

# Strategy of efficient estimation of soil organic content at the local scale based on the national spectral database

Hongyi Li<sup>1</sup>, Yuheng Li<sup>1</sup>, Meihua Yang<sup>2</sup>, Songchao Chen<sup>3</sup>, and Shi Zhou<sup>4</sup>

<sup>1</sup>Jiangxi University of Finance and Economics

<sup>2</sup>Yuzhang Normal University

<sup>3</sup>INRA

<sup>4</sup>Zhejiang University

September 28, 2021

## Abstract

The aim of this paper was to compare the prediction performance of three strategies: general global Partial least squares regression (PLSR) using CSSL with and without spiking samples, memory-based learning (MBL) using CSSL with and without spiking samples and general PLSR using only spiking samples to predict soil organic matter in the target area. When using spiked subsets, we also investigated the prediction performance of the extra-weighted subsets. A series of spiking subsets randomly selected from the total spiking samples were selected by conditioned Latin hypercube sampling (cLHS) from the target sites. We calculated the mean squared Euclidean distance (msd) of different spiking subsets with the distribution density function of their vis-NIR spectra only and statistically inferred the optimal sampling set size to be 30. Our study showed that when the number of spiking were lower than 30, the predicted accuracy derived from global PLSR using CSSL spiked with and without extra-weighted samples was greater than the predicted accuracy derived from the general PLSR using the corresponding number of spiking samples only (RMSE 5.57–5.98 v.s. RMSE 6.76). Global PLSR using CSSL spiked with the statistically optimal local samples can achieve higher predicted performance (with a mean RMSE of 5.75). MBL spiked with five extra-weighted optimal spiking samples achieved the best accuracy with an RMSE of 3.98, an R<sup>2</sup> of 0.70, a bias of 0.04 and an LCCC of 0.81. The msd is a simple and effective method to determine an adequate spiking size using only vis-NIR data.

## Hosted file

Document1.docx available at <https://authorea.com/users/308848/articles/539538-strategy-of-efficient-estimation-of-soil-organic-content-at-the-local-scale-based-on-the-national-spectral-database>

## Hosted file

Table.docx available at <https://authorea.com/users/308848/articles/539538-strategy-of-efficient-estimation-of-soil-organic-content-at-the-local-scale-based-on-the-national-spectral-database>









