5.02%



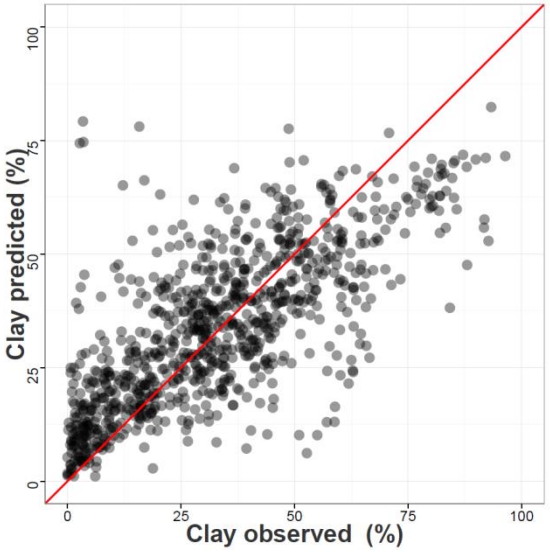
Regional scale

R^2 = 0.75; RMSE=8.03%



Global scale

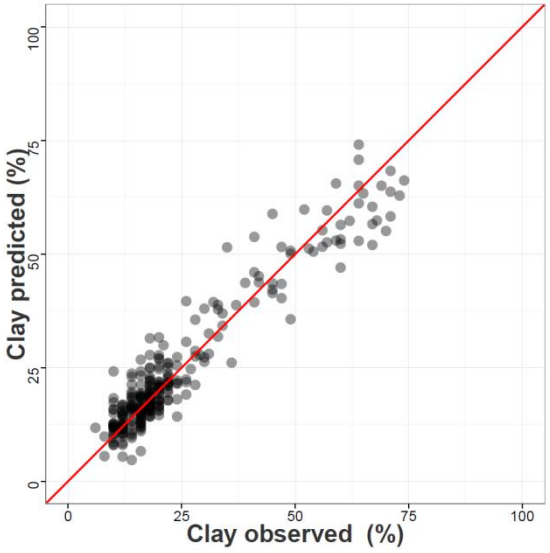
R^2 = 0.50; RMSE=15.53%



Soil vis–NIR libraries

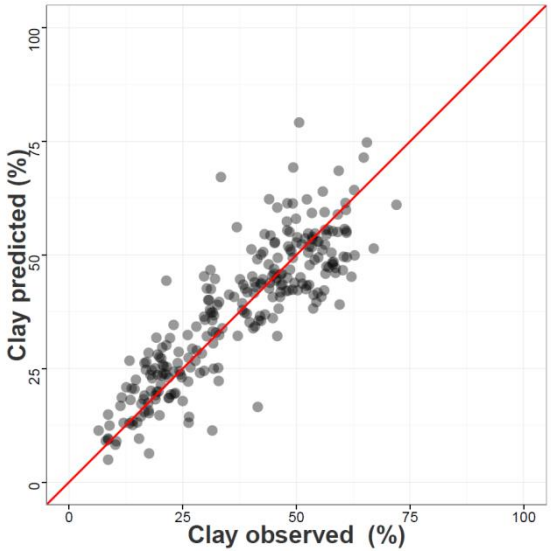
Field scale

R^2 = 0.91; RMSE=5.02%



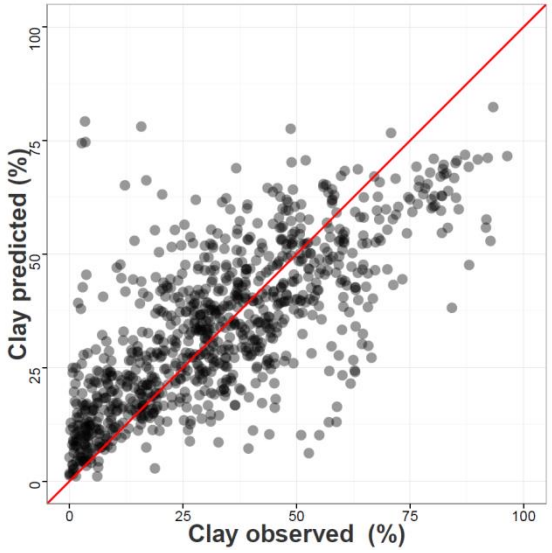
Regional scale

R^2 = 0.75; RMSE=8.03%



Global scale

R^2 = 0.50; RMSE=15.53%

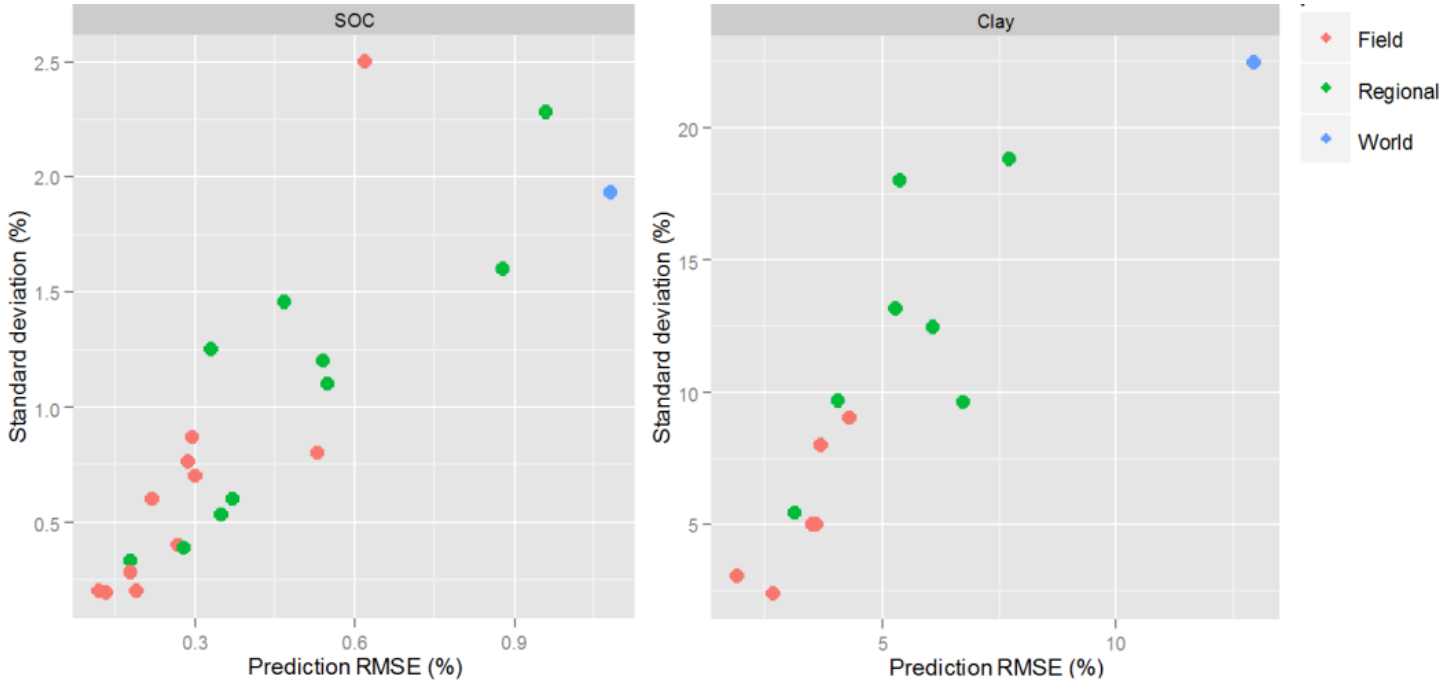


\− Data complexity +

Degradation of the accuracy

Soil vis–NIR libraries

Reported root mean square error (RMSE) of vis–NIR based predictions against the standard deviation (of the soil attribute) in the calibration sets

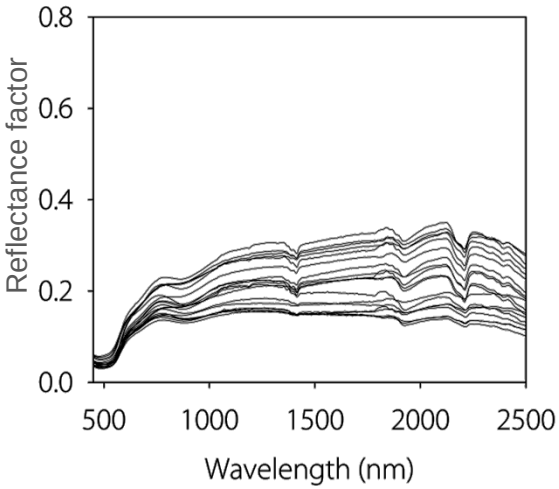


Analysis based on:

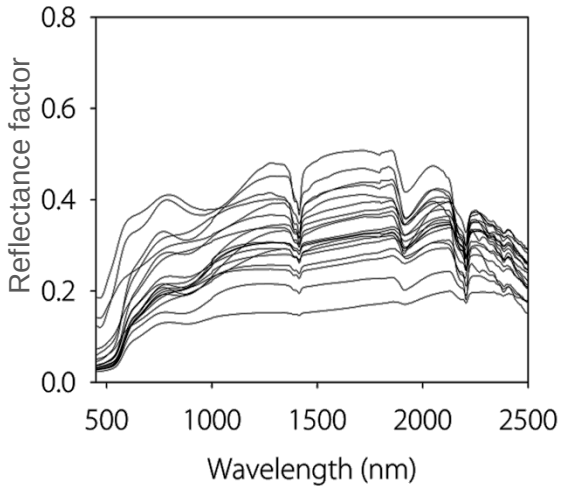
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. Visible and Near Infrared Spectroscopy in Soil Science. In: Donald L. Sparks, Ed. Advances in Agronomy, Vol. 107, Burlington: Academic Press, 2010, pp. 163-215.

Why are big soil vis–NIR libraries so complex?

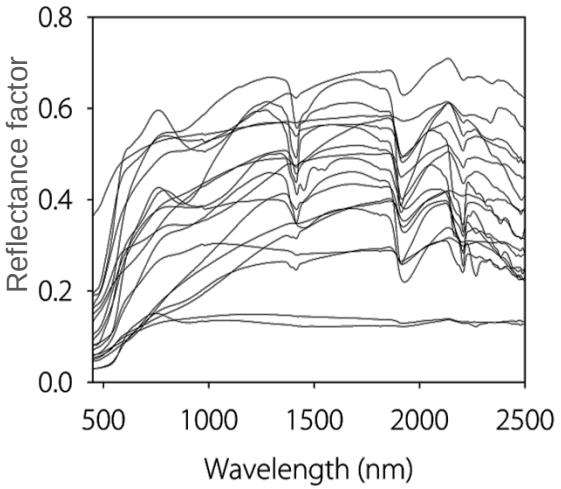
Field scale



Regional scale



Global scale



20 spectra sampled at random

Memory–based learning (MBL)

In contrast to other machine learning approaches, memory based learners do not attempt to derive a general target function. Instead, they offer instance-oriented solutions.

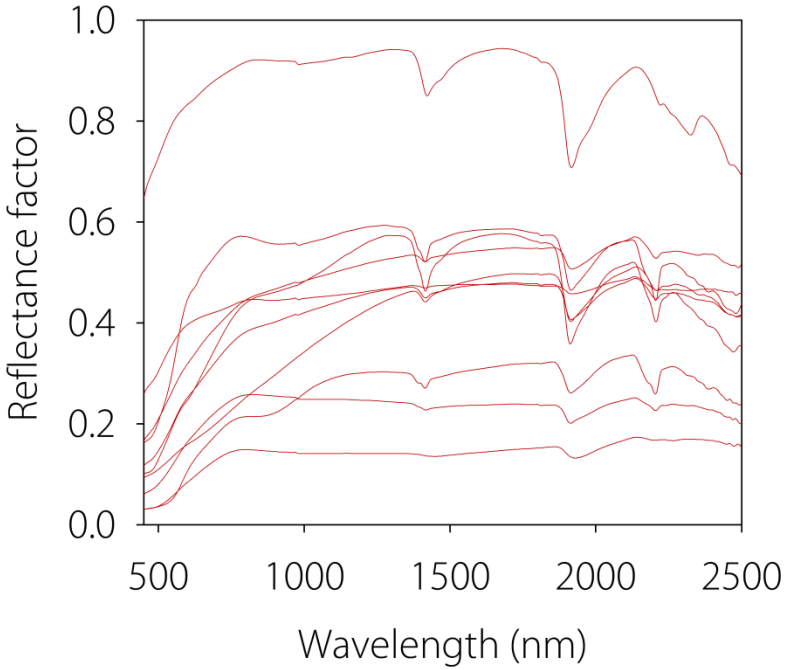
MBL is closely related to case based–reasoning (CBR) which emulates the human reasoning process:

1. Remember previous situations
2. Adapt them for solving the current problem
3. Examine the probability to solve the problem with the new solution
4. Memorize the experience for improving knowledge

MBL for soil spectral libraries

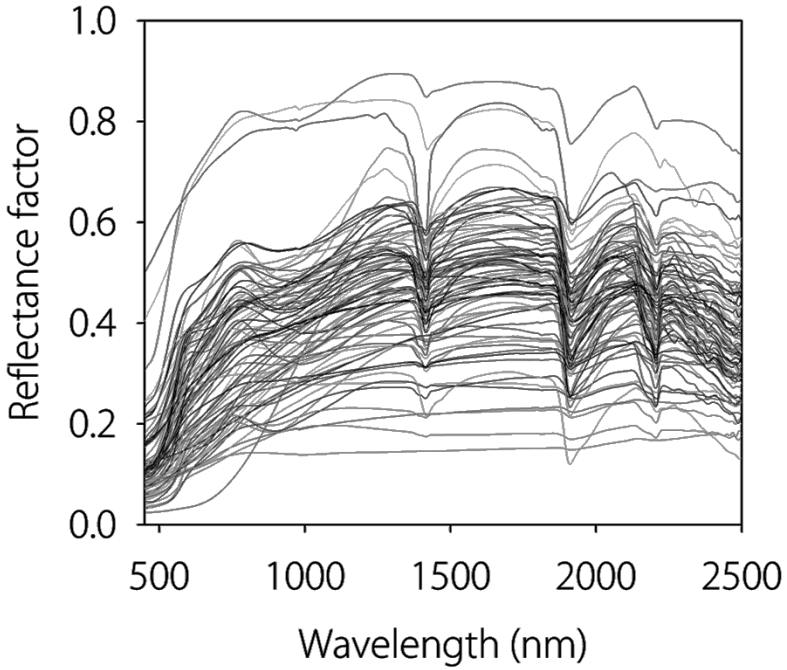
Soil attribute with unknown values

$$(\mathrm{Xu}, \mathrm{Yu}) = \{\mathrm{xu}_i, \mathrm{yu}_i\}_{i=1}^{m}$$



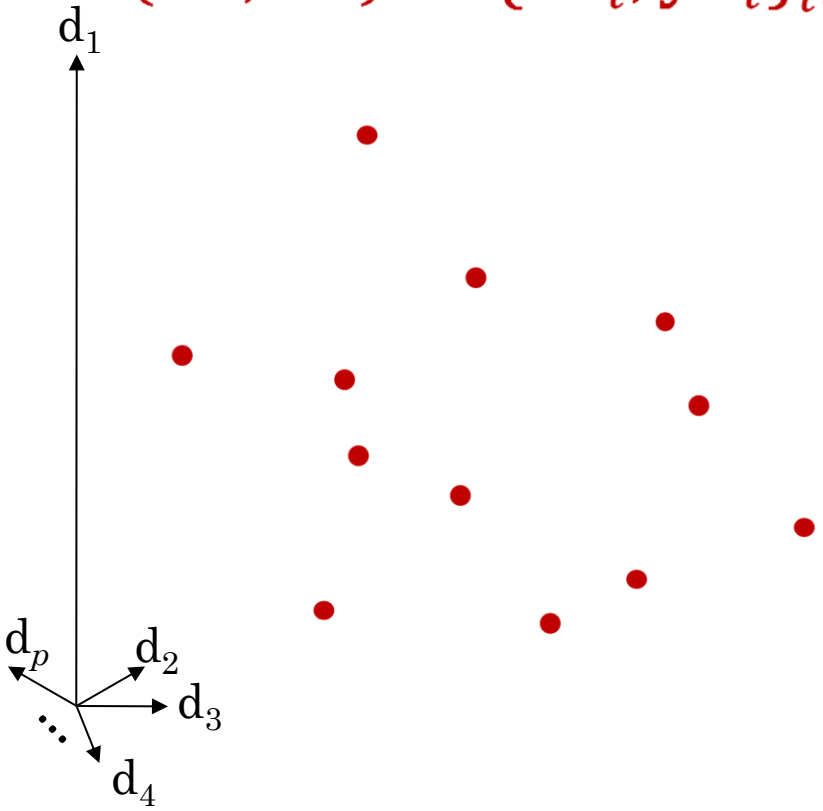
Soil spectral library

$$(\mathrm{Xr}, \mathrm{Yr}) = \{\mathrm{xr}_j, \mathrm{yr}_j\}_{j=1}^{n}$$

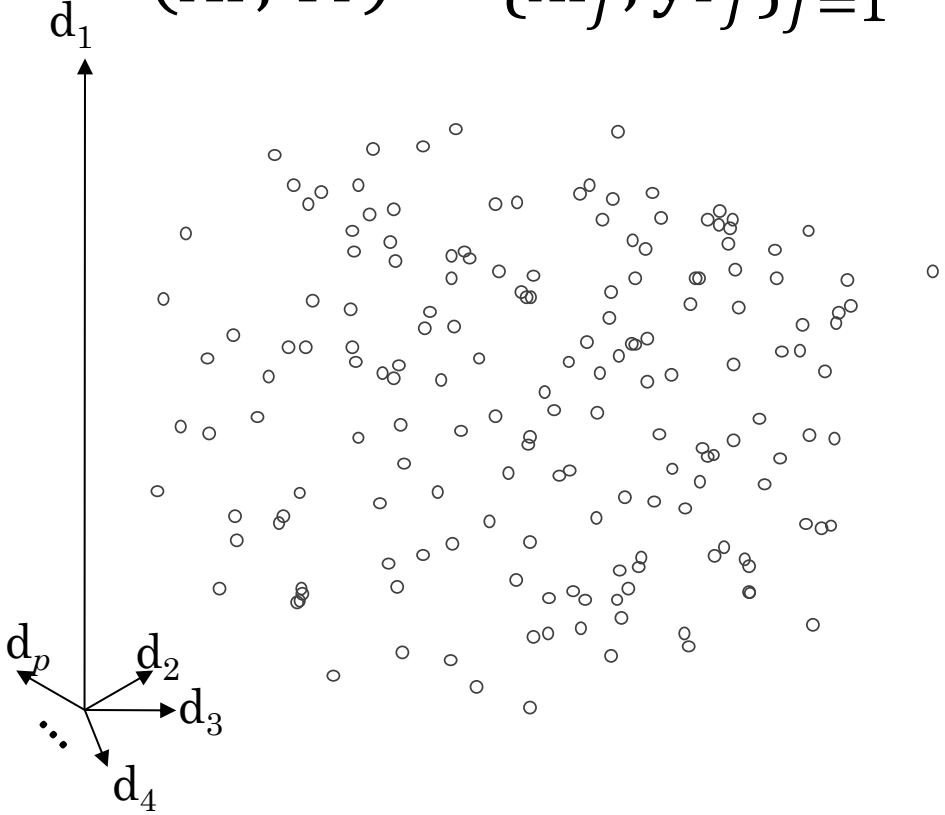


p–dimensional spectra

$$(Xu, Yu) = \{xu_i, yu_i\}_{i=1}^{m}$$



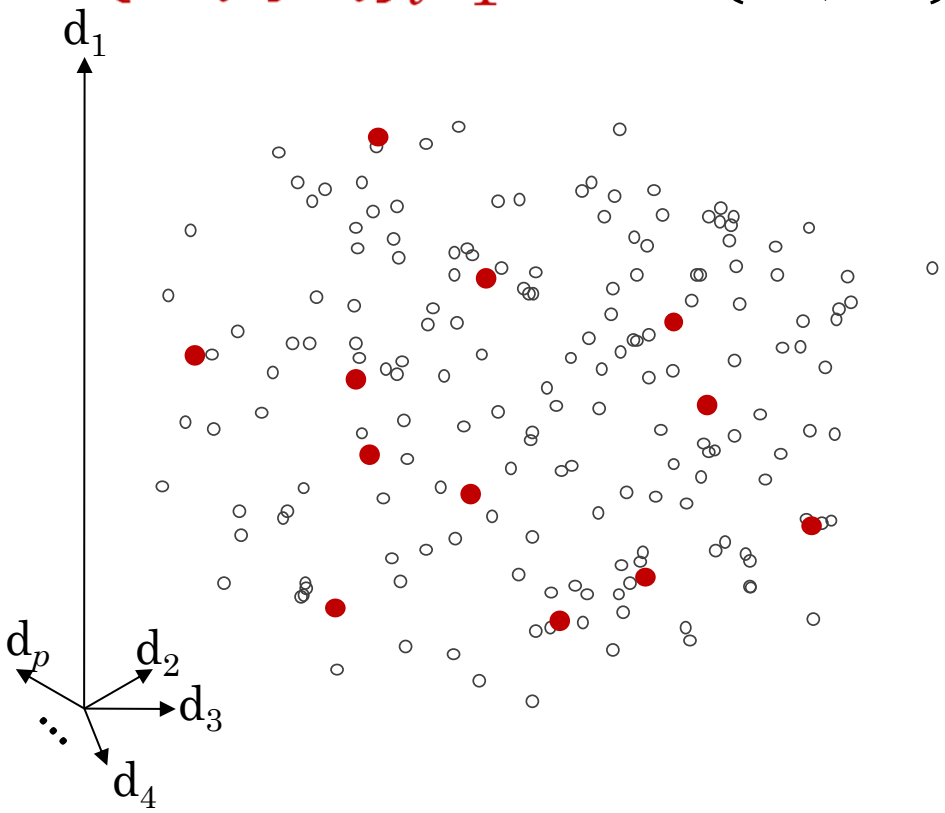
$$(Xr, Yr) = \{xr_j, yr_j\}_{j=1}^{n}$$



p–dimensional spectral feature space

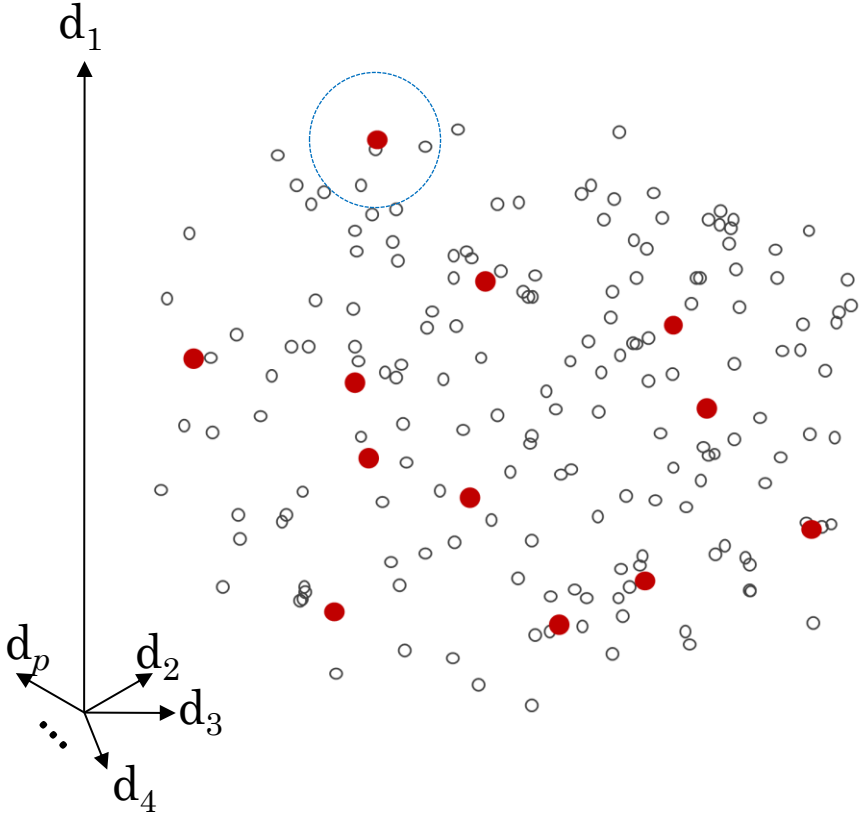
$$(\mathrm{Xu}, \mathrm{Yu}) = \{\mathrm{xu}_i, \mathrm{yu}_i\}_{i=1}^{m}$$

$$(\mathrm{Xr}, \mathrm{Yr}) = \{\mathrm{xr}_j, \mathrm{yr}_j\}_{j=1}^{n}$$



$$S(\mathrm{xu}_i, \mathrm{xr}_j)$$

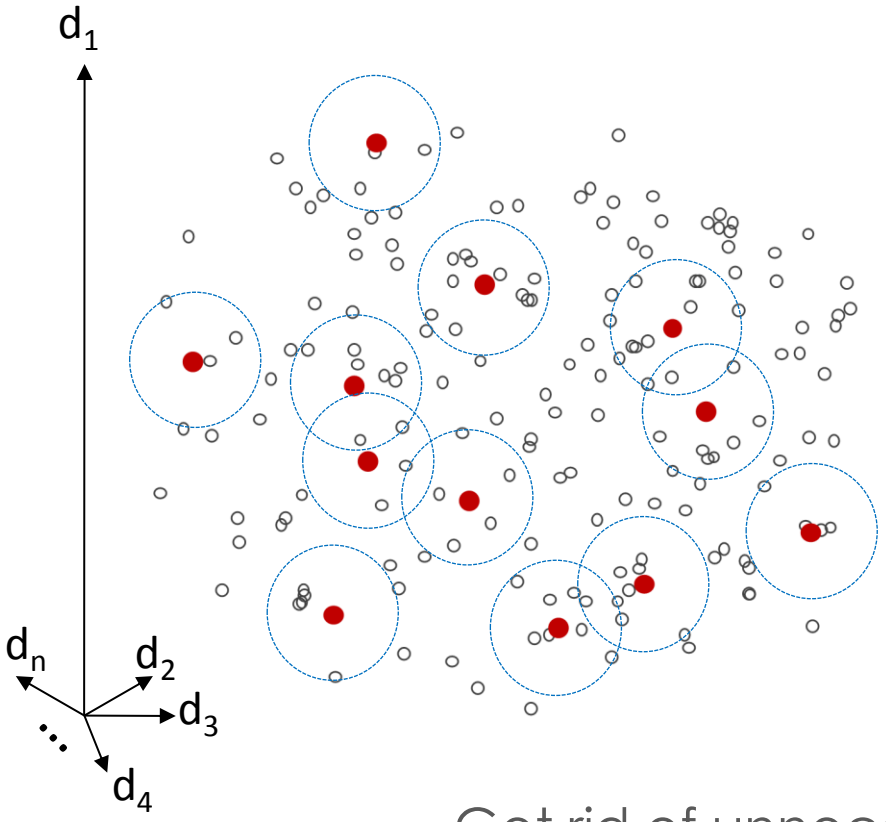
Sample neighbors



$$S(\mathbf{xu}_i, \mathbf{xr}_j)$$

a. Sphere neighbors OR *b*. *k*-NN

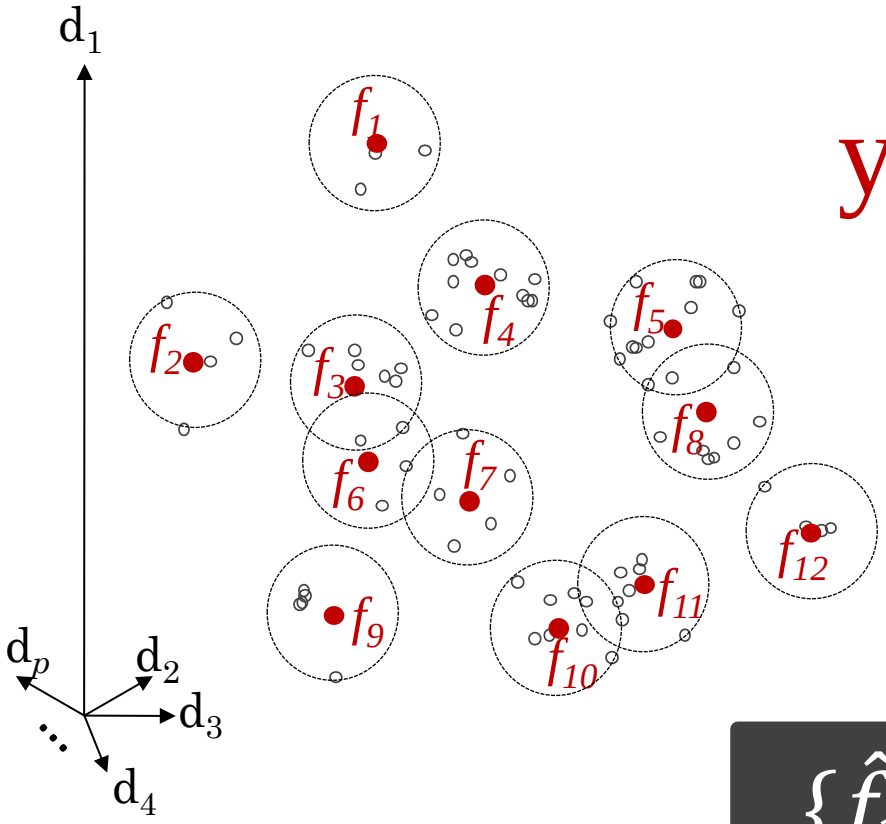
Sample neighbors



$$S(\mathbf{xu}_i, \mathbf{xr}_j)$$

Get rid of unnecessary samples

The solution to a complex task is constructed by a collection of simple local functions



$$y_i = \hat{f}_i(x;\theta) + \varepsilon$$

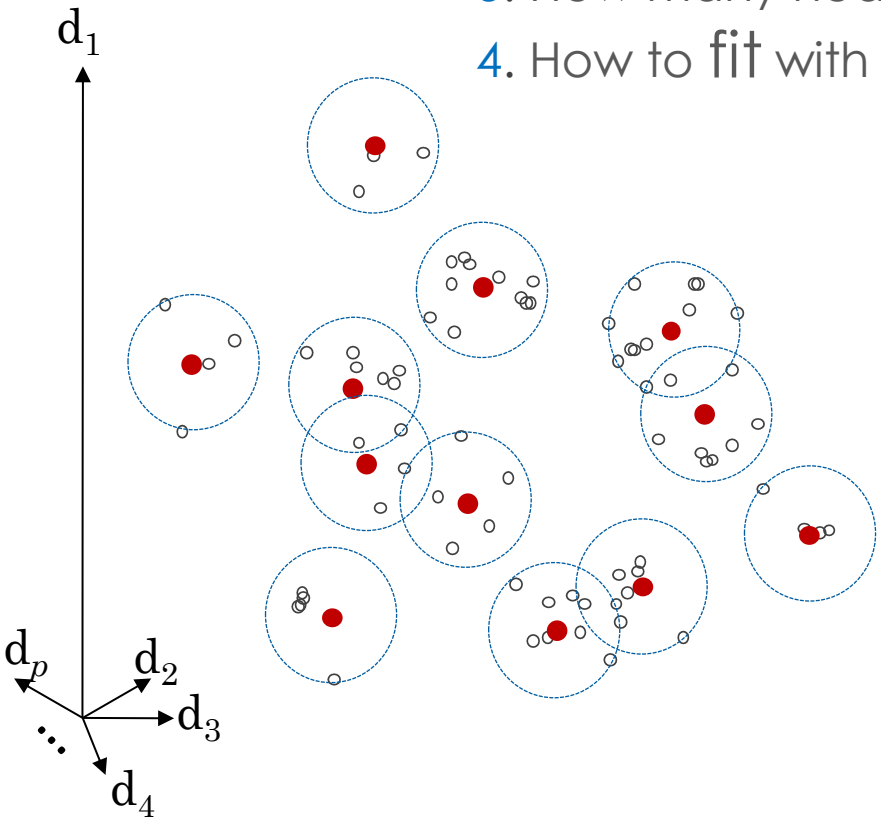
MBL

$$\{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m\}$$

Noisy and/or non-informative samples for *f* are ignored

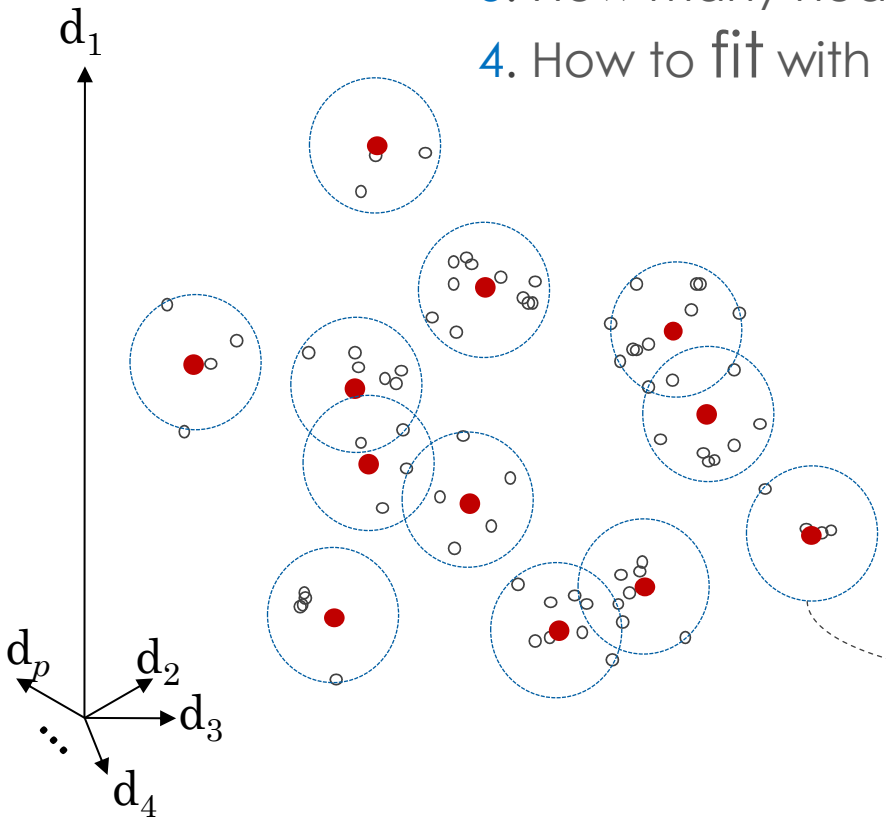
There are four basic aspects that must be defined for any MBL algorithm:

1. A similarity/dissimilarity metric
2. How to use the similarity/dissimilarity information
3. How many nearby neighbors to look at?
4. How to fit with the local points?



There are four basic aspects that must be defined for any MBL algorithm:

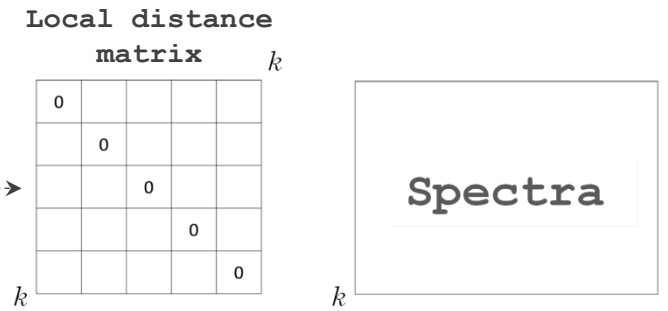
1. A similarity/dissimilarity metric
2. How to use the similarity/dissimilarity information
3. How many nearby neighbors to look at?
4. How to fit with the local points?



Option 1: Ignore it

Option 2: Use it for assigning weights

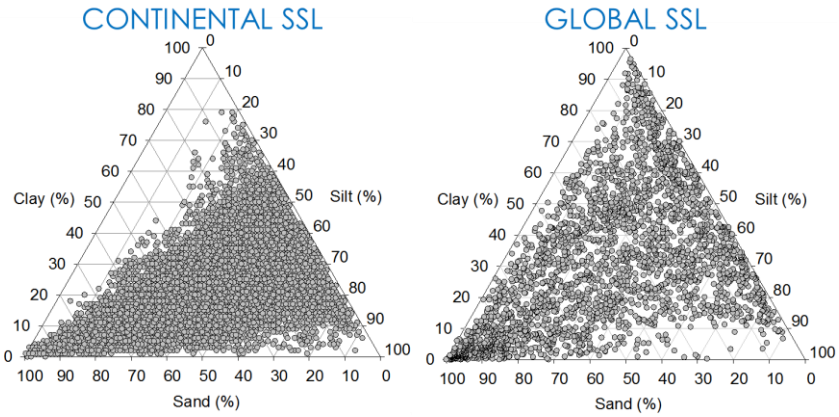
Option 3: Source of additional predictors…



Two (complex) soil vis-NIR libraries were used in order to test the performance of various MBL algorithms

	Total samples	‚Unknown set' (# samples)	Reference set (# samples)
Continental (LUCAS)*	19036	1000 (topsoil)	18036
World (ICRAF)**	3643	935 (168 profiles)	2078

*23 countries; 210 spectral variables
**55 countries; 216 spectral variables



Target attribute:

- Clay content

Algorithms tested:

- Partial least squares (PLS)
- Support vector machines (SVM)
- Random forest
- PLS-neural netwroks (PLS-NN)
- Cubist
- LOCAL
- Modified LOCAL
- Spectrum-based learner (SBL)
- Improved SBL(iSBL)

} Memrory-based learners

* Tóth, G., Jones, A., Montanarella, L. (eds.) 2013. LUCAS Topsoil Survey. Methodology, data and results. JRC Technical Reports. Luxembourg. Publications Office of the European Union, EUR 26102 – Scientific and Technical Research series – ISSN 1831-9424 (online); ISBN 978-92-79-32542-7; doi: 10.2788/979224

** World Agroforestry Centre (ICRAF) and ISRIC – World Soil Information. 2010. ICRAF-ISRIC Soil vis-NIR spectral Library. Nairobi, Kenya: World Agroforestry Centre (ICRAF).

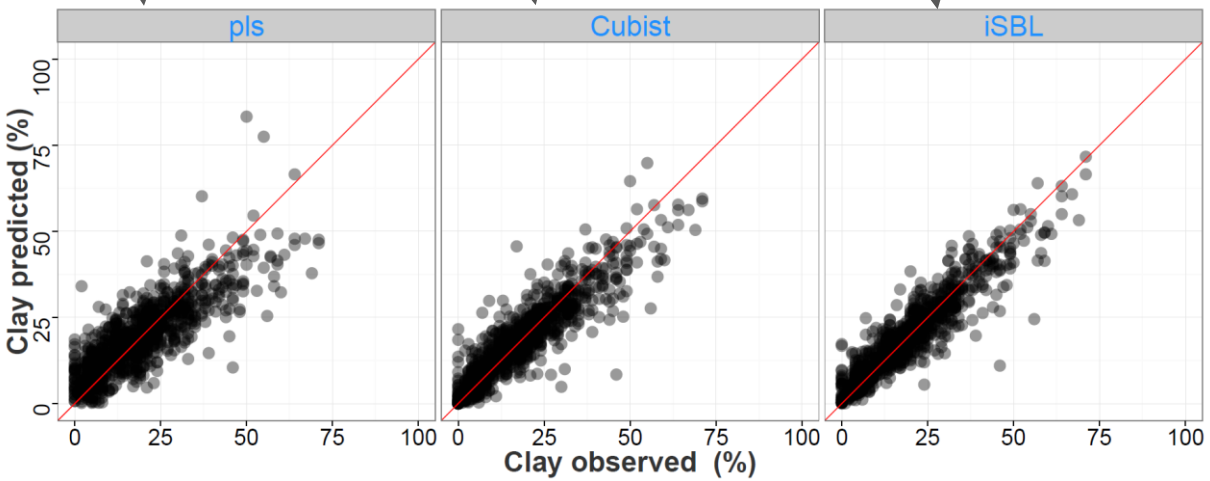
Case 1

Clay content prediction results

@ CONTINENTAL scale

	RMSE	R^2
iSBL	4.88	0.87
LOCAL	5.00	0.86
SBL	5.09	0.86
mLOCAL	5.33	0.85
Cubist	5.72	0.82
pls-NN	7.43	0.70
svm	7.59	0.69
pls	7.65	0.68
Rf	8.61	0.59

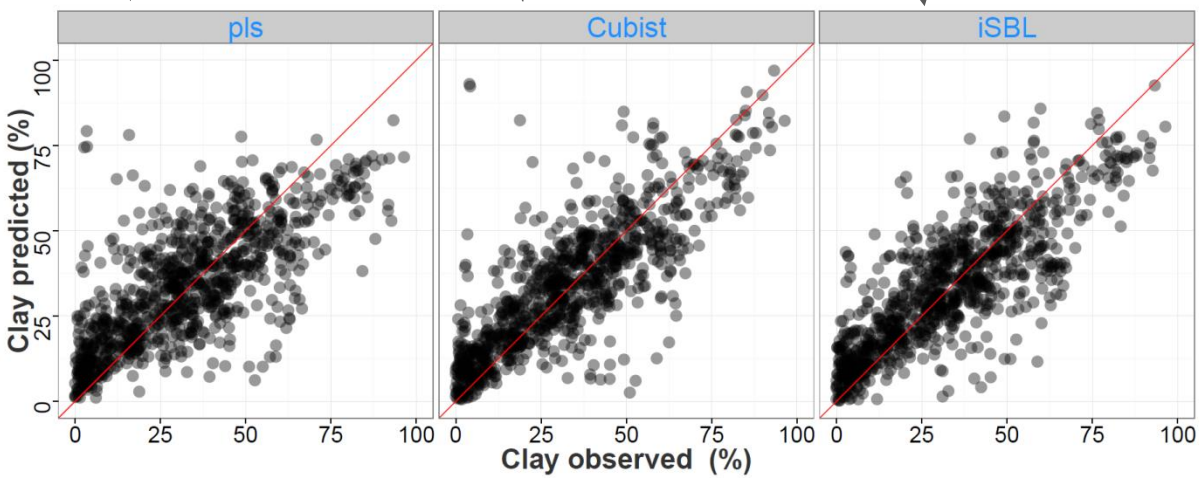
Memrory-based learners



Case 2
Clay content prediction results
@ GLOBAL scale

	RMSE	R^2
iSBL	12.50	0.67
SBL	12.53	0.67
mLOCAL	12.68	0.67
Cubist	12.80	0.66
LOCAL	13.19	0.64
pls-NN	14.60	0.59
svm	14.69	0.55
Rf	15.38	0.50
pls	15.53	0.50

Memrory-based learners



Spectroscopy-oriented software for MBL

Name	Platform	Algorithm(s) implemented
WinISI	FOSS software	LOCAL (Shenk *et al.*, 1997)
PLS_Toolbox	Matlab	Locally weighted PLS (LWR, Naes *et al.*, 1990)
resemble	**R** package	> 200 options (including LOCAL, LWR, mLOCAL, SBL, iSBL, etc)

Shenk, J.S., Westerhaus, M.O., Berzaghi, P. 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. Journal of Near Infrared Spectroscopy 5, 223–232.

Naes, T., Isaksson, T., Kowalski, B. 1990. Locally weighted regression and scatter correction for nearinfrared reflectance data. Analytical Chemistry 62, 664–673

Spectroscopy-oriented software for MBL

Name	Platform	Algorithm(s) implemented
WinISI	FOSS software	LOCAL (Shenk *et al.*, 1997)
PLS_Toolbox	Matlab	Locally weighted PLS (LWR, Naes *et al.*, 1990)
resemble	**R** package	> 200 options (including LOCAL, LWR, mLOCAL, SBL, iSBL, etc)

'resemble'

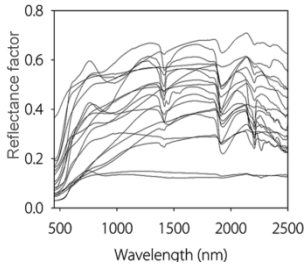
Methods for computing the spectral similarity/dissimilarity	Usage of the similarity/dissimilarity information	Local fit (regression algorithm)
• Euclidean • Mahalanobis • Spectral information divergence 1 • Spectral information divergence 2 • Correlation • Moving correlation • Cosine (spectral angle mapper) • 8 x Principal component • 8 x Partial least squares	• None • Predictors • Weights	• Gaussian process • Partial least squares (pls) • Weighted average pls 1 • Weighted average pls 2

Download the first (development) version which is available at :

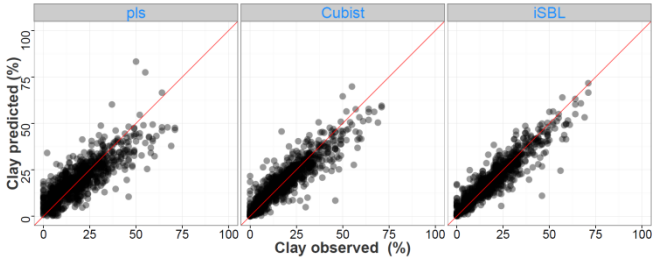
https://github.com/l-ramirez-lopez/resemble_v0.1

Shenk, J.S., Westerhaus, M.O., Berzaghi, P. 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. Journal of Near Infrared Spectroscopy 5, 223-232.

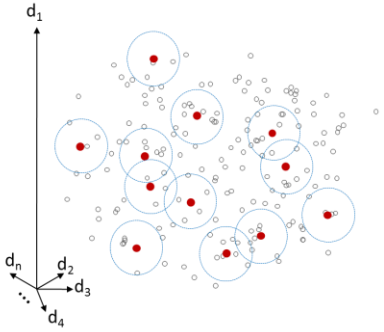
Naes, T., Isaksson, T., Kowalski, B. 1990. Locally weighted regression and scatter correction for nearinfrared reflectance data. Analytical Chemistry 62, 664-673



MBL offers a great opportunity to reduce the complexity problems associated with soil spectral modelling in large scale soil spectral libraries (SSL).



An adequate estimation and use of the soil spectral similarity information may lead to accurate soil vis–NIR predictions carried out by using MBL.



The analysis of the soil similarity (e.g. spectral, geographical, compositional, etc) should NOT be neglected in the management and modelling tasks involved in the use of any large scale SSL.

Try the '**resemble**' package in your SSL!

You combine it with the '**prospectr**' package for spectral preprocessing and calibration sampling

Memory based learning methods and tools: towards efficient modelling, predicting and managing tasks in large scale soil spectral libraries

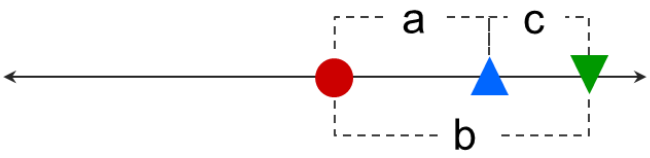
Leonardo Ramirez-Lopez [1,2]*; Antoine Stevens [3]

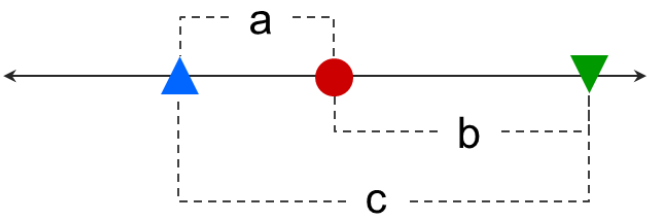
[1] Swiss Federal Institute of Technology, Zurich (ETHz), Switzerland
[2] Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Switzerland
[3] Universite Catholique de Louvain, Belgium

We thank ICRAF and JRC form making the Soil spectral libraries available

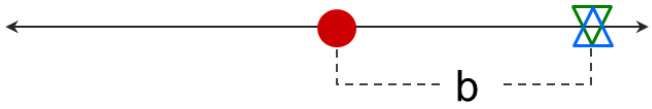
Grazie!

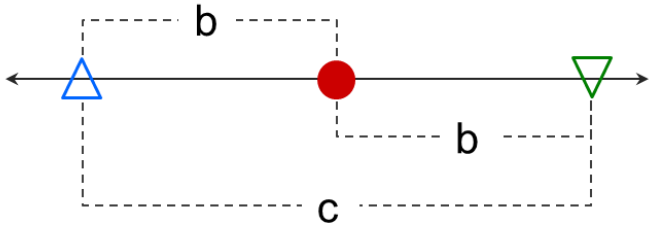
Example 1





Example 2

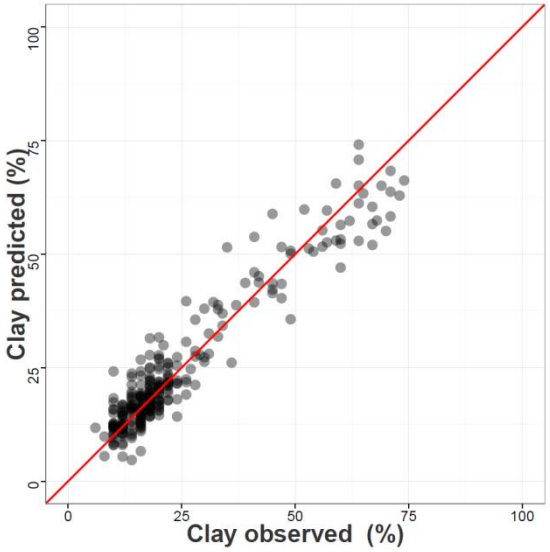




Soil vis–NIR libraries

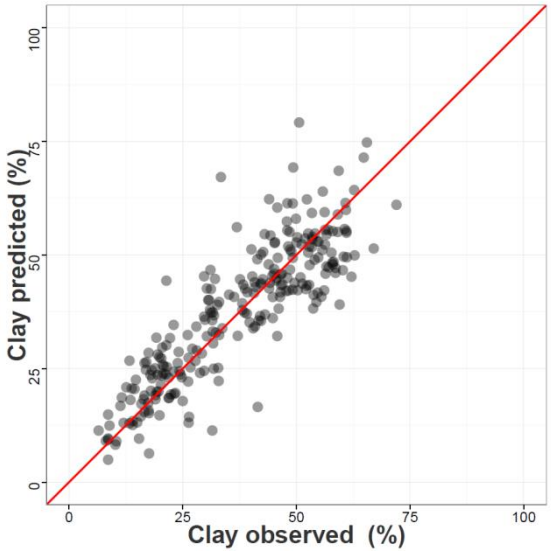
Field scale

R^2 = 0.91; RMSE=5.02%



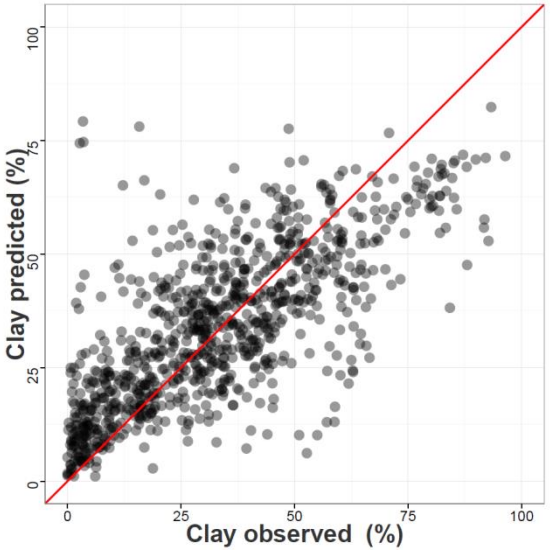
Regional scale

R^2 = 0.75; RMSE=8.03%

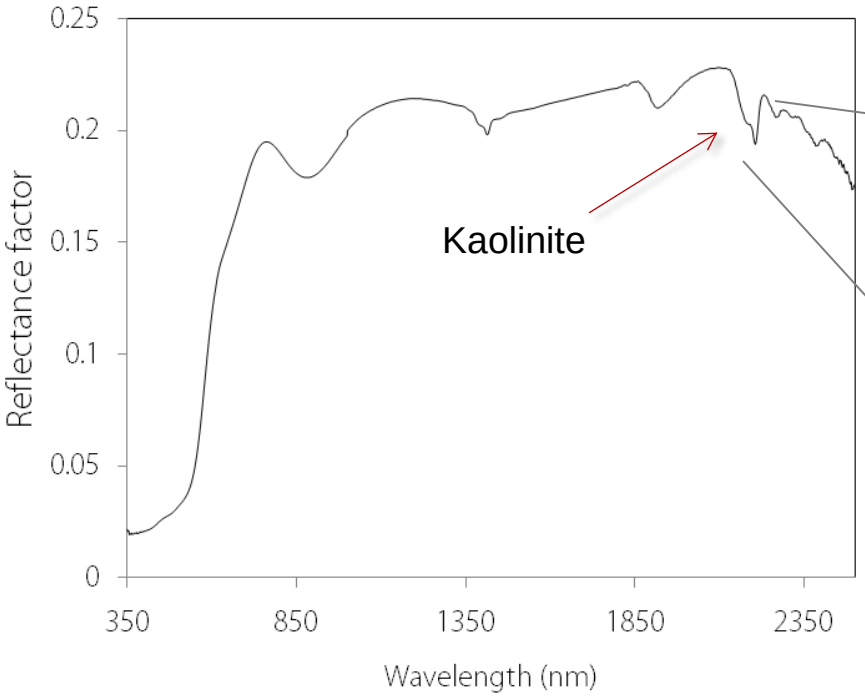


Global scale

R^2 = 0.50; RMSE=15.53%

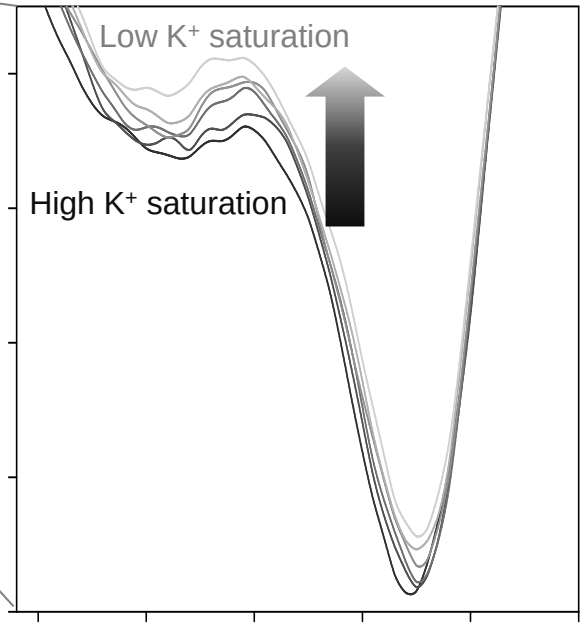


What makes big soil vis–NIR libraries so complex?



Another example…

Soil Kaolinite



The spectral similarity/dissimilarity methods employed in any MBL algorithm should attempt to reflect the compositional similarity/dissimilarity between soil samples.