Predicting Soil Conditions Using Machine Learning Algorithms

Vamsi Gondi
vkgondi@bsu.edu

David Hua
dhua@bsu.edu

Biju Raja Bajracharya
bajracharya@bsu.edu

CICS, CCIM
Ball State University
Muncie IN

Abstract

Soil conditions are identified and predictions are prioritized by various organizations and government agencies daily. Changes in global climate patters have resulted in increased tropical storms, hurricanes, flooding, and draughts. The challenges we face will become even more adverse in the future if we cannot use resources effectively. Unprecedented floods and food shortages are examples of the challenges being experienced at various locations around the globe. There are studies and instrumentation deployed across the globe for identification of current soil conditions and its applications for agriculture, hydrology, and flood management. Most of these models, however, cannot predict future conditions. This study utilized future weather forecast models and soil conditions to predict the impact on various applications such as agriculture, flash floods, and river flooding. In our study, we collected data from various locations remotely in real-time targeting three different characteristics of ground conditions; volumetric water content, conductivity, and soil temperature. Using future weather forecasts and current ground data conditions we trained various machine learning algorithms (IBM's SPSS, SVM) to predict all three characteristics of ground conditions for 15 days into the future. The predictions were then compared to the actual data collected over the prediction period and found reliable results that can be deployed for usage for agriculture, stormwater runoff models, and flash flood determination.

Keywords: Prediction Analytics, Soil Volumetric Moisture Content, Soil Temperature, Soil Conductivity, Soil PH, Machine learning algorithms, SPSS, Support Vector Machine (SVM).

1. INTRODUCTION

The climate changes and misuse of resources have resulted in extreme weather conditions and natural disasters across the world. This is well documented in various scientific studies which are caused by natural and manmade changes to the environment (three gorges dam, Hubei, China) (Srivastava A et.al, 2020) (Mujumdar M et.al., 2020) (Hafid 2020). Researchers are proposing new and innovative machine learning algorithms to predict the impact these events have on outcomes such as floods, droughts, water shortages, landslides, and flash floods.

In our study, we identified soil as one of the parameters which we can predict for various applications. The soil moisture content can help determine what will happen after heavy rains. Will the water be absorbed by the soil or will it runoff, if the land is mountainous, to create landslides? In the case of agriculture, more efficient and effective field watering schedules can be planned based soil conditions and weather forecasts.

In our study we used data from Clemson's NSF project based at Baruch Institute at Georgetown, SC. Clemson's researches deployed Decagon 5TE instruments to collect real-time data from the soil of volumetric water content in the soil, conductivity, and temperature of the soil. They also provide historical data for research purposes. We used these real-time and historical data to train our machine learning algorithms to predict soil conditions for 5 to 10 days.

The paper is organized as follows: Section 2 identifies the various machine algorithms for prediction models; Section 3 presents the SPSS results; Section 4 shows the data from SVM; Section 5 lays out our future research; and Section 6 provides concludes the paper.

2. PREDICTION MODELS

Prediction models come in various shapes and sizes. There are many methods and techniques that can be used to build a prediction model, and more are being developed all the time. The mostly frequently used predictive models are Linear Models, Neural Networks, Decision Tree Models, Cluster Vector Models, and Support Vector Machines. They are all machine learning algorithms. There are different ways to generate each of these models. Machine learning combines computer science and statistical analysis to improve prediction. High accuracy predictions can be achieved within human-computer interaction. Machine learning algorithms are categorized as supervised and unsupervised. Supervised algorithms learn on the basis of labeled data and produce the result. Unsupervised data does not use labeled data for learning (Kavitha S 2016).

IBM's SPSS (IBM SPSS 2020) is another tool that is widely used for statistical analysis, machine learning algorithms, and for big data. We used SPSS and SVM to predict soil conditions with historical data from sensors and weather and future weather conditions.

3. SPSS

Data was collected from sensors deployed at four different locations which consisted of different soil compositions. We fed 10 days of historical data with weather conditions on those days to the algorithm. We also fed 10 days of future weather conditions to the algorithm and predicted volumetric water content, conductivity, and soil temperature. We collected the real-time sensor data for the next 10 days and plotted the graphs from Figure 1 to 12.

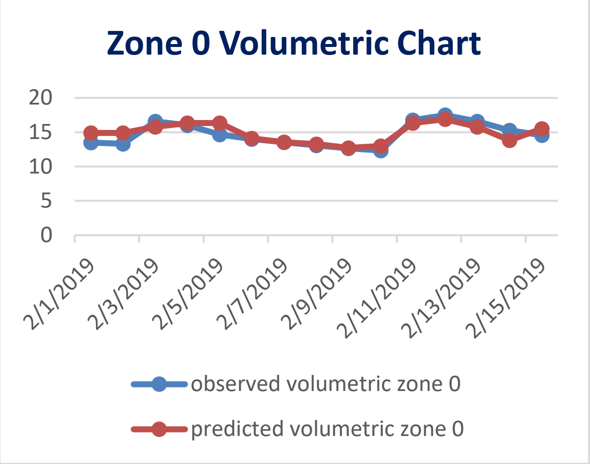

Fig. 1 Zone 0 Volumetric Water Content Chart

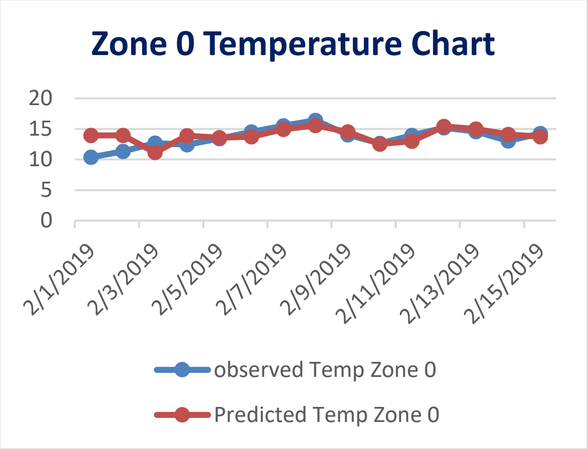

Fig. 2 Zone 0 Temperature Chart

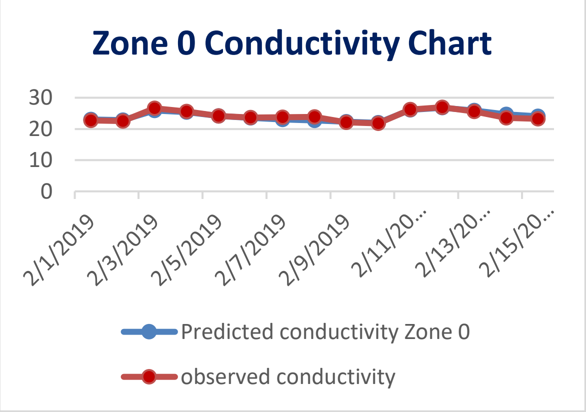

Fig. 3 Zone 0 Conductivity Chart

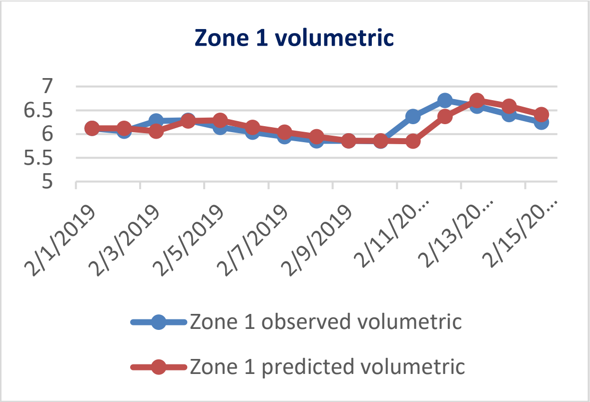

Fig. 4 Zone 1 Volumetric Water Content Chart

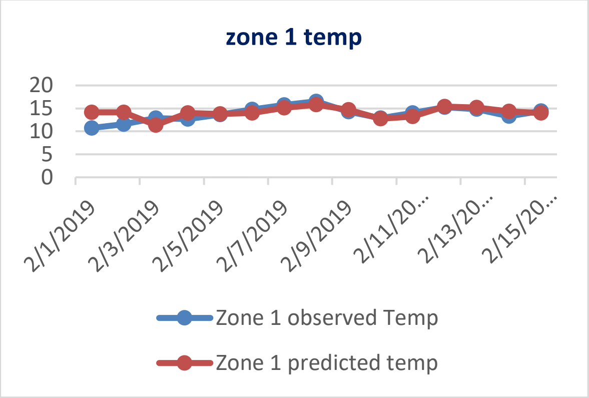

Fig. 5 Zone 1 Temperature Chart

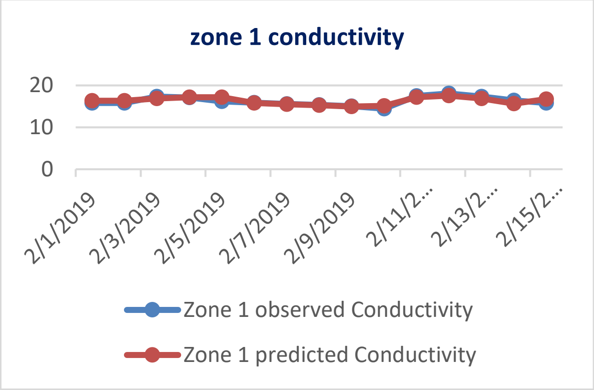

Fig. 6 Zone 1 Conductivity Chart

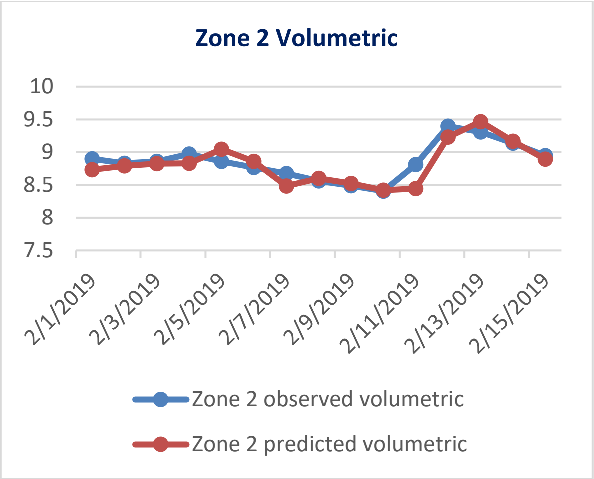

Fig. 7 Zone 2 Volumetric Water Content Chart

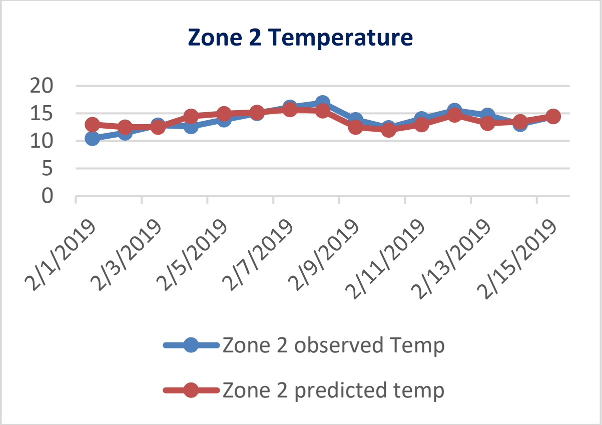

Fig. 8 Zone 2 Temperature Chart

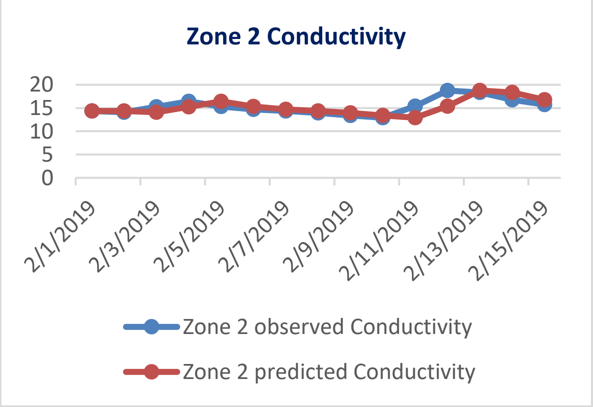

Fig. 9 Zone 2 Conductivity Chart

Charts for Zone 3

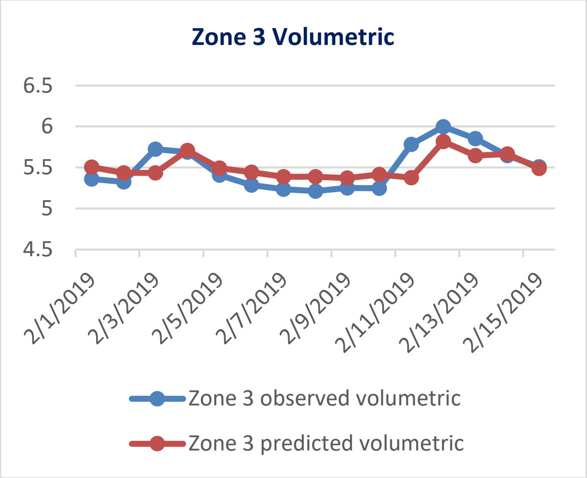

Fig. 10 Zone 3 Volumetric Water Content Chart

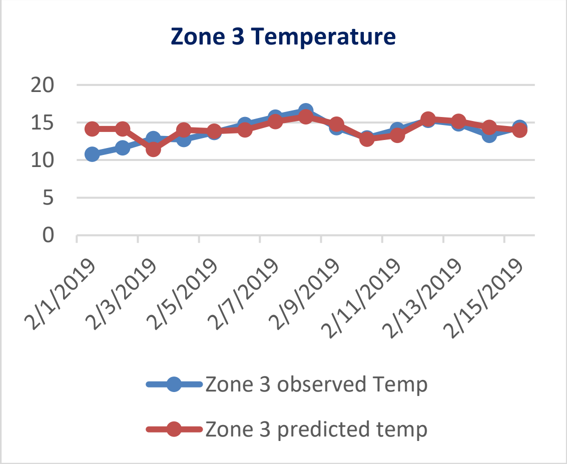

Fig. 11 Zone 3 Temperature Chart

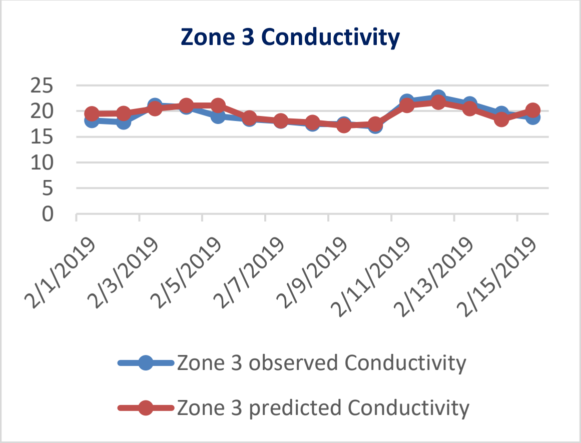

Fig. 12 Zone 3 Conductivity Chart

We also calculated R square and stationary R square for the predicted values to determine the accuracy of the prediction values for all four zones. Table 1 depicts the values negative values to near 1, negative values show the failure of the prediction and near 1 is the accuracy.

	Zone 0		
	Volumetric	Temperature	Conductivity
R square	0.766	0.490	0.872
Stationary R square	0.766	0.561	0.872
	Zone 1		
	Volumetric	Temperature	Conductivity
R square	0.425	0.551	0.730
Stationary R square	-0.002	0.594	0.730
	Zone 2		
	Volumetric	Temperature	Conductivity
R square	0.742	0.510	0.330
Stationary R square	0.571	0.510	-0.005
	Zone 3		
	Volumetric	Temperature	Conductivity
R square	0.484	0.536	0.735
Stationary R square	0.484	0.582	0.735

Table. 1 R Square and Stationary R Square values

4. SVM

The machine learning library scikit-learn was used for our prediction model. The prediction model was implemented in python. Three models were used for predictions: multivariate linear regression model, decision tree regression model, and Support Vector Regression (SVR) model. The best prediction was achieved from Support Vector Regression (SVR) model. Support Vector Regression (SVR) is a regression model which comes from the Support Vector Machine (SVM) for predicting continuous values instead of classification. SVM is a supervised machine learning algorithm that can be used for classification or regression problems (V. Anandhi 2013) (Alex J. Smola 2004) (SR Gunn 1998). It uses a technique called the kernel trick to transform data. It finds an optimal boundary between the possible outputs based on some extremely complex data transformations. There are mainly two kernels used which are linear and non-linear (Kavitha S 2016) (SR Gunn 1998). Linear kernel means the boundaries will be a straight line and non-linear means that the boundaries that the algorithm calculates don't have to be a straight line. The benefit of the non-linear kernel is that it can capture much more complex relationships between data points. Soil Temperature, Volumetric Water Content, and

Conductivity is predicted separately for a particular day where the independent variables are the Air Temperature and Rainfall of that particular day. For predicting soil temperature, we used a linear kernel in the SVR model and for predicting Volumetric Water Content, and Conductivity we used non-linear kernel Radial Basis Function (RBF) in the SVR model.

For our prediction model, we used 15 days of data. The first 10 days of data were used for training the model and the last 5 days of data for testing. We predicted the data for the future 5 days for each of the variables. Predicted values using a support vector machine (SVR) regression model and actual data are shown in Figures 13 to 15.

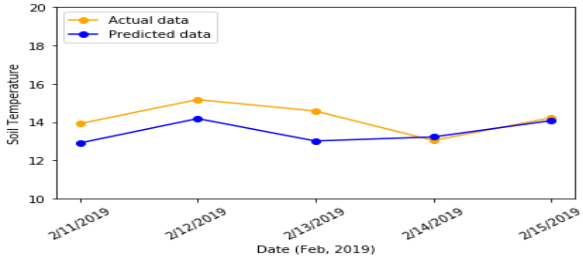

Fig. 13 Predicted Soil Temperature data vs Actual data

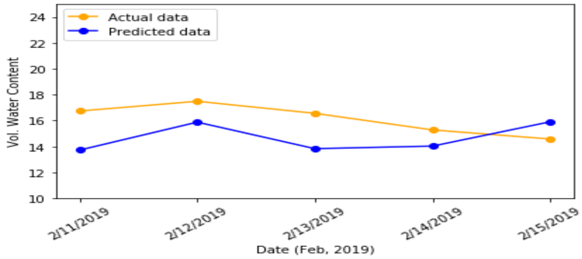

Fig. 14 Predicted Vol. Water Content data vs Actual data

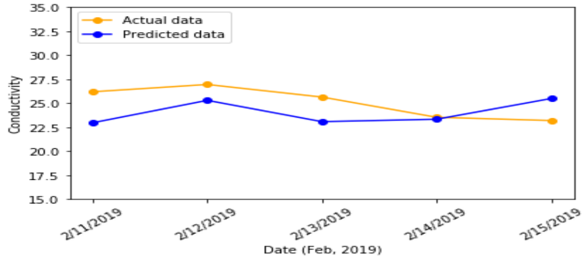

Fig. 15 Predicted Conductivity data vs Actual data

We considered MAE (mean absolute error) as an error measurement scale. MAE for predicted values of different variables are shown in Table 2.

Variables	MAE
Soil Temperature	0.78
Volumetric Water Content	1.98
Conductivity	1.99

5. Future Work

The data predicted in this study need to be fine-tuned before being applied commercially. The proposed prediction models can be implemented for agricultural purposes. Before using our model to predict potential natural disasters such as floods, landslides and flash floods, additional research investigating the many weather prediction models needs to be conducted. We are currently studying on how they can be incorporated for a wide range of applications.

6. CONCLUSIONS

With the food shortages across the globe and changes in climatic conditions, it is necessary to predict what ground conditions are going to be for saving lives, infrastructures, and food security. The prediction models proposed in this study can achieve adequate results, but these studies need to be extended to build complex models. The future models we plan to develop and collaborations with other researchers will produce groundbreaking new and complex machine algorithms for safeguarding future generations.

9. REFERENCES

Srivastava A., Singhal A., Jha P.K. (2020) Climate Change—Implication on Water Resources in South Asian Countries. Resilience, Response, and Risk in Water Systems. Springer Transactions in Civil and Environmental Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-15-4668-6_12

Mujumdar M. et al. (2020) Droughts and Floods. Assessment of Climate Change over the Indian Region. Springer, Singapore. https://doi.org/10.1007/978-981-15-4327-2_6

Hafida (2020) Climate change and natural disaster: A case study of flood affected women of Assam, north-east. In: Hafida Begum International Journal of Academic Research and Development, ISSN: 2455-4197, Volume 5; Issue 1; January 2020; Page No. 67-70

V. Anandhi and R. Manicka Chezian. (2013). Support Vector Regression to Forecast the Demand and Supply of Pulpwood, International Journal of Future Computer and Communication, Vol. 2, No. 3, June 2013.

Kavitha S; Varuna S; Ramya R. (2016). A comparative analysis on linear regression and support vector regression, 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 19-19 Nov. 2016.

Alex J. Smola, Bernhard Scholkopf (2004), "A tutorial on support vector regression" in Statistics and computing, Springer, 2004.

SR Gunn (1998), Support Vector Machines for Classification and Regression, ISIS technical report, 1998.

IBM SPSS (2020). IBM SPSS software. Retrieved from https://www.ibm.com/analytics/spss-statistics-software (On line: August 27, 2020)