

AgReFed Machine Learning Models

Machine learning for modelling and predicting agriculture systems and their uncertainties. This AgReFed Project is contributed by the Sydney Informatics Hub, The University of Sydney.

Project summary

The AgReFed Machine Learning Model project (AgReFed-ML) will contribute software that provides multiple machine learning workflows and tools for agriculture researchers. A particular focus is to develop machine learning models to map soil properties under sparse and uncertain input with support for spatial-temporal correlations and multi-covariates. While use-cases are developed for mapping soil bulk density, changes in carbon concentration, and soil moisture, the software can be used for a diverse range of soil property predictions such as sodicity, salinity, pH-values and many more.

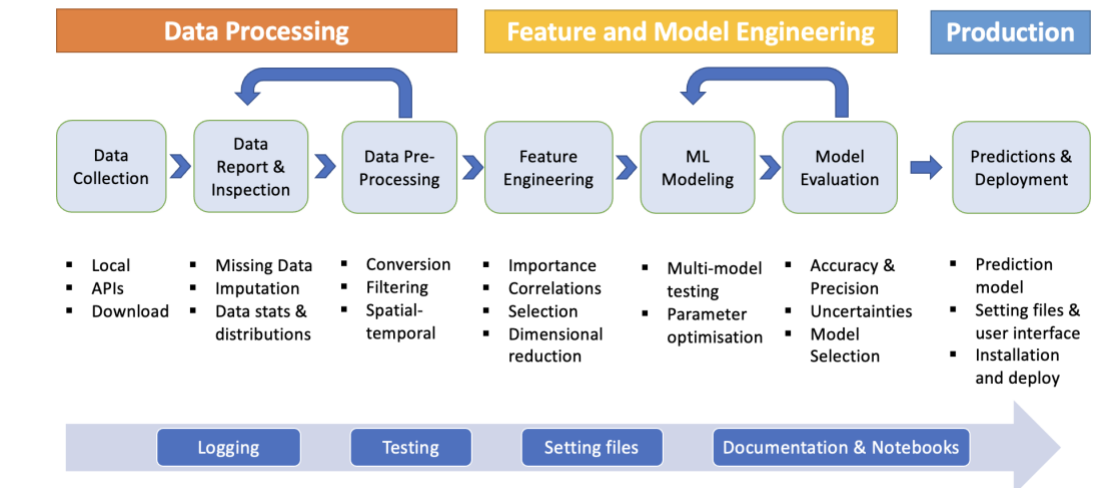
About

Problem: Currently agricultural researchers have models which are of high reuse value to the agricultural community. These models require inter-operable data flows of appropriately calibrated, cleaned data variables. Understanding the model limitations, assumptions and then interpreting the outputs is required. This takes time and a high level of expertise across a number of areas depending on the data type/s, data condition and model complexity.

Ideal Experience: Agriculture researchers will be able to extract appropriate inputs (e.g., weather, satellite) for user-defined locations and time periods, and to automatically convert these into a data cube (see project AgReFED Data Harvester) that is needed to run popular soil and agriculture models. These models can be either mechanistic (e.g., soil-physics) or of data-driven, statistical nature (e.g., Probabilistic Neural Nets, Bayesian Models, Random Forests), or a combination of both (e.g., for data-driven estimation/optimisation of mechanistic model parameters and their uncertainty). The output of these models, such as spatial-temporal predictions, can then be interrogated and used for the respective application (e.g., soil, yield, crops, animals).

Functionality

Machine Learning Process Overview



The main goal of the AgReFed-ML project is to enable researchers with reusable workflows and software tools to predict soil properties and uncertainties. The modelling approach should ideally have the following features:

- accommodate the spatial (-temporal) support of the observations
- accommodate the spatial (-temporal) auto-correlation of the observations
- accommodate measurement error of the observations
- incorporate cheap to measure and numerous variables as predictors (covariates)
- accommodate measurement error of the covariates
- when predicting give both a point and uncertainty (confidence interval) estimate
- be able to predict at any spatial (-temporal) support

As illustrated above, the typical workflow includes the following main steps:

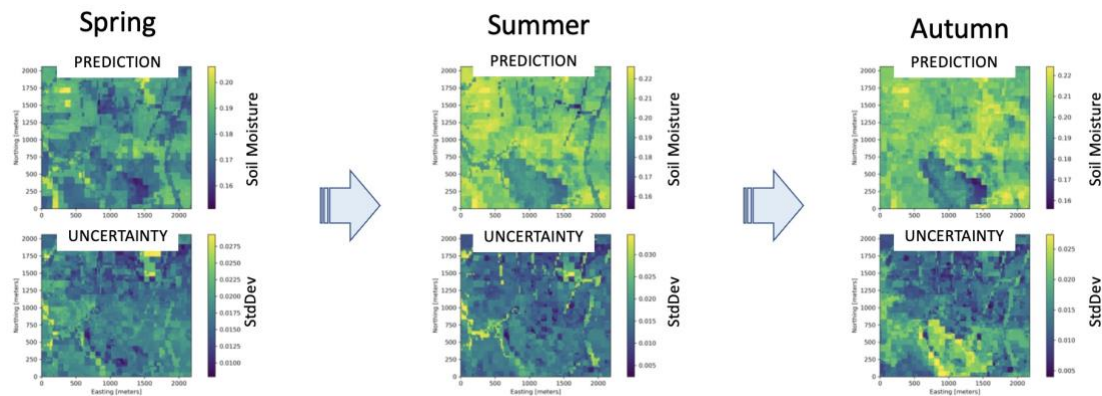
- Data aggregation (soil measurements, covariates) and processing
- Data exploration and feature engineering/selection
- Model exploration, parameter optimization (on training dataset), cross-validation (on test dataset), and model selection
- Generating prediction and uncertainty maps of soil properties

Use-case scenarios

Workflows and notebooks for three use case scenarios are provided as example applications for agricultural research:

- A. Static model: Focus on spatial modeling of soil properties for one given time.

- B. Change model for predicting for long-term change between two dates (e.g., change in organic carbon), including multiple averages of covariates before and after measurements
- C. Spatial-temporal model: Multiple time points regular space at smaller intervals (e.g., for soil moisture)



Project Details

Project Deliverables

- Python software package
- Documentation of package including functionality and installation
- Examples and use-case scenarios (presented in form of Python Jupyter notebook)

Project Plan

For this project we propose a workflow around the use of Gaussian Process Regression (GPR) which includes:

- a mean function relating the response to a data cube of predictors through a regression/ML model;
- a GPR on the residual to accommodate the measurement error and the spatial structure in the observations.

The output will be a map of soil properties on a grid of user define resolution and at point or block supports including uncertainty predictions. The machine learning models are demonstrated on three use-case scenarios.

Project Phase I

As part of the first stage of the project, the following software tools are developed and tested on farm soil data:

- tools for covariate feature selection and model evaluation

- Implementation and test of a range of mean function models, i.e. Random Forest and Bayesian Linear Regression
- GPR with custom spatial 3D kernel functions to include measurement uncertainties
- test of integrated GPR model on synthetic data set with spatial correlated fields and simulated uncertainties
- test on use-case scenario of spatial map for farm data of L'lara

Project Phase II

The goal of the second development phase is to develop an open-source software package which includes multiple additional ML feature improvements and use-cases:

- spatial-temporal modeling with focus on covariance between two time dates
- extraction of covariates with multiple time window averages
- use-case B: Map the change of soil properties (target: Organic Carbon stock) within a certain time period and estimate the uncertainty of change to allow hypothesis testing, including modeling of cross-correlation between 2 time points
- use case C: spatial-temporal modeling of multiple time points at smaller intervals (target: soil moisture)

More Information and Software Repositories

- Github Repository including Python packages, documentation and use-case example notebooks (To be released soon)
- [Feature importance package](#)

Feature Importance

