HydroSatML
Project Proposal

Michael Grant
Dane Jordan
Samir Patel
Rex Thompson

Contents

I. Executive Summary

We will develop a machine learning model that uses high-resolution multispectral satellite imagery to predict soil moisture in Eastern Washington. The model will be trained using in-situ soil moisture and weather observations from four farms over three growing seasons from 2012-2014, along with corresponding high-resolution satellite imagery provided by Planet Labs, Inc. (San Francisco, CA). The model will use new satellite images to predict soil moisture at locations across the satellite's field of view.

Our goal is to create a tool that provides farmers with useful insight into their fields' soil moisture at a scale, frequency and resolution that would otherwise be prohibitively expensive to achieve with today's technology. Such insight could be used to mitigate crop damage, improve yield predictions, and perhaps even improve yield itself.

II. Statement of Problem

Soil moisture is an important characteristic in agriculture as it has been shown to correlate strongly with plant health and crop yields. However, accurate soil moisture readings can only be obtained in-situ. Such measurements are expensive and impractical for capturing high-resolution variability in soil moisture at scale.

Remote sensing offers the promise of low cost measurements with high temporal and spatial resolution. However, there is currently no established method to leverage remote sensing, e.g. satellite imagery, to obtain accurate soil moisture values at scale.

III. Objective

We plan to produce the following two deliverables:

Model - A machine learning model or models that use high-resolution multispectral satellite imagery to predict soil moisture at a sub-field level. The model will be trained using in-situ soil moisture data and multispectral satellite images. It will predict soil moisture for different times/locations using additional multispectral satellite images.

Publication - A publication detailing the processes, methods and results of our work.

IV. Technical Approach

A. Identifying Needs of Customers

The primary customers for this project will be farmers in the Palouse region. Through the application of precision agriculture (PA), farmers can help optimize yields by efficient use of fertilizers, pesticides and water,which also can result in reduced costs and improved environmental sustainability.

B. Literature Review

Precision agriculture (PA) is the use of data to help drive decisions for farmers. This idea encompasses everything from using data to predict yields, to the precise application of pesticides, fertilizers and water. Currently, there is an overuse of resources in farming and as they become scarcer, practices like PA which reduce the overall inputs to the farm, will become necessary. A definition capturing these key aspects was provided by Dobermann et al. (2004), where he explains that PA is "a system[s] approach to managing soils and crops to reduce decision uncertainty through better understanding and management of spatial and temporal variability." From this definition, PA is a tool to help the farmer make better, more informed decisions with the objectivity of collected data.

In order to provide the services that PA can promise, sensors must be deployed or developed to collect the data of interest. A comprehensive review by Rossel et al. (2011) outlines a number of soil and plant sensors currently available or being researched while the review by Kim et al. (2009) focused on the technology behind soil sensors specifically. However, having the sensors is only part of the way there; on-the-go sensing is paramount to being able to quickly gather spatially diverse data. Technologies have been developed to ascertain on-the-go soil pH (Adamchuk et al. 2007; Adamchuk et al. 1999), and soil chemical and physical properties (Adamchuk et al. 2004; Adamchuk et al. 2005) although these systems are often expensive or unreliable.

In addition to proximal sensing, where measurements are either in contact or near the soil or plant, remote sensing leverages sensors mounted on drones, planes or satellites. These sensors often detect emitted or reflected light from wavelengths outside the visible spectrum. Light reflected from both visible and near-infrared (NIR) wavelengths contains information directly linked to both leaf structural and chemical properties (Yourek 2016). Vegetation indices (VI), produced by taking a ratio of a linear combination of different spectral bands, has been shown to correlate strongly with a number of important plant characteristic including leaf area index (LAI), leaf nitrogen content, chlorophyll, and biomass, to name a few (Haboudane et al. 2004; Thenkabail et al. 2000; Vories et al. 2014).

Soil moisture, or available water at the root zone, is one of the biggest determining factors of crop growth and yield. This single factor plays a big role in controlling the health of the surface vegetation and overall coverage (Schnur et al. 2010). However, collecting soil moisture is cumbersome and expensive due to the need for a large number of sensors placed in the field. Previous research has shown some promise using remote sensing to predict soil moisture, but often relied on basic regression techniques and/or not optimally chosen VIs (Adegoke & Carleton 2002; Schnur et al. 2010). Each of the previous studies used the normalized difference vegetation index (NDVI) which uses the NIR and red

spectral wavelengths. In this study, we will use the normalized difference red-edge index (NDRE) which uses the NIR and red-edge instead of the red band. This vegetation index has been shown to respond more strongly to chlorophyll in the plant (De Benedetto et al. 2013), which we feel will be a better indicator of soil moisture by more accurately analyzing the health and vigor of the surface vegetation. Additionally, we will use more advanced machine learning and/or deep learning techniques that we posit will improve upon previous research methods.

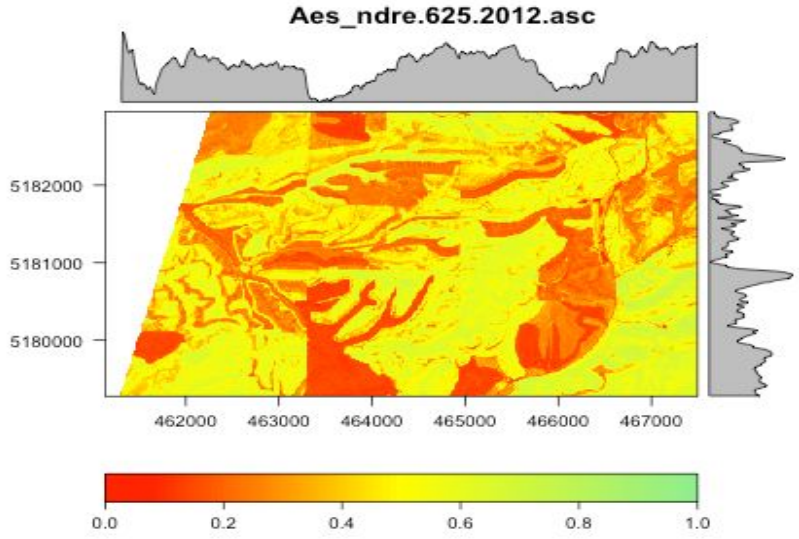
C. Design Objectives

1. Data Retrieval

We will obtain all relevant data from the project sponsor. We will also obtain a license for data retrieval from Planet.com through their “Planets Education and Research Program.” This will allow us to gather relevant satellite data that we may need in order to fill in missing data not made available by the sponsor. The additional data will be retrieved after performing an initial exploratory data analysis to determine what gaps exist in the data. We may also pull in additional data once we develop our first model.

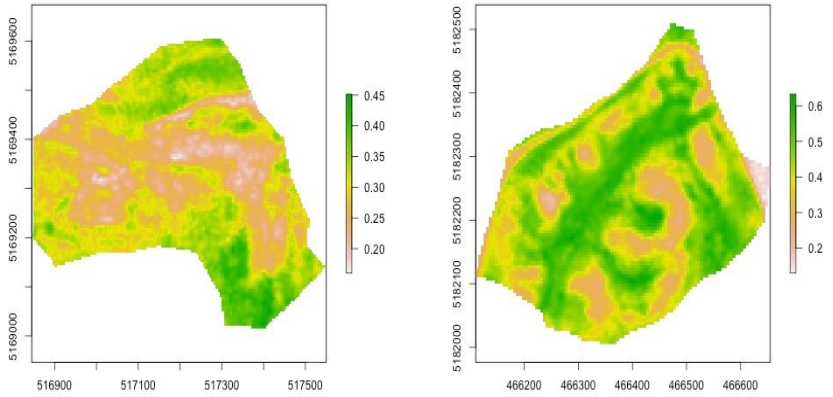
2. Exploratory Data Analysis (EDA)

We will perform an exploratory data analysis on the data received from the sponsor to determine what data we have and what data we think we may need (either additional requests from the sponsor or obtained separately from another source). This EDA will also allow us to explore how best to clean the data and develop our model.



3. Data Cleaning

Based on the exploratory data analysis, we will develop a functionalized cleaning process that can take in new raw data and clean it so it is in a format that is ready for modeling. Below are the cropped ndre images to the boundaries of two of the farms (part of the data cleaning process).



4. Data Augmentation

Because of the sparse nature of the in-situ soil moisture data, we plan to augment the in-situ soil moisture data by using a hydrological model to impute additional soil moisture values in between the locations of the actual soil moisture monitoring locations. The hydrological model will be based on a Digital Elevation Model, soil type, and precipitation. We will adjust the physical model output as necessary based on observed soil moisture readings.

5. Develop and Deploy First Neural Network (NN) Model

We will be working with Dr. Zaid Harchaoui to develop a model that best represents the data. Originally, we planned to use a time-series approach employing a Long Short-Term Memory (LSTM) recurrent neural network to model our data. However, due to the sparsity of the time points we have in our data this approach may not be obtainable. We will augment the data, as described above, but this will not increased the number of time points we have to work with. We still believe that a neural network will be best to model this data, but will likely have to work with a small multilayer perceptron instead of an LSTM to model this data. We understand that the time-series approach would likely be the best solution, but due to our data restrictions we may not be able to explore this avenue of research.

6. Iterations on Modeling and Validation

Depending on the results achieved from the first model (as determined through cross-validation and comparison to a held-out test set), we will iteratively review and modify our model to better fit the data, while remaining conscientious of potentially overfitting. Utilizing Dr. Zaid Harchaoui as a resource, we will change our model accordingly, with the intent to achieve more accurate results for previously unseen data.

With each iteration of a model, we will validate the results against a test set either as a subset of our data, or by gathering more sensor and spectral imaging data from other sources. Validation will ensure that the results we are achieving are accurate and representative of the sensor and spectral image data. Checking that the results are not overfitting the data will allow us to hopefully develop a model that can be used to analyze soil moisture given other spectral image data. In an effort to prevent overfitting we plan to utilize cross-validation and regularization techniques (e.g. L1, L2, dropout). We may also try model averaging and working with confidence intervals as opposed to estimated points.

As a final step of validating our model, we may compare model output for new satellite images against additional soil moisture data, which we would collect on our own or retrieve from others during the upcoming growing season.

7. Code Review

We will be performing a peer review as a quality assurance check on any source code created in the project. For Python code, we plan to follow the PEP8 style for all scripts and notebooks. For R code, the Google R style guide will be used. This process will help ensure code consistency and functionality, while allowing for improvements by all team members wherever necessary.

8. Presentation and Publication

As our final deliverable, we will present an overview of our project to the class during the first week of March 2018. The presentation will include an introduction, motivations, domain background, overview of processes used, results, conclusions and future steps.

We will also complete a first draft of a paper for publication in a journal.

D. Tools

We will be working with Python 3, an object-oriented programming language, popular for building data science tools, and machine learning models in particular. We will utilize the Jupyter Notebook environment for coding to analyze

soil moisture data provided by University of Idaho and normalized difference NDRE satellite image data from Planet Labs. Jupyter Notebooks are useful for collaborative data science work and in-line documentation. We anticipate using the following Python packages: matplotlib, numpy, pandas, and scikit-learn. We may also use the raster package in R, an open source programming language used for statistical computing and generating graphics, for initial data cleaning and aggregation.

We plan to utilize Amazon Web Services (AWS) Elastic Compute Cloud (EC2) instances to run Python models more efficiently. Data and models will be stored using Amazon Simple Storage Service (S3) for ease of importing and exporting between where the models will be run on EC2.

We anticipate using Keras, a deep learning library for Python which serves as a wrapper over Tensorflow, to develop our neural network. To validate the model we will be testing against a hold out test set of ground-truth data and calculate an appropriate error metric, likely a mean absolute error, or root mean squared error. We will also be comparing our approach to the naive physical model without machine learning augmentation. This will give us both an idea of how well our model is predicting this data, and how it compares to the current state of the art. Depending on the results we see, we will adjust our model(s) accordingly and to optimize performance.

We will use Git for version control. The project will be hosted on GitHub, and we will use branches as necessary to segregate different workflows.

V. Project Management

A. Timeline

HydroSatML Project Planner

B. Communication and Coordination with Sponsor

Our lead project sponsor is Dr. Dave Brown, a Professor in the Department of Crop and Soil Sciences at Washington State University in Pullman, WA. Also acting as co-sponsors are Dr. Erin Brooks, from Washington State University, and Matthew Yourek, from the University of Idaho.

Michael Grant will handle most communication between the team and the sponsors, mostly via email. The team and sponsors may occasionally hold conference calls, especially during the beginning of the project.

C. Team Qualifications

The team consists of Michael Grant, PhD, Dane Jordan, Samir Patel, and Rex Thompson. All are working towards a Master of Science in Data Science at the University of Washington. This project will serve as their Capstone.

Michael Grant has a PhD in Soil Chemistry from Washington State University. He is working on a precision agriculture project at at Microsoft Research. Michael will provide the bulk of the precision agriculture domain knowledge and deep learning experience.

Dane Jordan has an undergraduate degree in mathematics from Washington State University (2009). He worked as an analyst in the retirement industry from 2009-2017. Dane will provide feature engineering insight and quality assurance to maintain integrity.

Samir Patel worked in the semiconductor industry as a Chemical/Process Engineer at Intel, working on process development and improvement (optimizing yield, cost, output). He was at Intel from 2007-2016. Samir will provide project management experience and communication skills via presentation and technical writing.

Rex Thompson graduated with a B.S. in Atmospheric Science from the UW in 2009. He has previous experience with satellite data (MISR, MODIS), and also worked seven years as an air quality consultant providing emission reporting and regulatory support for clients in the energy industry. He brings experience using R to analyze satellite and raster data.

VI. References

(1) Adamchuk VI, Hummel JW, Morgan MT, Upadhyaya SK (2004) On-the-go soil sensors for precision agriculture. Comput. Electron. Agric. 44:71-91

(2) Adamchuk VI, Lund ED, Reed TM, Ferguson RB (2007) Evaluation of an on-the-go technology for soil pH mapping. Precis. Agric. 8:139-149

(3) Adamchuk VI, Lund ED, Sethuramasamyraja B, Morgan MT, Dobermann A, Marx DB (2005) Direct measurement of soil chemical properties on-the-go using ion-selective electrodes. Comput. Electron. Agric. 48:272-294

(4) Adamchuk VI, Morgan MT, Ess DR (1999) An automated sampling system for measuring soil pH. Trans. ASAE 42:885-891

(5) Adegoke JO, Carleton AM (2002) Relations between soil moisture and satellite vegetation indices in the US Corn Belt. Journal of Hydrometeorology 3:395-405

(6) De Benedetto D, Castrignanò A, Rinaldi M, Ruggieri S, Santoro F, Figorito B, Gualano S, Diacono M, Tamborrino R (2013) An approach for delineating homogeneous zones by using multi-sensor data. Geoderma 199:117-127

(7) Dobermann A, Blackmore S, Cook SE, Adamchuk VI Precision farming: challenges and future directions. In: Proceedings of the 4th International Crop Science Congress. 2004. vol 26.

(8) Haboudane D, Miller JR, Pattey E, Zarco-Tejada PJ, Strachan IB (2004) Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. Remote Sens. Environ. 90:337-352

(9) Kim HJ, Sudduth KA, Hummel JW (2009) Soil macronutrient sensing for precision agriculture. J. Environ. Monit. 11:1810-1824

(10) Rossel RAV, Adamchuk VI, Sudduth KA, McKenzie NJ, Lobsey C (2011) PROXIMAL SOIL SENSING: AN EFFECTIVE APPROACH FOR SOIL MEASUREMENTS IN SPACE AND TIME. In: Sparks DL (ed) Advances in Agronomy, Vol 113. Advances in Agronomy. Elsevier Academic Press Inc, San Diego. p 237-282

(11) Schnur MT, Xie HJ, Wang XW (2010) Estimating root zone soil moisture at distant sites using MODIS NDVI and EVI in a semi-arid region of southwestern USA. Ecological Informatics 5:400-409

(12) Thenkabail PS, Smith RB, De Pauw E (2000) Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. Remote Sens. Environ. 71:158-182

(13) Vories ED, Jones AS, Sudduth KA, Drummond ST, Benson NR (2014) SENSING NITROGEN REQUIREMENTS FOR IRRIGATED AND RAINFED COTTON. Appl. Eng. Agric. 30:707-716

(14) Yourek MA (2016) An Investigation of Crop Senescence Patterns Observed in Palouse Region Fields Using Satellite Remote Sensing and Hydrologic Modeling. In. University of Idaho.

VII. Appendix A: Resumés of Team Members

A. Michael Grant

B. Dane Jordan

C. Samir Patel

D. Rex Thompson