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

* Leonardo.Ramirez@usys.ethz.ch
Soil vis–NIR libraries

Field scale

\[ R^2 = 0.91; \text{RMSE}=5.02\% \]
Memory-based learning methods and tools: towards efficient modelling, predicting and managing tasks in large scale soil spectral libraries

Soil vis–NIR libraries

**Introduction**

MBL  Case studies  Tools  Final remarks

**Field scale**

$R^2 = 0.91; \text{RMSE}=5.02\%$

**Regional scale**

$R^2 = 0.75; \text{RMSE}=8.03\%$
Memory-based learning methods and tools: towards efficient modelling, predicting and managing tasks in large scale soil spectral libraries

Soil vis–NIR libraries

Field scale

\[ R^2 = 0.91; \quad \text{RMSE}=5.02\% \]

Regional scale

\[ R^2 = 0.75; \quad \text{RMSE}=8.03\% \]

Global scale

\[ R^2 = 0.50; \quad \text{RMSE}=15.53\% \]
Memory-based learning methods and tools: towards efficient modelling, predicting and managing tasks in large scale soil spectral libraries

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

**MBL for soil spectral libraries**

Soil attribute with unknown values

\[(X_u, Y_u) = \{x_{ui}, y_{ui}\}_{i=1}^{m}\]

Soil spectral library

\[(X_r, Y_r) = \{x_{rj}, y_{rj}\}_{j=1}^{n}\]

\(p\)-dimensional spectra
\[(X_u, Y_u) = \{x_{u_i}, y_{u_i}\}_{i=1}^m\]

\[(X_r, Y_r) = \{x_{r_j}, y_{r_j}\}_{j=1}^n\]

\[p\text{-dimensional spectral feature space}\]
\[(X_u, Y_u) = \{x_{ui}, y_{ui}\}_{i=1}^{m}\] 

\[(X_r, Y_r) = \{x_{rj}, y_{rj}\}_{j=1}^{n}\] 

\[S(x_{ui}, x_{rj})\]
Sample neighbors

\[ S(x_{u_i}, x_{r_j}) \]

a. Sphere neighbors  \hspace{1cm} OR \hspace{1cm} b. k-NN
Sample neighbors

Get rid of unnecessary samples

$S(xu_i, xr_j)$
The solution to a complex task is constructed by a collection of simple local functions.

The equation is given by:

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

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

<table>
<thead>
<tr>
<th></th>
<th>Total samples</th>
<th>‘Unknown set’ (# samples)</th>
<th>Reference set (# samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continental (LUCAS)*</td>
<td>19036</td>
<td>1000 (topsoil)</td>
<td>18036</td>
</tr>
<tr>
<td>World (ICRAF)**</td>
<td>3643</td>
<td>935 (168 profiles)</td>
<td>2078</td>
</tr>
</tbody>
</table>

*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 networks (PLS-NN)  
- Cubist  
- LOCAL  
- Modified LOCAL  
- Spectrum-based learner (SBL)  
- Improved SBL (iSBL)  

Memory-based learners

### Case 1

**Clay content prediction results @ CONTINENTAL scale**

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSBL</td>
<td>4.88</td>
<td>0.87</td>
</tr>
<tr>
<td>LOCAL</td>
<td>5.00</td>
<td>0.86</td>
</tr>
<tr>
<td>SBL</td>
<td>5.09</td>
<td>0.86</td>
</tr>
<tr>
<td>mLOCAL</td>
<td>5.33</td>
<td>0.85</td>
</tr>
<tr>
<td>Cubist</td>
<td>5.72</td>
<td>0.82</td>
</tr>
<tr>
<td>pls-NN</td>
<td>7.43</td>
<td>0.70</td>
</tr>
<tr>
<td>svm</td>
<td>7.59</td>
<td>0.69</td>
</tr>
<tr>
<td>pls</td>
<td>7.65</td>
<td>0.68</td>
</tr>
<tr>
<td>Rf</td>
<td>8.61</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Memory-based learners

![Clay content prediction scatter plots](image-url)
Case 2
Clay content prediction results
@ GLOBAL scale

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSBL</td>
<td>12.50</td>
<td>0.67</td>
</tr>
<tr>
<td>SBL</td>
<td>12.53</td>
<td>0.67</td>
</tr>
<tr>
<td>mLOCAL</td>
<td>12.68</td>
<td>0.67</td>
</tr>
<tr>
<td>Cubist</td>
<td>12.80</td>
<td>0.66</td>
</tr>
<tr>
<td>LOCAL</td>
<td>13.19</td>
<td>0.64</td>
</tr>
<tr>
<td>pls-NN</td>
<td>14.60</td>
<td>0.59</td>
</tr>
<tr>
<td>svm</td>
<td>14.69</td>
<td>0.55</td>
</tr>
<tr>
<td>Rf</td>
<td>15.38</td>
<td>0.50</td>
</tr>
<tr>
<td>pls</td>
<td>15.53</td>
<td>0.50</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSBL</td>
<td>12.50</td>
<td>0.67</td>
</tr>
<tr>
<td>SBL</td>
<td>12.53</td>
<td>0.67</td>
</tr>
<tr>
<td>mLOCAL</td>
<td>12.68</td>
<td>0.67</td>
</tr>
<tr>
<td>Cubist</td>
<td>12.80</td>
<td>0.66</td>
</tr>
<tr>
<td>LOCAL</td>
<td>13.19</td>
<td>0.64</td>
</tr>
<tr>
<td>pls-NN</td>
<td>14.60</td>
<td>0.59</td>
</tr>
<tr>
<td>svm</td>
<td>14.69</td>
<td>0.55</td>
</tr>
<tr>
<td>Rf</td>
<td>15.38</td>
<td>0.50</td>
</tr>
<tr>
<td>pls</td>
<td>15.53</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Spectroscopy-oriented software for MBL

<table>
<thead>
<tr>
<th>Name</th>
<th>Platform</th>
<th>Algorithm(s) implemented</th>
</tr>
</thead>
<tbody>
<tr>
<td>WinISI</td>
<td>FOSS software</td>
<td>LOCAL (Shenk et al., 1997)</td>
</tr>
<tr>
<td>PLS_Toolbox</td>
<td>Matlab</td>
<td>Locally weighted PLS (LWR, Naes et al., 1990)</td>
</tr>
<tr>
<td>resemble</td>
<td>R package</td>
<td>&gt; 200 options (including LOCAL, LWR, mLLOCAL, SBL, iSBL, etc)</td>
</tr>
</tbody>
</table>

Spectroscopy-oriented software for MBL

<table>
<thead>
<tr>
<th>Name</th>
<th>Platform</th>
<th>Algorithm(s) implemented</th>
</tr>
</thead>
<tbody>
<tr>
<td>WinISI</td>
<td>FOSS software</td>
<td>LOCAL (Shenk et al., 1997)</td>
</tr>
<tr>
<td>PLS_Toolbox</td>
<td>Matlab</td>
<td>Locally weighted PLS (LWR, Naes et al., 1990)</td>
</tr>
<tr>
<td>resemble</td>
<td>R package</td>
<td>&gt; 200 options (including LOCAL, LWR, mLOCAL, SBL, iSBL, etc)</td>
</tr>
</tbody>
</table>

Methods for computing the spectral similarity/dissimilarity
- 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

Usage of the similarity/dissimilarity information
- None
- Predictors
- Weights

Local fit (regression algorithm)
- 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

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

Example 1

Example 2
Soil vis–NIR libraries

Field scale

$R^2 = 0.91; \text{RMSE}=5.02\%$

Regional scale

$R^2 = 0.75; \text{RMSE}=8.03\%$

Global scale

$R^2 = 0.50; \text{RMSE}=15.53\%$
What makes big soil vis–NIR libraries so complex?

Another example...

Soil Kaolinite

Low K⁺ saturation

High K⁺ saturation

Memory-based learning methods and tools: towards efficient modelling, predicting and managing tasks in large scale soil spectral libraries
The spectral similarity/dissimilarity methods employed in any MBL algorithm should attempt to reflect the compositional similarity/dissimilarity between soil samples.