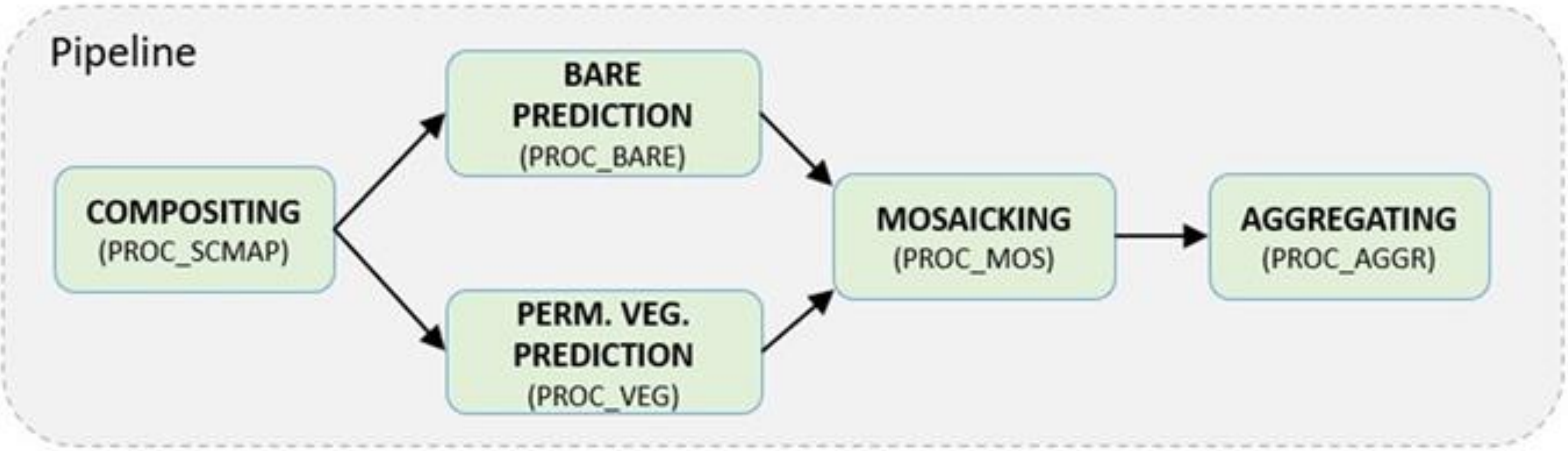
Session 3 – SOC prediction algorithms for Non-Vegetated areas

Nikolaos Tsakiridis, AUTh
WORLDSOILS consortium

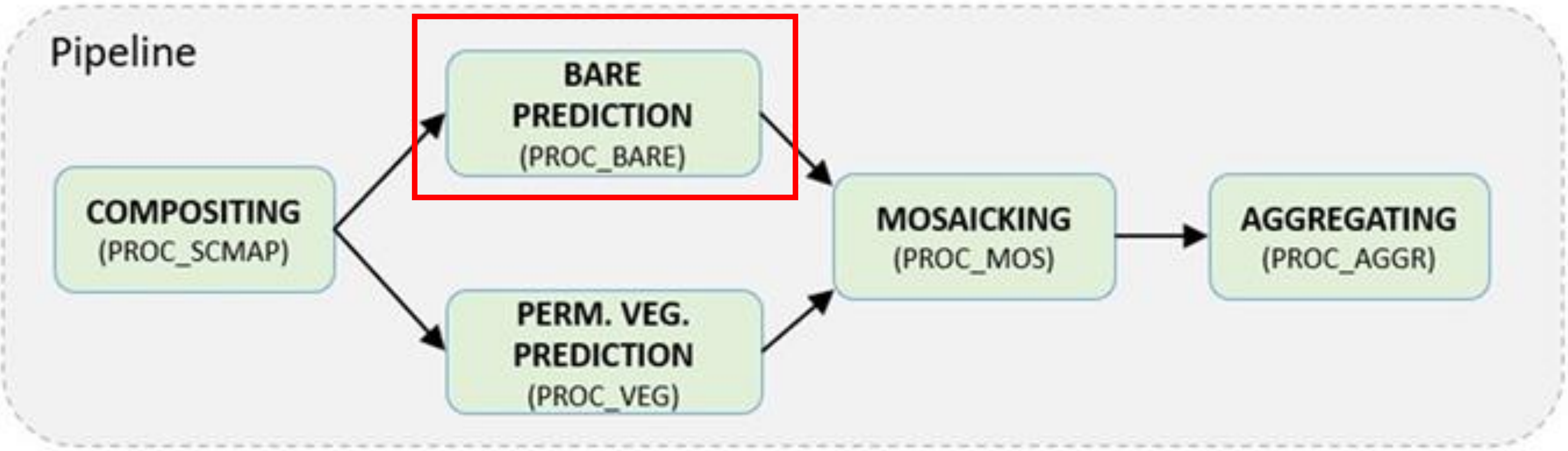
ESA Symposium on Earth Observation for Soil Protection and Restoration

Contract 400131273/20/I-NB

0. WorldSoils – general framework

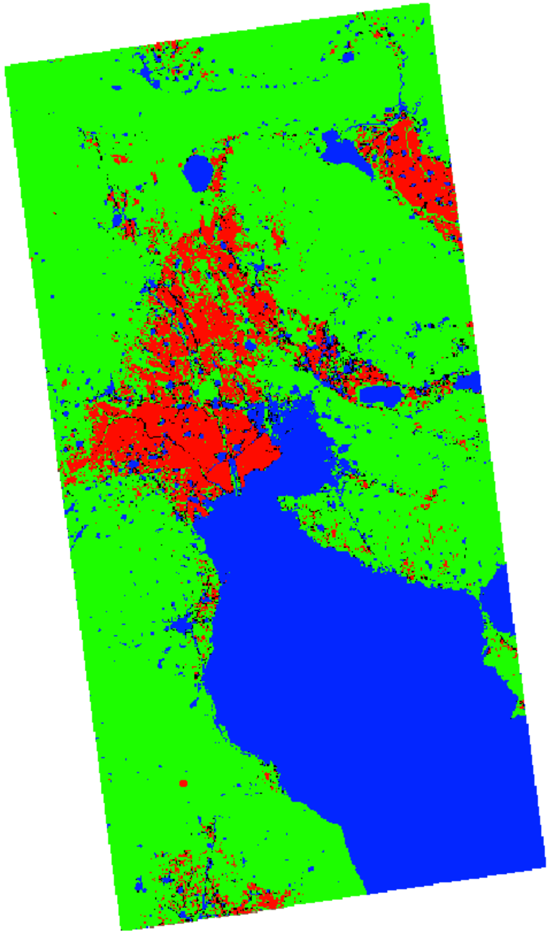


0. WorldSoils – general framework

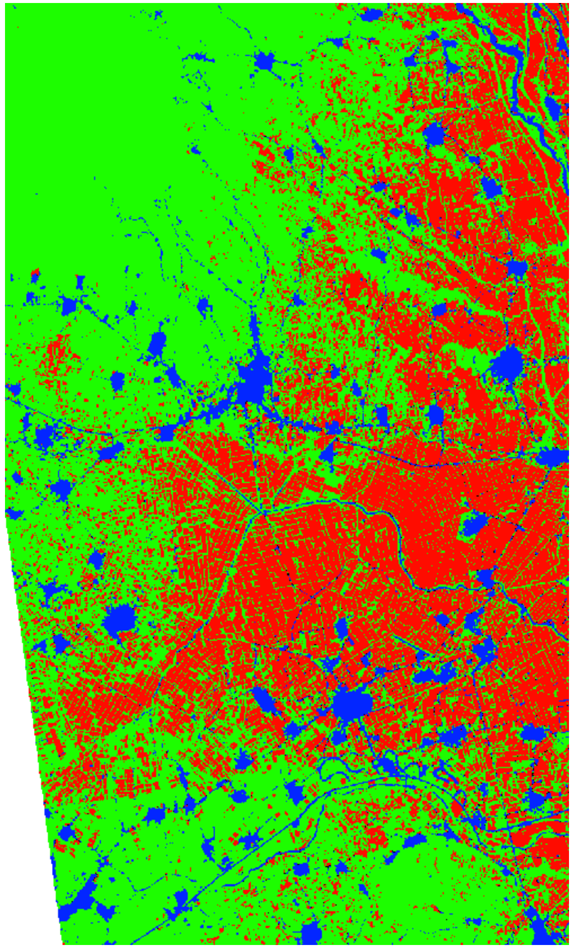


1. SCMaP processing

Example of region in Central Macedonia

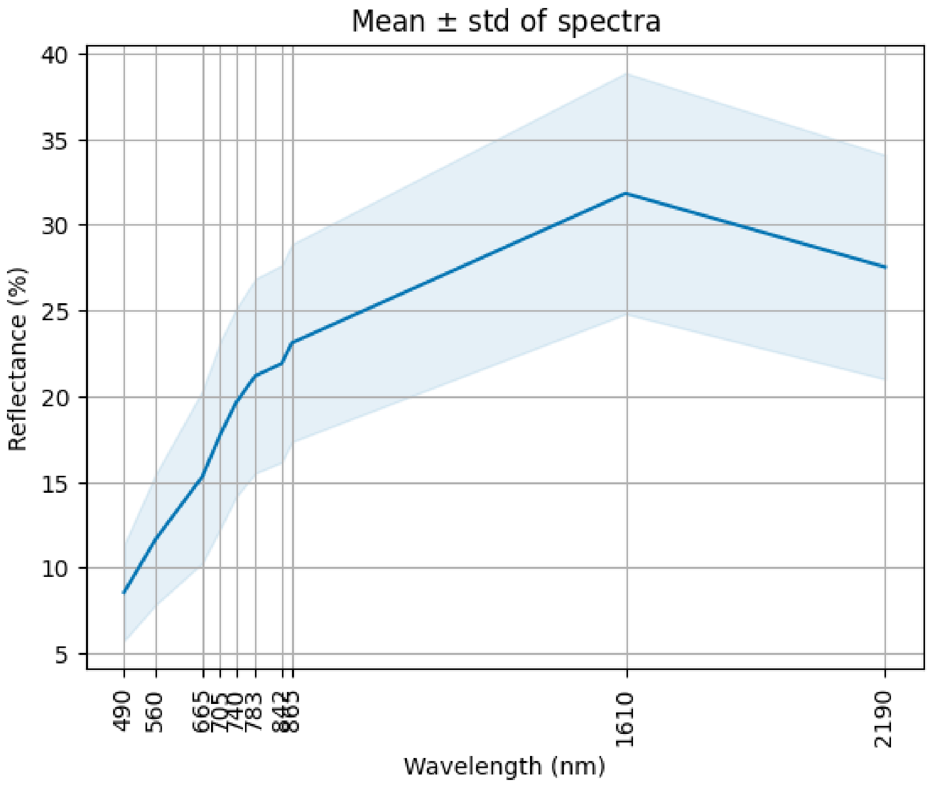


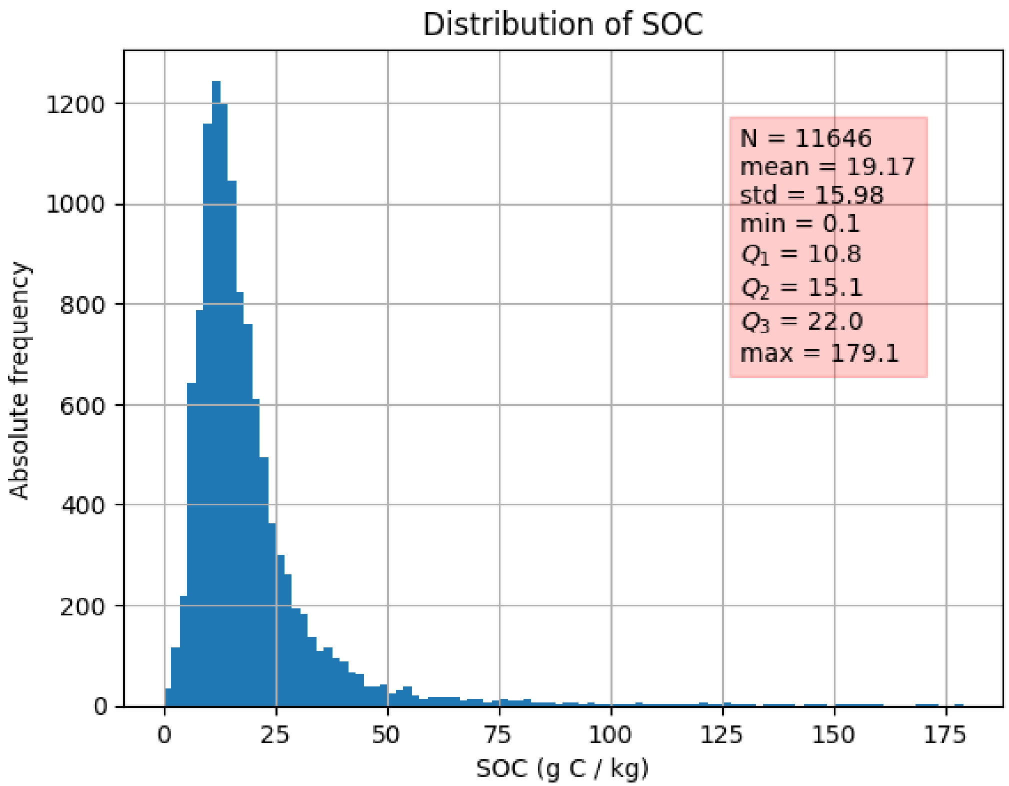
Example of region in Central Macedonia (zoomed in)



1. Data considered

LUCAS topsoil data for non-vegetated areas (continental level)

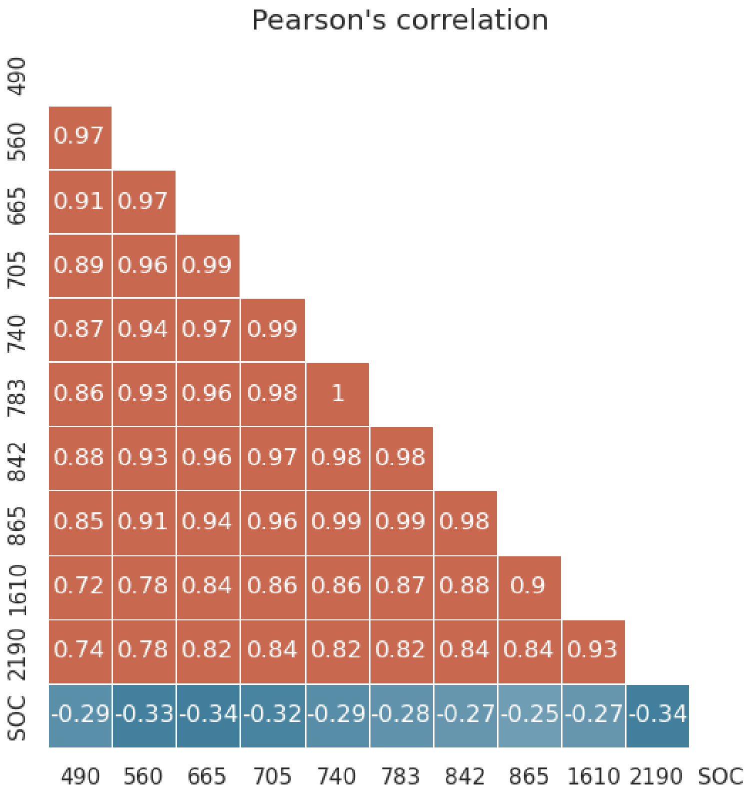


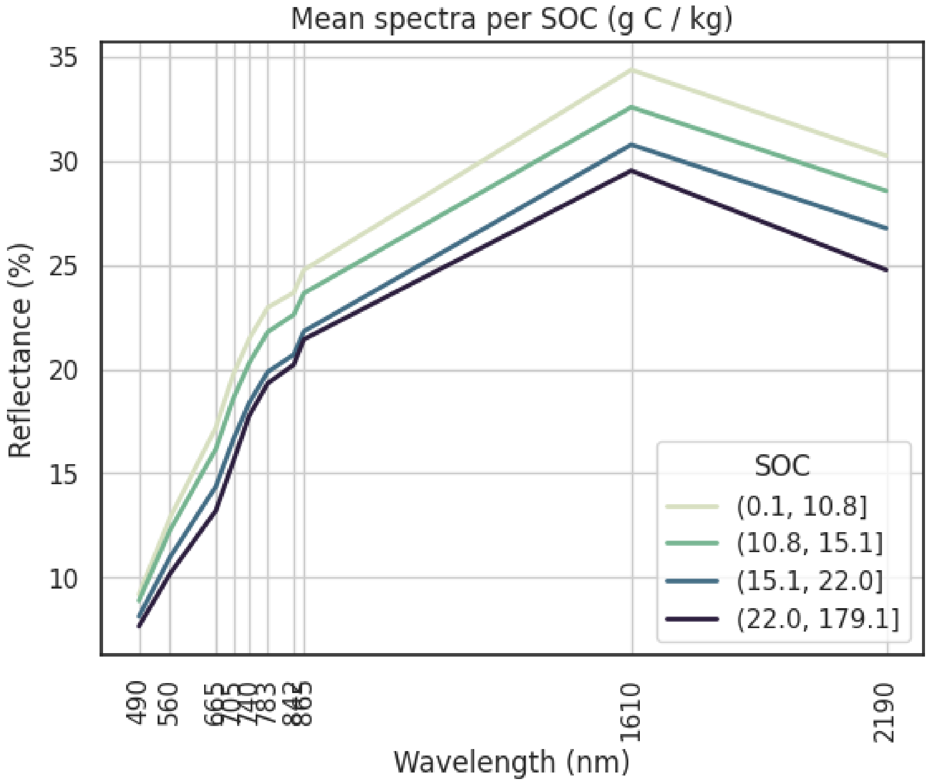


WORLDSOILS

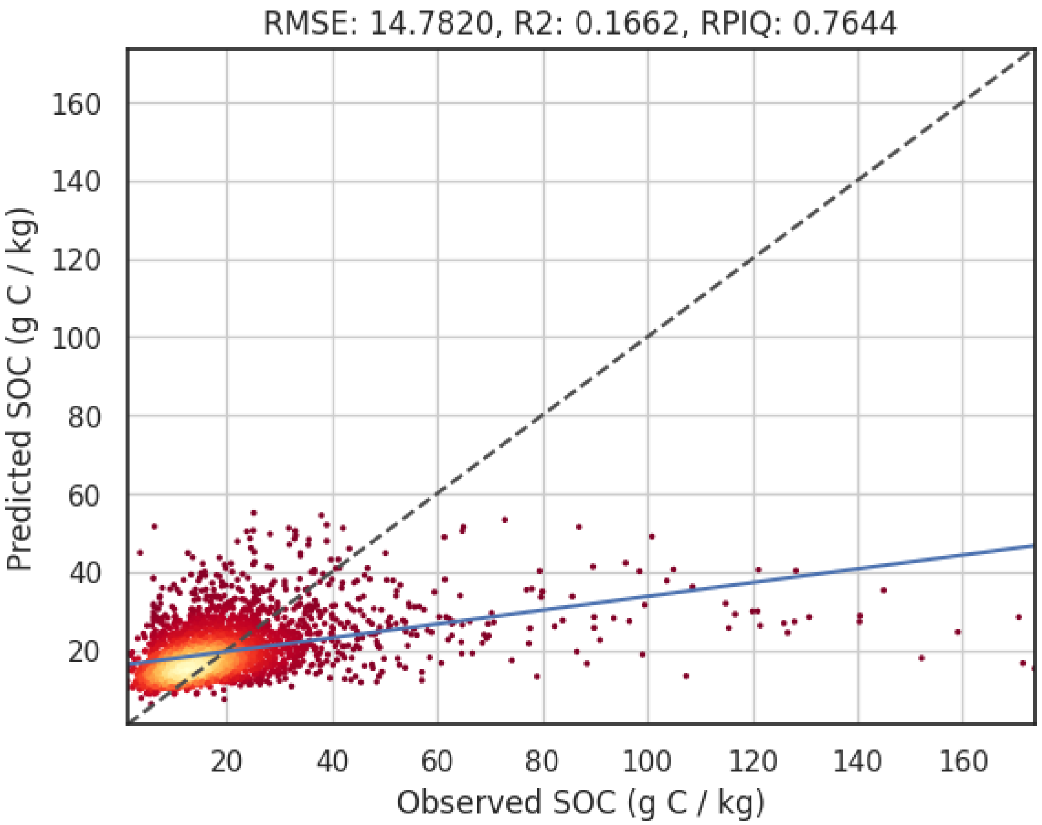
1. Data considered

LUCAS topsoil data for non-vegetated areas (continental level)





2. First modeling approach

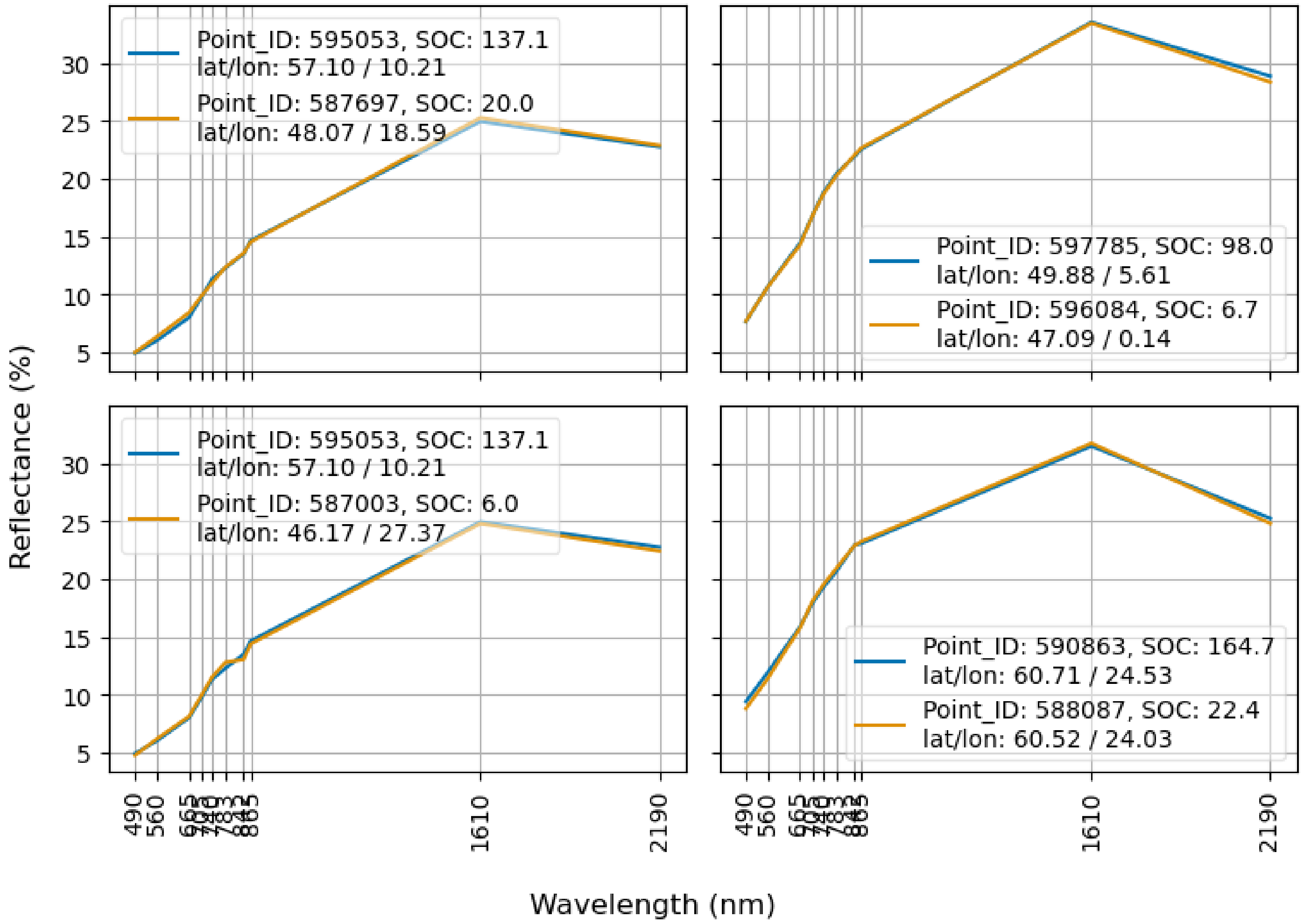


- 70 / 30 split into calibration and independent test set
- Weighted Random Forest for imbalanced regression
- Weights determined by arbitrarily splitting the data into three classes
- Parameters optimized via 5-fold CV
- Train / test split via using k-means to create 300 spectral clusters, and random 75% / 25% split within each cluster
- Poor performance but also: **no high values!**

2. First modeling approach – further considerations

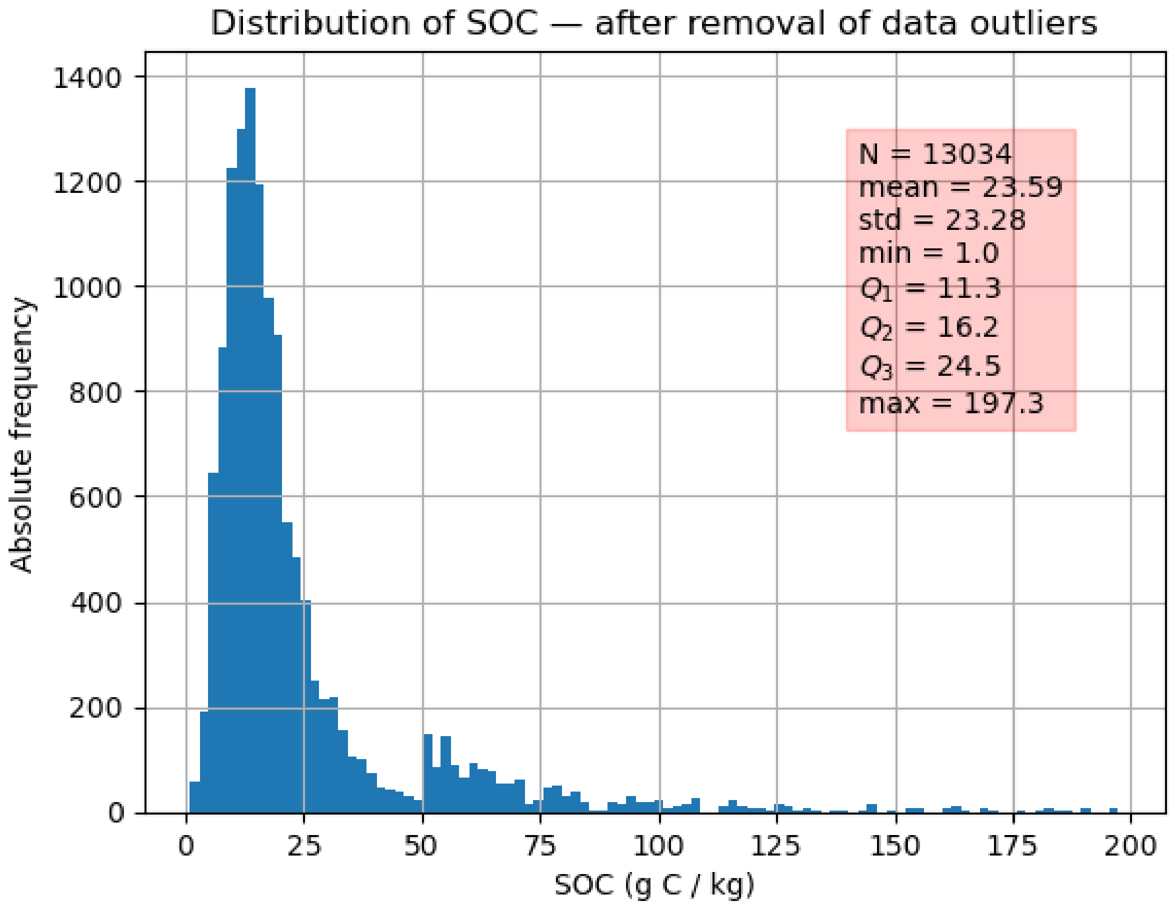
- Similar spectra ⇨ similar SOC
- Dissimilar spectra ⇨ dissimilar SOC

Not always the case!



3. Data augmentation for high SOC content

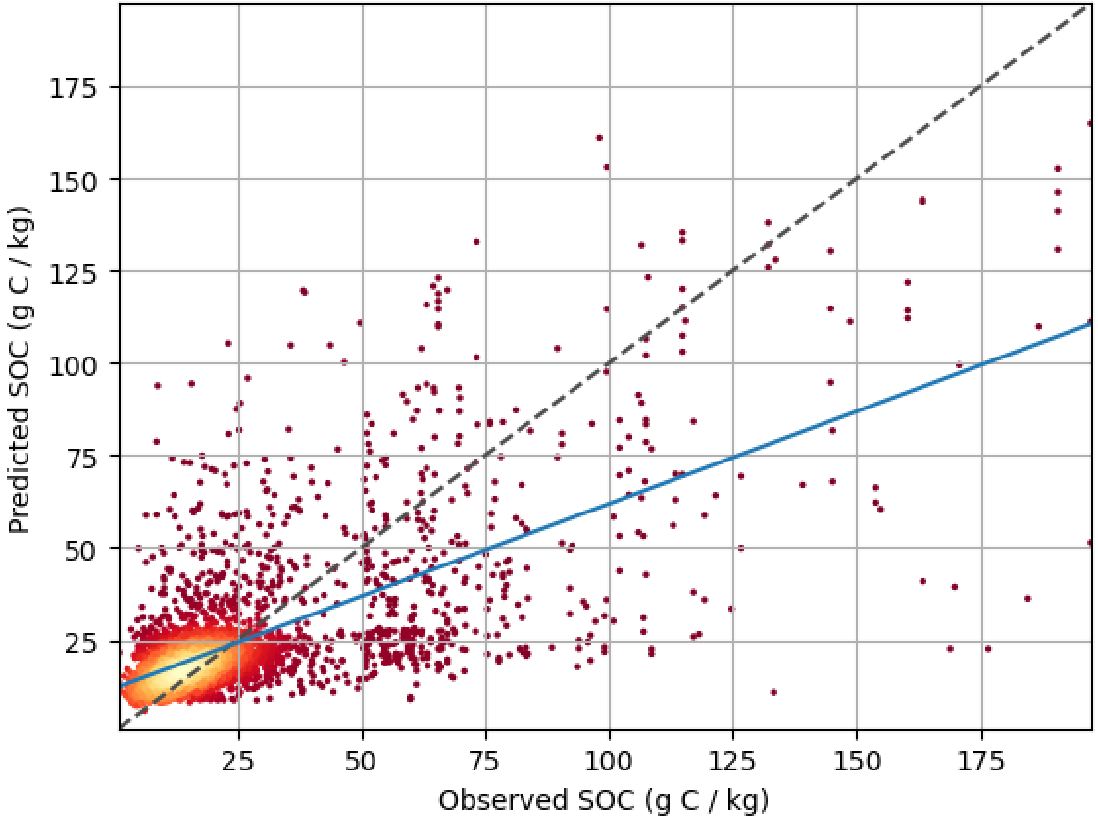
LUCAS topsoil data for non-vegetated areas – augmented



- Augmenting high SOC content values (i.e., over 50 g / kg)
- Geostatistics: close points ≈ similar SOC
- 3x3 grid around the central LUCAS sample
- Distribution improved, but not dramatically

3. Data augmentation for high SOC content

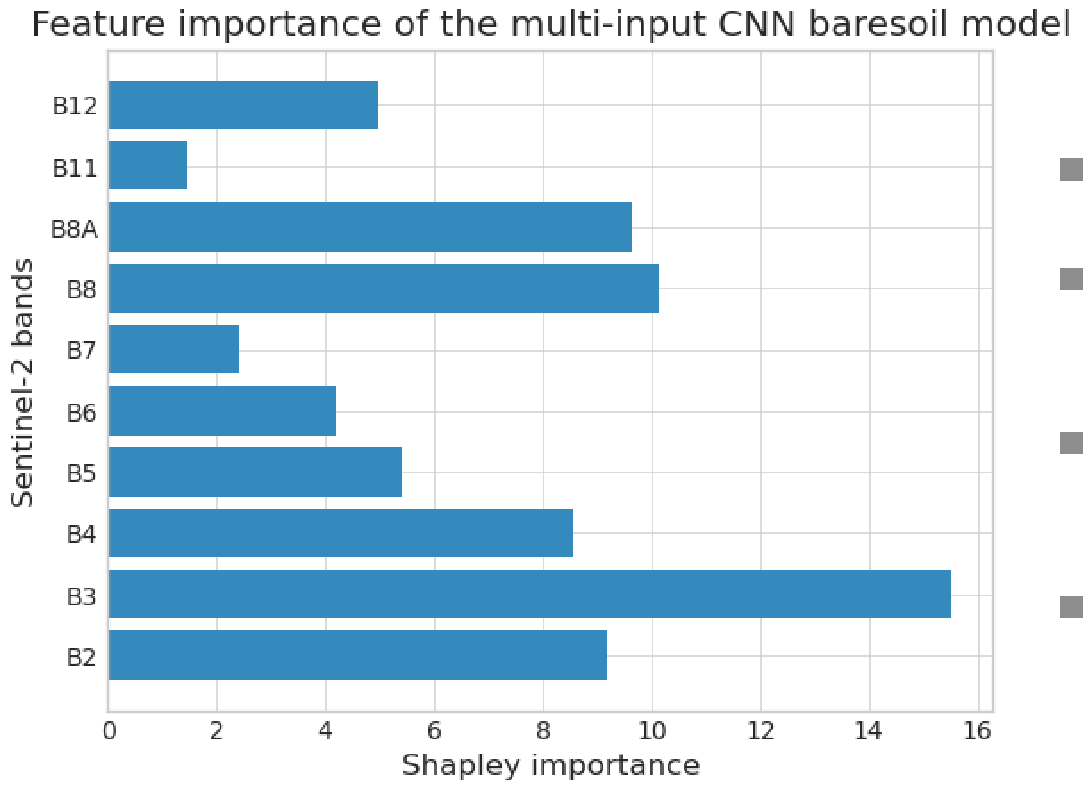
LUCAS topsoil data for non-vegetated areas – augmented



- Custom CNN developed
  - Multi-input (Ref. and Ref. SNV)
  - Hyperparameters optimized using the HyperBand algorithm
  - 3 convolutional layers and 5 fully connected layers
  - Custom loss function to penalize large samples
- Model uncertainty using bootstraps (PICP: 0.93)
- Improved accuracy

R^2	RMSE (g / kg)	NRMSE	CCC	Bias	RPIQ
0.41	18.07	0.77	0.46	0.08	0.73

LUCAS topsoil data for non-vegetated areas – augmented



- CNN = blackbox
- Feature importance calculated using model-agnostic Shapley values
- Largest importance ascribed to B3 and B8 in the visible and near infrared, respectively
- Less so in the SWIR, with B12 at 2.19 µm

Open-source software used

- Neural network: keras and tensorflow
- Raster loading and calculations: rasterio and gdal
- Connectivity: boto3
- Visualizations: pandas, matplotlib and seaborn

- All models developed in self-contained docker containers, can be run on different infrastructures and can be deployed in different architectures (e.g., arm64)

- Inference process optimized using integer quantization (int16) and parallelized to take advantage of multi-core CPUs

5. Final considerations – summary

The road forward

- Multi-spectral data have limited capacity to predict very accurately the topsoil SOC content, particularly given its high skewness; simulated Sentinel-2 from laboratory LUCAS data also yield similar results
- For a pure spectral-based model, it is necessary to utilize the forthcoming hyperspectral missions (e.g., CHIME, HyperField) whose continuous monitoring will provide a plethora of more data
- Food for thought: the bare soil model requires the presence of bare soil; good agricultural practices mandate the presence of cover crops in the winter months. Will we ever be able to see all (or rather, a good percentage of) bare fields in the future using only spring and autumn months with a revisit time of 12.5 days (CHIME)?

Thank you!

Nikolaos Tsakiridis, tsakirin@auth.gr

ESA Symposium on Earth Observation for Soil Protection and Restoration

Contract 400131273/20/I-NB