Data Analytics for Modeling Soil Moisture Patterns across United States Ecoclimatic Domains

Thomas Kitson
Computer Science
University of Delaware

Paula Olaya
Computer Science
Javeriana University

Elizabeth Racca
Computer Science
University of Delaware

Michael R. Wyatt II
Computer Science
University of Delaware

Mario Guevara
Plant & Soil Sciences
University of Delaware

Rodrigo Vargas
Plant & Soil Sciences
University of Delaware

Michela Taufer
Computer Science
University of Delaware

***Abstract*—Our poster presents a data analytics strategy to enable scientists to model patterns of soil moisture data at different resolutions across the United States. We build upon previous work of Guevara and co-authors with three contributions. First, we introduce divisions of soil moisture into the climatic regions proposed by the National Ecology Observatory Network. Second, we reduce the topological parameters used in modeling soil moisture using Principal Component Analysis. Third, we present an efficient workflow for modeling and visualizing soil moisture data.**

Keywords— k-Nearest Neighbors algorithm, Surrogate Based Model, Data prediction, Workflow.

I. Introduction

There is an increasing need for soil moisture information at scales relevant for environmental policy. Soil moisture data is used in land and resource management applications, for example, informing farmers of how much water they need to use on their crops. Soil moisture data is also a key factor in our understanding of climate change. The current standard in soil moisture data comes from environmental research satellites. Satellites provide nearly global coverage of soil moisture at a 27 km x 27 km resolution. This source of data has two main shortcomings. The first is the coarse resolution of the satellites' sensors. The other is the satellites' inability to measure soil moisture in areas of dense vegetation or snow cover, which produces gaps in the data [1].

To make this source of data more useful to scientists, we need a workflow for increasing the resolution of the data set and predicting values in areas of missing data. Guevara and co-authors addressed the problem by modeling the soil moisture data at the lower resolution of 1 km x 1 km supported by a k-nearest neighbors algorithm (k-NN) and a set of 15 topographical parameters describing soil lithodiversity, soil topodiversity and soil cronodiversity [2]. Specifically, to make fine-grain predictions for soil moisture, Guevara utilizes separate testing and training data sets. The training data is at a coarse grain and contains soil moisture labels. The testing data is at a fine grain and is missing soil moisture labels. In order to increase the resolution of the data, the scientists assign a soil moisture prediction to each point in the testing set that is the weighted average of the soil moistures of the nearest neighbors in the training set. It optimizes the choice of k (i.e., number of neighbors considered) and kernel (i.e., weighting function) for each given year over a range of x years (from 1978 to 2013) using 10-fold cross validation [2]. This work, however, still leaves open questions about the use of parameters, including how many are necessary to make good predictions. Furthermore, additional optimization of kNN may improve the predictions.

We leverage Guevara's method in three ways. First we split the data into climatic regions, controlling for the effects of climate. Second, we condense the topographic information used as predictors by Guevara *et al.* using Principle Component Analysis. Third, we develop a workflow for predicting finer-grained soil moisture at a 1 km x 1 km resolution. Specifically, our workflow consists of four steps. We divide training and testing data into regions. Then, the topographical data undergoes Principal Component Analysis (PCA) preprocessing to reduce the set of 15 topographical parameters in size to few principal components. Next, we assess the effectiveness of two modelling tools such as kNN and HYPPO [3] and apply the tools to the training data. kNN assigns each point in the testing set a soil moisture that is the weighted average of the soil moisture values of its neighbors. HYPPO or Hybrid Piecewise Polynomial combines KNN with a Surrogate Based Model, a global polynomial model of a surface. HYPPO uses a polynomial approximation in each neighborhood of k nearest points to predict soil moisture. After that we evaluate our model on the testing data. Finally, we visualize our predictions. Each step in our workflow is described in more detail in the following sections.

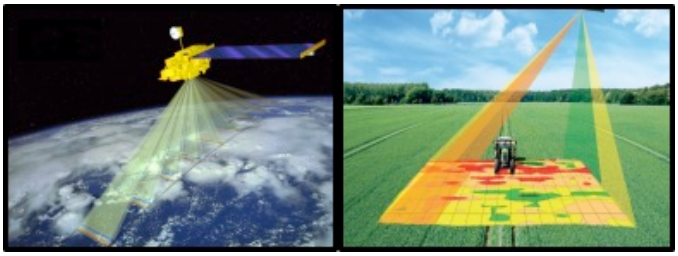


Figure 1. A satellite collecting pixelated data [4].

II. Input Data

The input data we used in our work comes from two main sources: satellite soil moisture data and a Digital Elevation Model (DEM). The satellite data, made available by the

European Space Agency as part of its Climate Change Initiative, consists of average annual soil moisture values across the United States at a 27 km x 27 km resolution. The information is collected as pixels covering an area on the Earth's surface, as shown in Figure 1. The data contains many gaps, representing areas where factors such as dense vegetation or snow cover prevented the satellites from getting a good reading. The DEM provided values of several derived topographic parameters at both 27 km x 27 km and 1 km x 1 km resolutions [5].

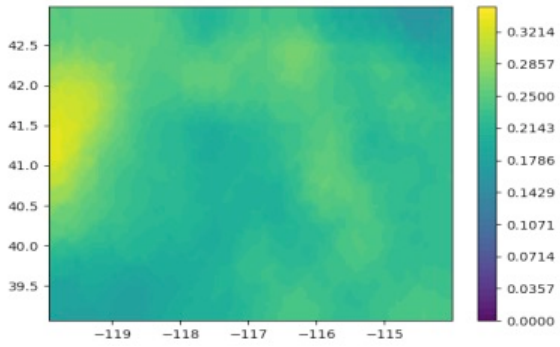
III. Predictive Workflow

We build a workflow for prediction that includes four steps: (1) we split the data into 20 Eco-climatic regions; (2) we apply Principal Component Analysis (PCA) to the data; (3) we test soil moisture prediction of k-NN and HYPPO algorithms; and (4) we visualize the actual soil moisture data and the differences from ground truth produced by k-NN and HYPPO.

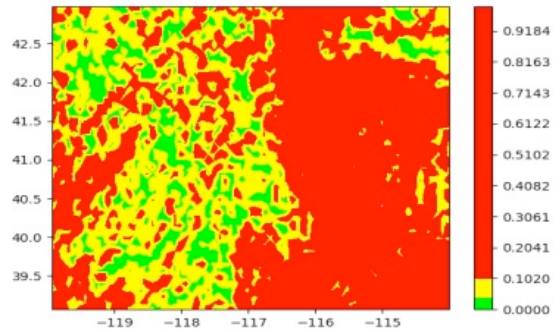
We use the data described in Section II. The training data consists of a series of vectors, one for each pixel in the satellite data. Each pixel is represented by a vector consisting of the latitude and longitude of the centroid of the pixel in the satellite data, the average soil moisture values of each year from 1978 to 2013 for that pixel, and the values of 15 topographic parameters from the DEM evaluated at the centroid. The testing data consists of a series of vectors representing 1 km x 1 km pixels from the DEM. Each vector consists of the centroid of the pixel in the DEM and the values of the 15 topographic parameters describing soil lithodiversity, soil topodiversity, and soil cronodiversity at that location [2].

There are two stages to our preprocessing: splitting the data into regions and compressing it using PCA. These are performed identically on both the training and testing data. The division into regions is motivated by our understanding that climate is the primary factor influencing soil moisture and topography is the second most important factor [2]. By dividing our data into the 20 Eco-climatic proposed by the National Ecological Observatory Network, we are able to consider regions in which the climate does not vary considerably. Thus, we are able to isolate the effects of climate and move into a scale where the effects of topography dominate and evaluate our topographically-based models at that level. This partition of the data is followed by a reduction of the number of predictors under consideration, using PCA. This is intended to minimize the number of additional features for modeling introduced by the addition of the topographic parameters from the DEM. This transformation replaces the topographic parameters with a set of derived components, with no impact on either the (x,y) coordinates or soil moisture values. The final product of the preprocessing stage is a separate set of training and testing data for each region under consideration. The models described below can be run on either the raw data, i.e. that used as the starting point for PCA, the processed version output by PCA, or a minimal version of the data, in which only longitude and latitude are available as predictors.

In generating models for the training data, we considered two algorithms: k Nearest Neighbors (kNN) and Hybrid Piecewise Polynomial (HYPPO). kNN assigns each point in the testing set a soil moisture that is the weighted average of the soil moisture values of the closest points to it in the training data. In our work, we use Johnston *et. al.*'s implementation of kNN in Python and Guevara et. al.'s implementation of kNN in R. The difference in kNN in Python versus kNN in R is explained by choice of k, choice of kernel, and the fact that the Python implementation uses a random sampling of points in the training set as opposed to the full set [2,3]. We use Johnston *et. al.*'s implementation of HYPPO in Python as our second model. HYPPO attempts to improve upon kNN by using polynomial approximations in each neighborhood of k nearest points [3]. This is different from the concept of kernel weighting used in kNN since it differentiates between the various features in the data and provides a more complex model.



(a)



(b)

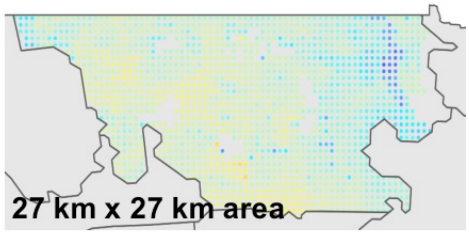
Figure 2. Output from our basic plotting tool (a) and our comparison tool (b).

We create two tools for visualization of our results. The first is plotting tool for soil moisture data. Our plotting tool displays soil moisture using a colormap, as in Figure 2 (a). On this scale 1.0 represents complete saturation of the soil and 0.0 represents the complete absence of water in the soil. The second is a model comparison tool. This software works on two models, evaluated at the same resolution, displaying points at which the models' predictions are within 5 percent of each other as green, within 10 percent as yellow, and more than 10 percent apart as red, as seen in Figure 2 (b). We utilize ArcGIS to generate maps for our poster [6].

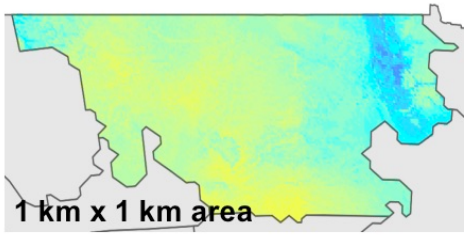
IV. Results

We ran our tools to model and visualize our data using a commodity desktop with an Intel i5 processor and 8 GB of memory. We used the original data sets used in the work of Guevara et. al. described in Section II and the methodology outlined in Section III to produce our results. Additionally, we selected an arbitrary year, in this case 1996, from which to draw all soil moisture values from the training data.

To visually assess the output of our workflow, we select the Northern Plains as a region of interest. We choose this region because it has a range of soil moisture values that captures the variation in the data and because it contains a large number of training points to sample from. Figure 3 shows the soil moisture for the Northern Plains region with the satellite resolution of 27 km x 27 km (a) and the soil moisture predictions with our workflow supported by the HYPPO model (b). Our poster presents further prediction and comparisons Guevara and co-authors' method.



(a)



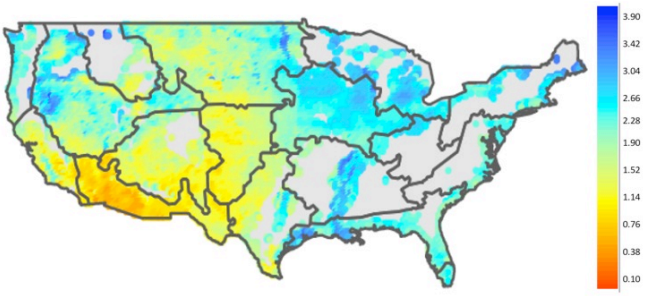
(b)

Figure 3. Comparison of 27 km x 27 km soil moisture data measured with satellites (a) and 1km x 1 km soil moisture data predicted with our workflow (b).

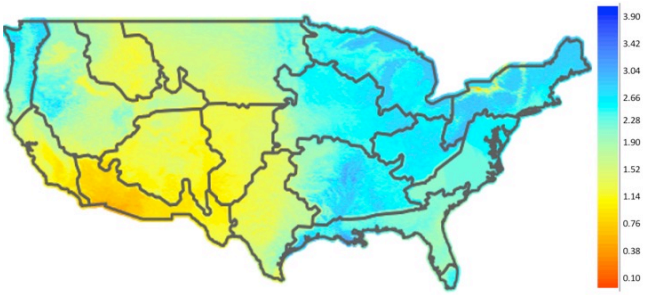
We present visualization of our soil moisture predictions for the continental US. We selected Guevara et. al.'s implementation of kNN with PCA preprocessing to see the effect of our refinements within the context of the state of the art practice. Figure 4 displays a visual comparison between the original training data (a) and our selected model (b). By implementing our model with PCA preprocessing, we are able to make fine mesh predictions of soil moisture on all the points in the testing set. A similar result could be produced from our other models.

V. Conclusions

We provide a workflow to the process of modeling soil moisture based on satellite data that takes into account the effects of climate on soil moisture by dividing the modeled area into climate regions and uses Principal Component Analysis to condense the topographic parameters with minimal loss of information. We contribute in the development of tools for scientists to use to analyze soil moisture data and visualize results. In the future, we will develop a means for validating our models and explore the use of other modeling techniques to make our soil moisture predictions. We envision that our tools can be scaled up to make real-time predictions of soil moisture in areas of missing data. To accomplish this, we eventually plan to parallelize and automate certain components of our workflow.



(a)



(b)

Figure 4. Comparison of 27 km x 27 km soil moisture data measured with satellites (a) and 1 km x 1 km soil moisture data predicted with our workflow (b).

References

[1] Liu YY, Parinussa RM, Dorigo WA *et al.* (2011) Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals. Hydrology and Earth System Sciences, 15, 425–436.

[2] Guevara M. and Vargas R. Soil moisture decline in conterminous United States. In Review.

[3] Johnston, T., Zanin, C., Taufer, M. (2016). HYPPO: A hybrid, piecewise polynomial modeling technique for non-smooth surfaces. In Proc. of ICPADS Conference.

[4] From https://directory.eoportal.org/web/eoportal/satellite-missions/t/terra

[5] Becker JJ, Sandwell DT, Smith WHF *et al.* (2009) Global Bathymetry and Elevation Data at 30 Arc Seconds Resolution: SRTM30_PLUS. Marine Geodesy, 32, 355–371.

[6] Environmental Systems Research Institute (ESRI). (2012). ArcGIS Release 10.1. Redlands, CA.

ACKNOWLEDGMENTS: With the support of NSF OAC #1724843.