Course M-GFP3

Imaging and non-imaging spectroscopy: Techniques and analysis

Exam – Term Paper

Topic:

Strategies to enhance predictive modeling of soil organic carbon (SOC) using the LUCAS topsoil spectral library

General information:

- Working in groups of two or three.
- Each group member's contributions must be clearly identified, and all members are expected to contribute roughly the same amount of work.
- Ensure that all results are reproducible by providing a script (e.g., in R, Python, etc.).

Submit the following as your final deliverables:

1. A written report in PDF format, including:
    - Cover sheet with course information, group member names & dataset
    - The actual report, incl. results and brief discussion
    - List of references used, if applicable
2. The accompanying script file(s) (e.g., R, Python, etc.)
3. Alternatively, you may submit a single Markdown document that integrates both the text and code.

Please ensure that your submission is clear, well-structured, and follows academic standards.

Instructions

The European Soil Data Centre (ESDAC) has constructed a comprehensive database of topsoil characteristics at the European scale. In the 2009 sampling campaign, approximately 20,000 topsoil samples were collected and subjected to extensive analysis. The database encompasses a multitude of physicochemical reference values and the associated VNIR spectra. For further details, please refer to the following link: https://esdac.jrc.ec.europa.eu/projects/lucas.

Please select a dataset (available in Moodle) and proceed to complete the following tasks:

Data splitting (5 P):

- Split your data into a calibration data set (~70%) and an independent test data set (~30%). Show that both are representative of the full data set. For procedures with randomized approaches, please define and note the seed (in R: `set.seed( )`) to make

the split reproducible for the instructors. From this point onward, the composition of the test data set must remain constant and unchanged for all subsequent tasks.

Baseline model (5 P):

- Develop a global **baseline PLSR model** using the calibration dataset (**entire VNIR range** from 500 nm to 2499 nm in steps of 2 nm) without applying any spectral preprocessing. The target variable is **soil organic carbon (SOC)**. Perform internal optimization to determine the optimal number of latent PLS variables and report your selected value. Apply the optimized model to the independent test set. Compute the validation metrics (R^2, RMSE, bias, and RPD), visualize the results in a scatter plot (observed vs. predicted values) and assess the model's performance.

Model Improvement Strategies (5 P per strategy):

- Develop and evaluate three distinct strategies to improve the baseline model, using the same independent test set for validation. For each strategy, report the validation metrics (R^2, RMSE, bias, and RPD), visualize the best result in a scatter plot (observed vs. predicted values) and assess the performance of these alternative models. Use the same independent test set for all strategies to ensure that validation metrics are directly comparable.

  IMPORTANT: Testing two or more spectral preprocessing methods is considered one strategy, not multiple strategies. Similarly, testing one or more alternative regression algorithms counts as one strategy, not multiple.

Discussion of Results (5 P):

- Briefly discuss your results and interpret them based on the validation metrics for the test set. Compare your findings with those of published studies in a similar context. Evaluate whether soil VNIR reflectance spectroscopy could serve as a complementary approach for large-scale soil organic carbon assessment in Earth (system) science.

Deadline: 06/03/2025 – 11:59 p.m.

Wishing you much success!