BIOAg Project Report

Report Type

PROGRESS

Title

Scalable assessment of soil organic carbon for carbon incentive programs

Principal Investigator(s) and Cooperator(s)

Kirti Rajaopalan (PI)
Jana Doppa, Deirdre Griffin LaHue, Dani Gelardi, Jordan Jobe (co-PIs)

Abstract

Incentive programs to promote climate mitigation and soil health often resort to incentivizing practice adoption and crudely estimated benefits, rather than the actual soil carbon accrual. A transition to incentivizing the benefit itself, aligned with the principles of true cost accounting, is critical. Our interdisciplinary team will tackle this aspect in the Pacific Northwest dryland systems. As part of this seed grant, we will develop and evaluate a machine learning model for predicting soil organic carbon in the 0-30 cm of soils by integrating satellite imagery with environmental and management covariates. Two hypotheses will be tested: (a) we can successfully predict soil organic carbon (SOC) in the 0-30 cm of soils with satellite observations, and (b) a gaussian-processes-based machine learning approach is better than traditional generative imputation approaches for small datasets. We believe the proposed outputs will strengthen an extramural resubmission to USDA NIFA.

Project Description

The goal is to develop and evaluate a machine learning model suitable for small datasets for predicting soil organic carbon (SOC) in the top 30 cm of soils by integrating multispectral satellite imagery, environmental and management covariates.

The model will be evaluated to test two hypotheses.

(a) We can successfully predict SOC in the 0-30 cm of soils with satellite observations.
(b) A gaussian-processes-based machine learning approach is better than traditional generative imputation approaches for our small dataset context.

Outputs

Overview of Work Completed and in Progress

Two graduate students have been recruited and weekly project meetings have been set up. We are on track to wrap up the data collation Task 1 (see Methods section below) by January 2025. In the meantime, we are starting to make progress on the model development Task 2 (see Methods section below) using tillage class data that is already compiled so that the basic modeling framework will be ready by the time Task 1 is completed. We are aiming for a related paper submission in January 2025.

Co-PI Jobe has started coordinating with Wenatchee Valley College and Heritage University on the workshop planned as part of the social dimension piece (see Methods section below).

We have also submitted a related extramural grant proposal to USDA NIFA's DSFAS program. Title and other details are noted below.

Methods

Task 1: Assemble response and predictor variables from multiple data sources.

Environmental covariates: SOC stocks are a balance between carbon inputs, losses derived from decomposition processes in the soil, and management practices. Key drivers compiled in multiple comprehensive reviews include those related to climate (temperature, precipitation, and vapor pressure deficit), organisms, terrain, soil properties, and agricultural land use and management practices. The following will be collated and regridded to the unit of analysis: daily weather variables at a 4 km resolution from 1979-current from the gridMET data product; soil properties at a 30 m resolution; topographical information from a 30 m resolution Digital Elevation Model; crop and rotation information from the WSDA and USDA Crop Data Layers.

Multispectral satellite imagery: Satellite imagery from Sentinel 1 (radar), Sentinel 2 (optical), and Landsat (optical and thermal) constellations will be used. Raw bands as well as over 80 vegetation, soil moisture, and tillage related indices as listed in the literature will be compiled. The reflectance data are available at 10 m to 60 m resolution depending on the satellite and reflectance band. Spectral features in the visible and shortwave infrared bands have been noted to be useful for SOC prediction. Recent work has demonstrated that the spatial, spectral, and temporal resolution of Sentinel imagery is sufficient to capture SOC variability both within field and at regional scales. While environmental covariates have been used in the literature, an important less explored aspect that can increase predictive capacity are the management practices. We plan to account for the crop type and rotation, and tillage practices in this seed grant.

Response variable: The Washington Soil Health Initiative (WaSHI) dataset includes the ground truth SOC response variable.

Common unit of analysis: Multi-scale data need to be brought to a common unit of analysis. The unit of analysis is the 40-hectare site definition used in State of the Soils Assessment. Data at smaller spatial resolutions than the unit of analysis will be aggregated via averages and percentile distributions (e.g., soil and terrain characteristics, satellite bands). Data available at larger spatial resolutions (e.g., gridded weather data at a 4 km resolution; management practice at field-scale) will apply to all units of analysis contained within the larger resolution.

Task 2 Develop and evaluate a machine learning model for SOC prediction.

The primary goal is to develop a model that can produce spatiotemporal predictions and uncertainty envelopes of SOC based on limited data by synergistically integrating scientific knowledge (e.g., higher precipitation zones with higher carbon inputs can be expected to have higher SOC interacting with management). This problem can be viewed as *imputing missing data* (i.e., predicting unknown data from known data)—e.g., given SOC measurements at selected sites and partial information about management practices, environmental covariates and numerous satellite imagery indices at those and other sites, we want to fill-in-the-blanks (or impute) missing information. There are three characteristics of the desired imputation models: 1) imputations should be spatially and temporally consistent 2) they should produce meaningful uncertainty estimates for informed decision-making; and 3) they should seamlessly interface with biophysical models in order to provide key inputs for future work.

By leveraging advances in generative AI, we propose to learn *generative imputation models*, which learn to directly generate samples from the underlying distribution of the missing data. Note that we can generate the most likely SOC that is conditioned on the measurements available in that region. Importantly, by running the models multiple times, any number of samples can be generated, allowing for Monte-Carlo estimates of uncertainty.

Unlike traditional generative models which are based on deep modeling and rely on enormous, complete datasets (e.g., millions of images), our application involves a *limited data regime* (e.g., hundreds/thousands measurements). To overcome this small data challenge, we propose to follow an alternative approach that leverages Bayesian modeling.

Gaussian processes for spatio-temporal imputation. We propose to cast generative imputation in a Bayesian framework. In particular, given a set of data D from a region R, limited SOC measurements and satellite imagery, our approach will produce a posterior distribution $\Pr(Y \mid D)$ where Y is a random variable corresponding to the missing/imputed values of interest in R that are not included in D. We focus on Gaussian Processes (GPs), which are a powerful modeling tool in this context. Importantly, given a learned GP model, it is straightforward to generate samples of Y, which is the key requirement of our application. GPs are also appealing because they also allow for various ways to incorporate domain knowledge (e.g., mass balance constraints and other input from biophysical models) unlike traditional approaches.

Task 3: Hands-on workshop on remote sensing and data science targeting 10-20 undergraduate students from underserved groups including first-generation college students from farm worker families.

The students will utilize the Google Earth Engine and explore an agricultural land use mapping research question. Students in agricultural domains often perceive that they do not have the skills needed to explore AI and machine learning applications, and additional encouragement and support is needed to build confidence. The aim of this workshop is to address this skill barrier in a friendly setting and introduce high-paying agricultural technology careers to students from underserved communities.

We will partner with the AgAID Institute and target students from Institute members Heritage University and Wenatchee Valley College—both serving students from underserved groups. AgAID Institute's partnership with these colleges is for its summer internship program. While the program has successfully recruited from underserved communities, it has been a challenge to find and encourage as many applicants as the institute would like to support. If we can break the perceived skill barrier through workshops targeting a larger pool of students, these students can become a direct recruitment pool for the AgAID Institute's internship program and provide a complementary benefit to the institute's social equity priority as well.

Publications, Handouts, Other Text & Web Products:

None to report yet.

Outreach & Education Activities:

None to report yet

Impacts

- Short-Term: Graduate students are being trained in interdisciplinary research.
- Intermediate-Term: None to report yet
- Long-Term: None to report yet

Additional funding applied for/secured

Title: Big insights from small data in climate-smart agriculture.
Program: USDA NIFA DSFAS Program
Status: Pending
Budget: $650K
Duration: 4 years

Graduate students funded

Md. Amine Gharsallaoui, PhD Student, Computer Science
Amin Norouzi Kandelati, PhD Candidate, Biological Systems Engineering

Recommendations for future research

None to report yet.