Project Scope: AgReFed Machine Learning Tools I

Project Administration

Title: **Mechanistic and data-driven models under uncertainty for agricultural systems**

alternative title: Machine learning tools for modelling and predicting agriculture systems and their uncertainties

Prepared by: Sebastian Haan

Date Prepared: Sep 2021

Project Manager: Sebastian Haan

Primary Client: AgReFed

Funding: 50/50 SIH/AgReFed

Project Summary

Research Context

Problem: Currently agricultural researchers have models which are of high reuse value to the agricultural community. These models require interoperable 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).

This project will contribute software scripts that provide multiple machine learning workflows and tools for agriculture researchers. One first pilot project will be to develop a software tool to map soil properties under sparse and uncertain input. While this tool will be tested first on mapping soil bulk density and carbon concnetration, it can be used for a diverse range of soil property predictions such as sodicity, salinity, pH-values and many more.

Client Needs

Software tool 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
- 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

- Optional: model multiple prediction targets simultaneously by taken into account the correlations between them.

Project Implementation

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.

A breakdown of the development steps is outlined in the following. First milestone (6 week FTE):

- extract and process soil and covariate data available for Llara (1 week FTE, depending on assistance from USYD researcher)
- exploratory data analysis (1 week FTE)
- develop pre-processing tools for covariate feature selection, coordinate conversions etc (1 week FTE)
- Implement and test a range mean function models, e.g. Random Forest, Bayesian Linear Regression, NN (1 week FTE)
- develop GPR with custom spatial-temporal kernel functions to include measurement uncertainties (1 week FTE)
- test GPR on synthethic data set with spatial correlated fields and simulated uncertainties (1 week FTE) #- test combined model on soil carbon and generate prediction maps (1 week FTE)

Second milestone (4 weeks FTE):

- Add mechanistic soil model (e.g. pedo-transfer function), either in mean function or on top of prediction output (1 week FTE). THe latter need to take into account merged covariances
- develop prediction tools for spatial area and temporal range predictions rather than for point locations only (1 week FTE)
- Documentation and installation guides (1 week)
- Package, test, and review software (1 week)

Data Availability

Soil data is available for L'lara (USYD, requires assistance) and covariates can be extracted from public data-sources

Scheduling and Availability

- Project update at weekly AgReFed meeting

Deliverables

- Python software package
- Documentation of package including functionality and installation
- Examples and use-case scenarios

SIH Skills required

- Python

- Data science, statistics and linear algebra knowledge
- Probabilistic machine learning, Gaussian Processes

In Scope

- Software package for data aggregation and processing
- Software testing and review
- Documentation of package including functionality and installation guide

Out of Scope & Exclusions

- Long-term maintenance and updates of delivered Software tools
- Installation and software assistance beyond testing phase and documentation
- Publication (can be considered in future depending on availability and need)
- User interface (potential as future optional update)

Time Estimate

Total Time: 10 weeks FTE