Machine Learning for Predicting Saturated Hydraulic Conductivity

William Duy, Ty P.A. Ferre
University of Arizona, Department of Hydrology & Atmospheric Sciences

Introduction

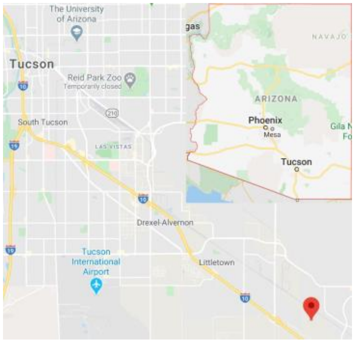
Saturated Hydraulic Conductivity (K_{sat}) is an important factor in planning and development for water projects to predict solute transport. There are ways to measure K_{sat} such as the falling head test or Constant head test. Tests such as these require potentially expensive equipment to be brought into the field to measure on site or, for deeper measurements core samples need to be collected and returned to the lab for analysis.

Objective:

The purpose of this project was to use machine learning to create models to predict saturated hydraulic conductivity (R^2) using fewer, more easily measured and collected predictive variables.

Site Description

The samples were collected at the University of Arizona Tech Park site located 10 miles SE of Tucson. The sample site is surrounded by the T.E.P. solar farm with very little ground cover.



UA Tech Park
Global Solar Energy
S Rita Rd
10
Google

Methods

The samples were collected via Continuous Split Spoon Sampling to a depth of 85 feet every 5 feet. They were then analyzed to determine 9 modeling variables, particle size distribution in seven bins, porosity, and dry bulk density. Using 70% of the data for training Random Forest and Gradient Boosting models were created for all combinations of the variables over 10 different selections of training data. The remaining 30% of the data was then tested with the models and the predicted K_{sat} value was compared to the measured K_{sat} to determine the R^2 value. The best models for each number of variables were then aggregated and the Importance value for each variable summed and divided by the number of times used to create a standardized importance for each variable.

After running the models with a random selection of samples from all of the boreholes, the model was recreated while using all samples from a single borehole to test the robustness of the model.

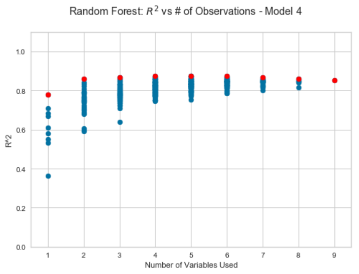
Results and Conclusions

When using a random selection of training data from all boreholes, all iterations of the models for both Random Forest and Gradient Boosting showed the increase in the R^2 value becomes negligible when using more than 4 variables. Both modeling methods do yield R^2 values over 0.85 when using 4 variables. According to the standardized importance of the 4 variable models the most important variables are 0mm, 0.063mm, 0.125mm, 0.25mm.

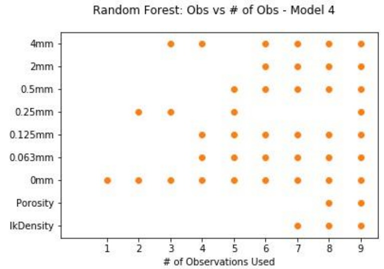
When using a single hole for training the model the results are not as good but show a similar pattern of smaller increases in R^2 as more variables are added. In this case the R^2 doesn't have a noticeable increase after using more than four variables, 0mm, 0.063mm, 0.125mm, and 0.5mm

The results show that a reasonably accurate model can be made from training on a single borehole and only using particle size data to estimate K_{sat} of the other nearby boreholes.

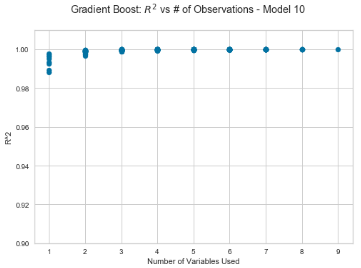
Random Forest: R^2 vs # of Observations - Model 4



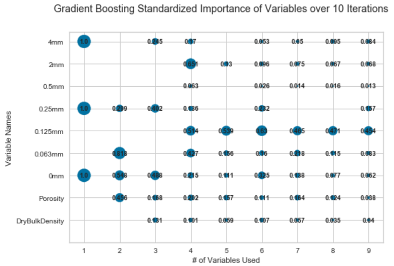
Random Forest: Obs vs # of Obs - Model 4



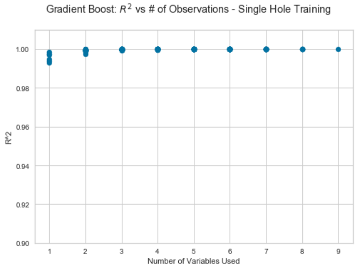
Gradient Boost: R^2 vs # of Observations - Model 10



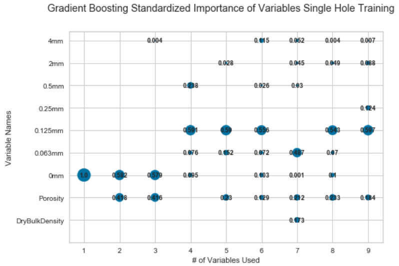
Gradient Boosting Standardized Importance of Variables over 10 Iterations



Gradient Boost: R^2 vs # of Observations - Single Hole Training



Gradient Boosting Standardized Importance of Variables Single Hole Training



Acknowledgments: Thank you to Ty Ferre, Steve Farrington, and Jeff Klakovich for this opportunity and guidance on this project