Technische Hochschule Ingolstadt

AgroLens

A Soil Quality Evaluation System

01.02.2025

AKI-M WS24/25

Project specification:

- From MI4 People
- Vision: *Creation of a free of charge **MI system** that can predict most important **quality indicators for soil without** performing **expensive chemical lab tests** and can be directly used by farmers of any educational level. The system should be able make these predictions based on various data **inputs**, like **satellite imagery**, **infra-red spectral measurements data**, **etc.**, and will be able to improve its predictions whenever it is able to cross-validate its predictions with chemical test data.*

Scope:

- Study project by Team AgroLens
- 7 team members
- Start at end of October

MI4People

2. Agenda

1. Introduction
2. Agenda
3. Methodology
4. Model for Europe
5. Extended Model for Europe
6. Model for World
7. Summary & Outlook
8. Discussion

3. Methodology

Overview: Input Data

Input Data

BASE	12 Sentinel-2 satellite image bands (single pixel)
SURR	12x 3x3 Sentinel-2 pixels (3x3 pixels for 12 bands)
CLAY	1024 Clay AI Embeddings
WTHR	9 Weather Features
CRY	27 Crop Yield Scores

Notes

BASE & SURR	Near the date of the corresponding target soil samples
CLAY	Clay AI embedding from 12x 9x9 Sentinel-2 pixels
WTHR	Time of BASE
CRY	Dated 2010
Model for Europe:	BASE
Extended Model for Europe:	BASE + Options

Locations correspond to the location of the target soil sample.

Target Data

- **Key soil properties**
  - **pH for H2O and CaCl2**
  - **Phosphorus (P)**
  - **Potassium (K)**
  - **Nitrogen (N)**

Data set	Timeframe	Timestamp	Profiles	Landsat 7/8 (> 03/2008)	Sentinel-2 (> 06/2015)
LUCAS	2018	Yes	18,984	18,471	18,471
AfSIS	2009-2018	No	20,704	0	0
WoSIS	1920-2023	Partially	228,000	3,641	0

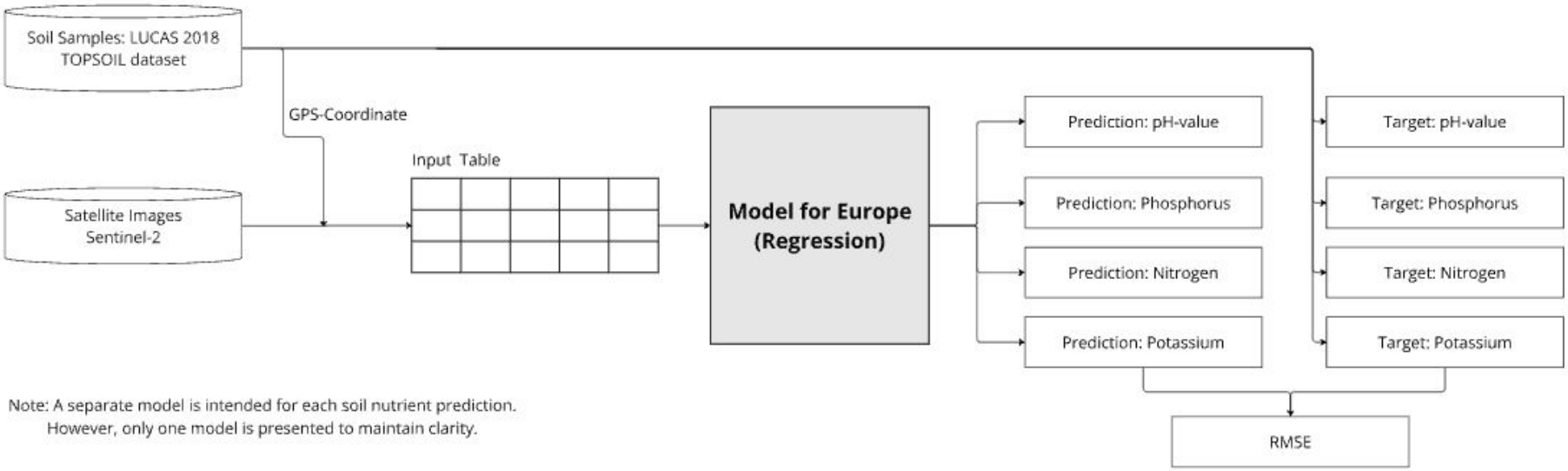
Prediction of soil nutrient levels → **Regression problem**
Five nutrient predictions → **Five separate regression models**

Selected regression models:

1. **XGBoost (Extreme Gradient Boosting)**
   - Sequential tree boosting for error correction of the previous trees
   - Final prediction is obtained by combining all the trees.
   - High accuracy, robust to large/noisy datasets
2. **Fully Connected Neuronal Network**
   - Using Neurons with arranged in layers
   - Captures non-linear interactions between input features
3. **Random Forest**
   - Ensemble model independent tree predictors
   - Robust, interpretable, suitable for limited data

4. Model for Europe

Concept



Note: A separate model is intended for each soil nutrient prediction.
However, only one model is presented to maintain clarity.

4. Model for Europe

Target Data: LUCAS 2018 TOPSOIL dataset

LUCAS Programme Overview (2006)

- **Purpose:** Land use and cover across the EU, every 3 years by Eurostat
- **Coverage:** 2x2 km grid intersections (~1 million georeferenced points)

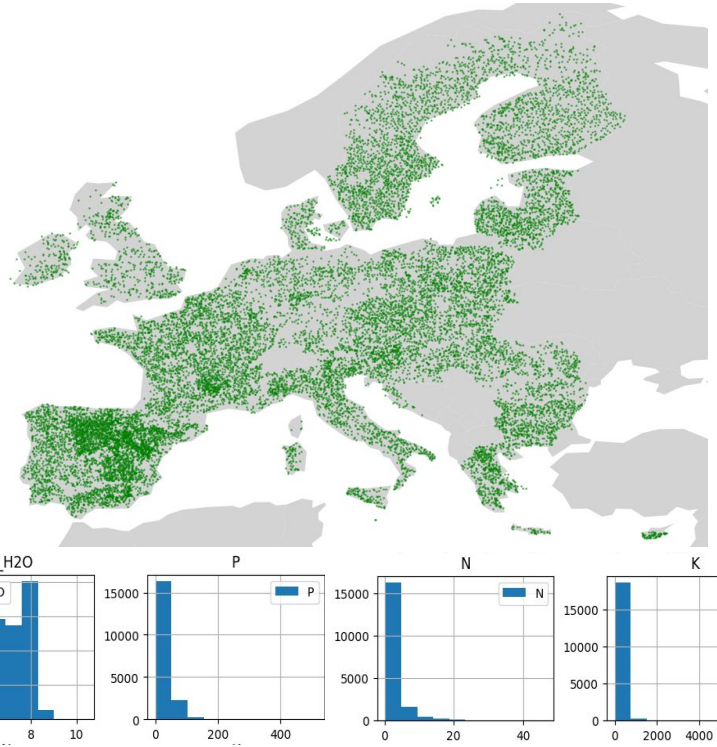
LUCAS 2018 TOPSOIL Dataset

- **Scope:** Soil property data from 18,984 samples across the EU and UK
- **Properties: pH** ($CaCl_2$, H_2O), **nitrogen**, **phosphorus**, **potassium**

Data Preprocessing

- Limits of Detection (LOD):
  - pH: 2–10
  - Nitrogen: 0.2 g/kg
  - Phosphorus & Potassium: 10 mg/kg
- Potassium samples < LOD → 4,945 → Avg. imputing Potassium = 5

Final dataset → 18,471 samples per element



4. Model for Europe

Input Data: Sentinel-2 satellite images

Sentinel-2 Mission Overview

- Launched in 2015/2017 by ESA
- Processing of the whole world in five days
- 100x100 km tiles
- 13 bands (Visible, NIR, SWIR)
- Resolutions: 4x10m, 6x20m, 3x60m
- Level 2A with atmospheric correction (no Band 10)
- Available on Copernicus Data Space

Preprocessing in two steps

- Downloading of raw images and cropping to 101x101 pixels around a GPS coordinate
- Generation of a data table with band values

Enhanced TCI

Band 8

Evaluation Metric

- Root Mean Squared Error (RMSE)

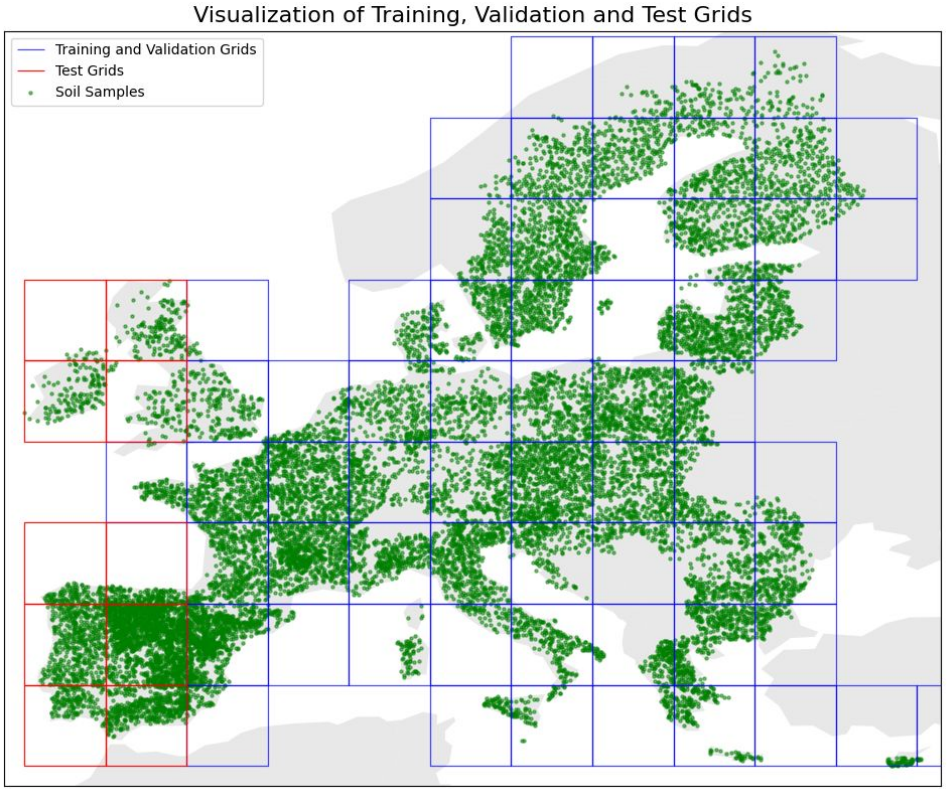
Single Split

Ratio 80:20, Random Split into:

- Training Dataset: 14776 samples
- Test Dataset: 3695 samples

Spatial Cross Validation

- Split into training, validation and test dataset
- 5-Fold Cross Validation → Grid–Based 4° x 4°



4. Model for Europe

Result Overview

Nutrient	Unit	Mean ± StdDev	RMSE (Test Dataset, Single Split)		
			XGBoost	Random Forest	Neural Network
pH in $CaCl_2$	-	5.71 ± 1.40	**1.09**	**1.09**	1.12
pH in H_2O	-	6.26 ± 1.32	1.03	**1.02**	1.08
Phosphorus	mg/kg	26.95 ± 27.02	26.53	26.50	**25.50**
Nitrogen	g/kg	3.15 ± 3.70	3.63	3.63	**3.44**
Potassium	mg/kg	204.83 ± 208.25	216.48	216.06	**178.20**

4. Model for Europe

XGBoost: Spatial Cross Validation

Nutrient	Unit	Mean ± StdDev	XGBoost RMSE (Test Dataset, Single Split)	XGBoost Spatial Cross Validation	
				Average RMSE (5 Folds) Grid = 4x4	RMSE (Test Dataset) Grid = 4x4
pH in $CaCl_2$	-	5.71 ± 1.40	1.09	1.09	1.15
pH in H_2O	-	6.26 ± 1.32	**1.03**	**1.03**	**1.10**
Phosphorus	mg/kg	26.95 ± 27.02	26.53	26.98	26.32
Nitrogen	g/kg	3.15 ± 3.70	3.63	**3.72**	**2.46**
Potassium	mg/kg	204.83 ± 208.25	216.48	**207.68**	**177.52**

4. Model for Europe

Neural Network: Optimization

Training

- Hyperparameter Tuning by using Optuna
- Start hidden layer max. 5

Model Optimization

- Deeper network necessary to avoid underfitting

pH in $CaCl_2$ vs. pH in H_2O:

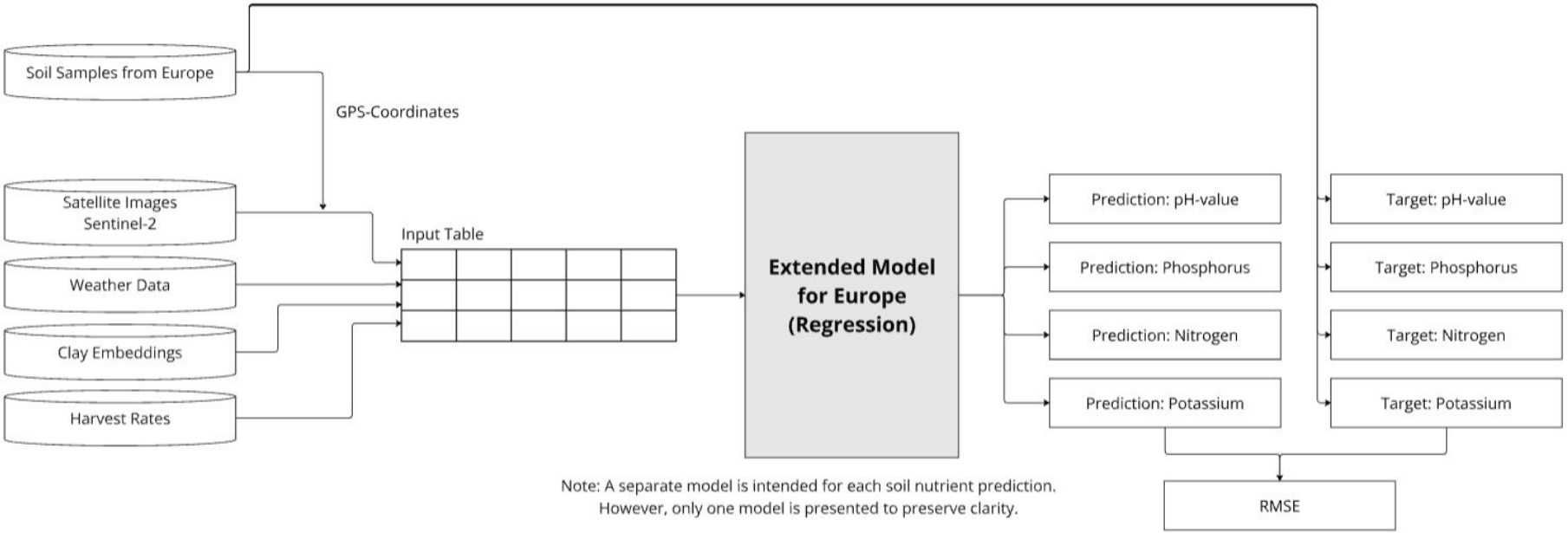
- Comparable accuracy for predicting pH, but the model is simpler for H_2O

Model I - 2 hidden layer

Model II - 8 hidden layer

5. Extended Model for Europe

Concept



5. Extended Model for Europe

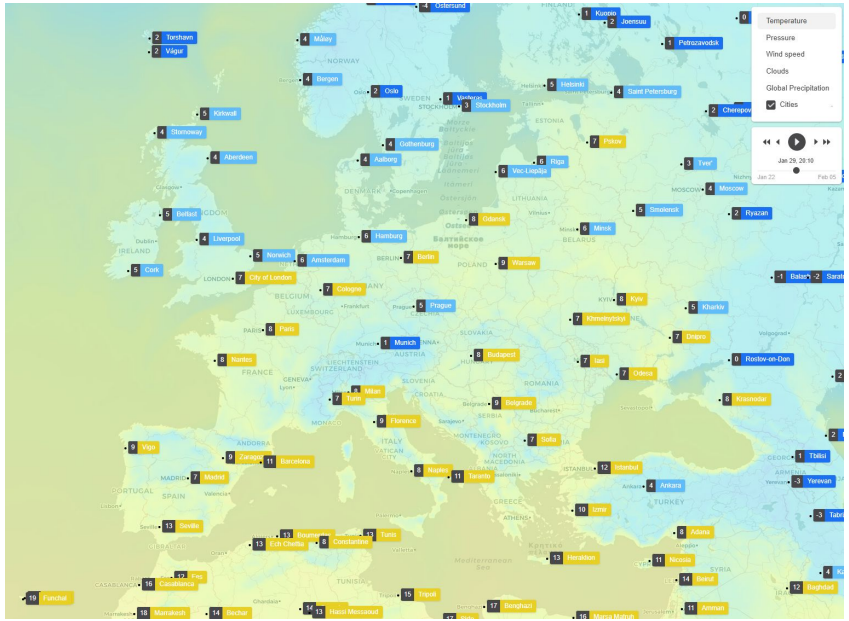
Weather Data

Historic weather data has been purchased from https://www.openweathermap.org.

It comprises the following used features:

- Cloud coverage
- Dew point
- Humidity
- Pressure
- Temperature
- Feels like temperature
- Wind direction
- Wind speed
- Daylight duration

Feature importances may not align with intuition. Therefore, the decision for omitting training features should be evidence based.



5. Extended Model for Europe

Agricultural Yield and Production

Features have been retrieved from the Food and Agriculture Organization of the United Nations Global Agro-Ecological Zones Data Portal:
https://gaez-data-portal-hqfao.hub.arcgis.com/

They comprise “Theme 5: Actual Yields and Production” for various types of important crops, which have been harvested in the year 2010.

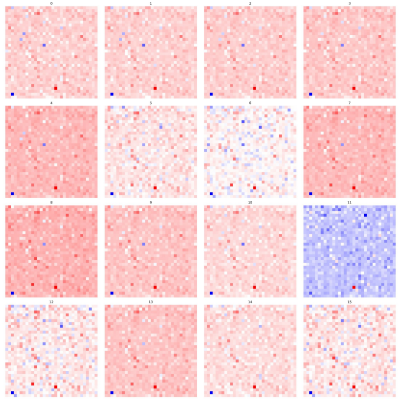
5. Extended Model for Europe

Clay's Model - Usage of Embeddings as Additional Features

Clay's Model is an open-source foundation model for Earth observation.

Vision:

- Makes data cheaper
- Easy-to-use
- More accessible to everyone working on climate and nature



Input Data:

- pixels of sentinel-2 bands: (n,12,9,9)
- time: (n,4)
- latlon: (n,4)
- waves: wavelengths of 12 bands
- gsd: 60

Clay's Model:

- ClayMAEModul
- checkpoint:"clay-v1.5.ckpt"
- model size: "large"
- kernel size: 8x8

Output:

- embeddings (n, 1024)
- 1024 additional features per data set

- n: number of data sets
- image:visualization of the embeddings of the first 16 data sets using "bwr" color map

5. Extended Model for Europe

XGBoost Performance

Nutrient	Unit	Mean ± StdDev	RMSE Model Results		
			BASE	Previous + SURR, WTHR, CRY	Previous + CLAY
pH in $CaCl_2$	-	5.71 ± 1.40	1.09	**0.86**	0.90
pH in H_2O	-	6.26 ± 1.32	1.03	**0.81**	0.85
Phosphorus	mg/kg	26.95 ± 27.02	26.53	**24.88**	25.05
Nitrogen	g/kg	3.15 ± 3.70	3.63	**3.40**	3.44
Potassium	mg/kg	204.83 ± 208.25	216.48	**200.42**	202.03

5. Extended Model for Europe

Fully Connected Neural Network Performance

Nutrient	Unit	Mean ± StdDev	RMSE Model Results		
			BASE	Previous + SURR, WTHR, CRY	Previous + CLAY
pH in $CaCl_2$	-	5.71 ± 1.40	1.12	0.93	**0.91**
pH in H_2O	-	6.26 ± 1.32	1.12	**0.87**	0.90
Phosphorus	mg/kg	26.95 ± 27.02	27.12	24.37	**23.06**
Nitrogen	g/kg	3.15 ± 3.70	3.44	**3.27**	3.35
Potassium	mg/kg	204.83 ± 208.25	185.31	**159.03**	166.36

5. Extended Model for Europe

Random Forest Performance

Nutrient	Unit	Mean ± StdDev	RMSE Model Results		
			BASE	Previous + SURR, WTHR, CRY	Previous + CLAY
pH in $CaCl_2$	-	5.71 ± 1.40	1.09	**0.85**	0.89
pH in H_2O	-	6.26 ± 1.32	1.02	**0.80**	0.84
Phosphorus	mg/kg	26.95 ± 27.02	26.50	**24.60**	24.80
Nitrogen	g/kg	3.15 ± 3.70	3.63	**3.37**	3.42
Potassium	mg/kg	204.83 ± 208.25	216.06	**192.01**	197.36

6. Model for World

Input Data: Landsat 7&8 satellite images (1/2)

Landsat 7 & 8

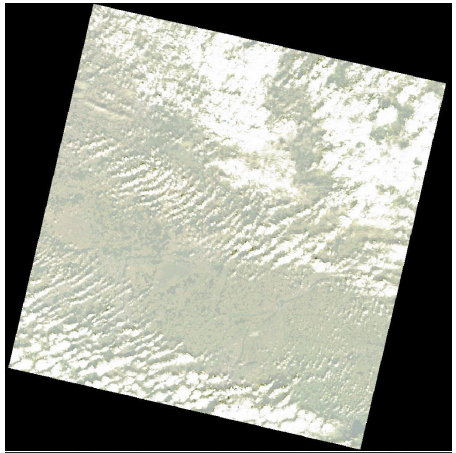
- NASA & USGS
- Processing of the whole world in 16 days
- 150x150 km tiles

Landsat 7

- Launched in 1999
- 8 bands (Visible, NIR, SWIR)
- Resolutions: 1x15m, 6x30m, 1x60m

Landsat 8

- Launched in 2013
- 9 bands (Visible, NIR, SWIR)
- Resolutions: 1x15m, 8x30m



Whole tile

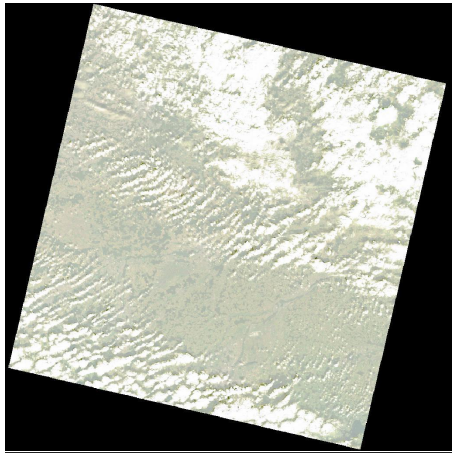
Band 3

6. Model for World

Input Data: Landsat 7&8 satellite images (2/2)

Problems with Landsat data

- Bigger time spans between images
- Nearly any Landsat 7 data available in Copernicus Data Space
- > 95% of WoSIS outside of Europe older than Landsat 8
- Sparse Landsat 7 data outside of Copernicus Data Space
- Downloads from official USGS API no longer possible



Whole tile

Band 3

6. Model for World

Target Data: WoSIS dataset

WoSIS (World Soil Information Service) Programme Overview

- **Purpose:** global initiative led by ISRIC with the goal of harmonization, collection, and dissemination of soil data
- **Coverage:** 228,000 profiles (1920-2020)

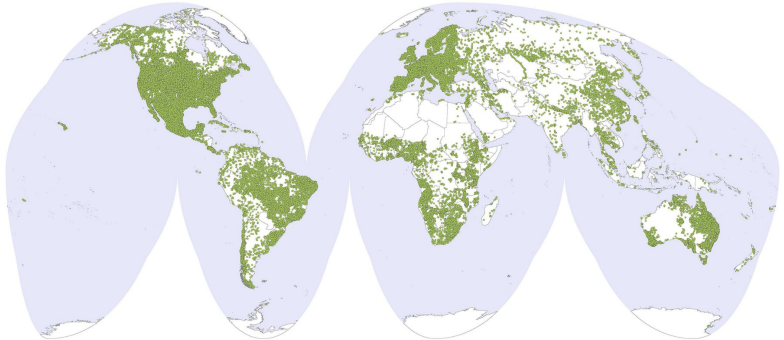
WoSIS 2023 Snapshot Dataset

- **Scope:** Geo-referenced soil profiles from across the world
- **Properties: pH**, organic carbon, **nitrogen**, **phosphorus**, bulk density, … But: **no potassium**

Data Preprocessing

- Timestamp availability and Landsat 7/8 coverage (>03/2008)
- Soil sample depth <= 25 cm
- Drop Europe

Final dataset → pH: 3,130; P=837; N=1,826



6. Model for World

Apply Europe Model to the World

Approach so far

- Model for Europe
- Extended model for Europe

} → Generalization for Africa?

Problem

- Only few soil data samples available for Africa
- No soil data samples with daily or monthly basis timestamp and hence no accurate satellite data available

Idea

1. Use Landsat 7/8 satellite data combined with WoSIS 2023 snapshot
2. Break world data into 5 continents
3. Validate and evaluate the extended model for Europe with each continent
4. If RMSE is similar as with LUCAS 2018 TOPSOIL Dataset
   → Model generalizes well

Continent	pH	P	N
Africa	1	0	1
Asia	34	26	34
Northern America	2,708	806	1,615
Oceania	366	2	155
South America	21	3	21

WoSIS soil data with timestamp

7. Summary

BASE	12 Sentinel-2 bands (single pixel)
SURR	12x3x3 Sentinel-2 image pixels
CLAY	1024 Clay AI Embeddings
WTHR	9 Weather Features
CRY	27 Crop Yield Scores

I	XGBoost
II	Fully Connected NN
III	Random Forest

Nutrient	Ref	Base	Extended
pH	5.99 ±1.36	1.02	0.80
P	27.0 ±27.0	25.5	23.1
N	3.15 ±3.70	3.44	3.27
K	204 ±208	178	159

Europe → Rest of World → ? → Africa

7. Outlook

Suggestions for further improvements I

- **African soil data with timestamp**
- The team behind the open weather data source we used is interested in the conducted research and any further projects
- More and newer data like satellite date from future measure campaign like LSTM on temperature and CIMR on climate change
- There are not open source sources for satellite data with higher resolution (up to 30 cm per pixel)
- Further additional information that can be delivered from the user can be crop rotation, former fertilization and adding of artificial water

More than 10 additional suggestions can be found in our project report

7. Outlook

Suggestions for further improvements II

Recommendations for the final app and rollout

- ML-Ops for
  - Tracking additional data
  - Resulting model updates
  - Delivered versions

- *“The real problem is understanding the right path to the best nutrient management recommendation and making them accessible and understandable to farmers”*

  Ennaji et. al.: Machine Learning in nutrient management: A review. Artificial Intelligence in Agriculture, 9:1-11, 2023

Discussion