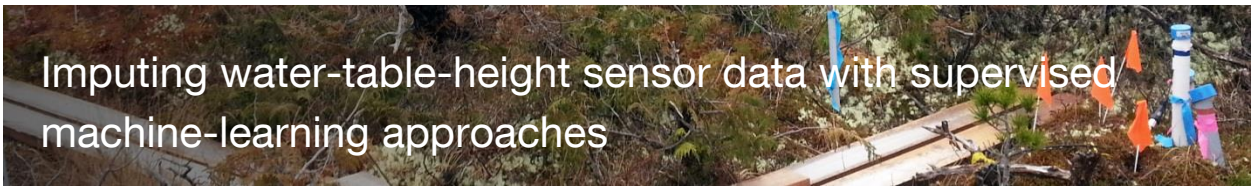

Imputing water-table-height sensor data with supervised machine-learning approaches

David Levy-Booth,
UBC Microbiology

Summary

Machine-learning approaches were tested for imputation of missing soil water table depth measurements from pressure transducer sensors. Of the methods tested, gradient boosting regression improved prediction accuracy over linear methods, could handle missing values in the model and provided a robust estimation of prediction confidence.

Problem

The Hakai terrestrial sensor network uses a suite of pressure transducers for continuous monitoring water table depths at three locations: TSN1: Pond Margin Fen, TSN2: Bog Forest and TSN3: Deep Bog. Such sensor networks allow real-time monitoring of environmental conditions but are subject to occasional interruption due to malfunctions, servicing and wildlife.

Computational methods to determine when data is missing, or when a sensor is returning data that is different from expected values, can be used to select data to be filled (imputed). Imputed data can be associated with confidence intervals to determine reliability of the fitting method.

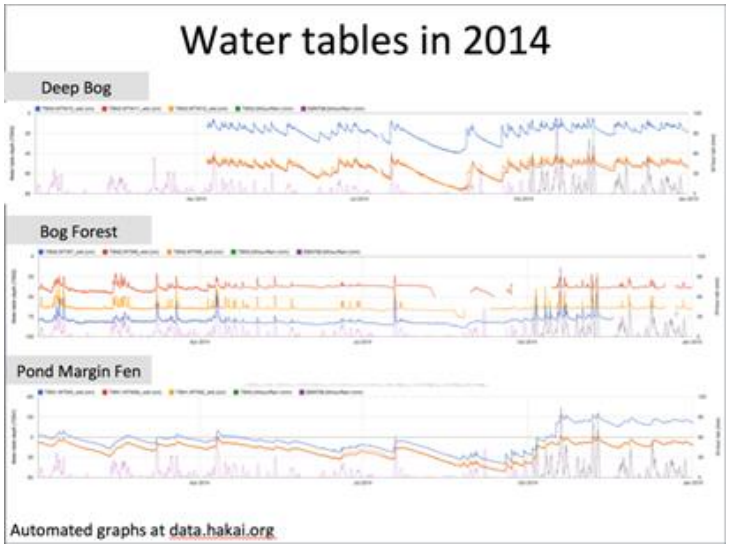

Figure 1. Raw data showing areas of missing values.

The focus of this document is to compare supervised machine-learning imputation methods.

Approach

A variety of supervised machine-learning regression approaches to sensor data imputation are investigated and benchmarked alongside simple linear regression methods. For all methods the same categorical variables were used to fit the water table height at TSN 2-Well 7.

Method	Description	R Package
Gradient boosting machine regression (GBM)	Produces an ensemble of weak classification and regression tree (CART) models using a random subsample of a training dataset.	gbm (Friedman, 2001)
Support vector machine regression (SVM)	Plots n predictor values in n-dimensional space, and finds the n-1 dimensional hyperplane that describes relationships between predictor variables and response.	e1071 (Cortes and Vapnik, 1995)
Random Forest Regression (RF)	Similar to GBM, produces an ensemble of CART models using random selection of features and training data cases.	randomForest Breiman, 2001)
Artificial Neural Network (ANN)	Uses "hidden layers" of non-linear transformations to find relationships between predictor and response variables. Requires data pre- and post-processing.	nnet (Ripley, 1996)

Machine Learning Imputation

To assess the accuracy of machine-learning predictions, a subset of data for TSN2 well 7 was predicted and compared against known values. Predicted data were rounded to the first decimal place. Models were tuned to minimize root mean squared error (RMSE). The area under the curve (AUC) of a receiver operating characteristic (ROC) curve was maximized for an accurate prediction (AUC = 1 for perfect prediction). Additionally, the computational intensity of a model fitting was measured in the time the function took to fit a subset of 4685 observations that had no missing values (run in 32-bit R 3.0.3 on a 3 GB Intel Pentium Dual CPU T3400 2.16Ghz computer with Windows 7).

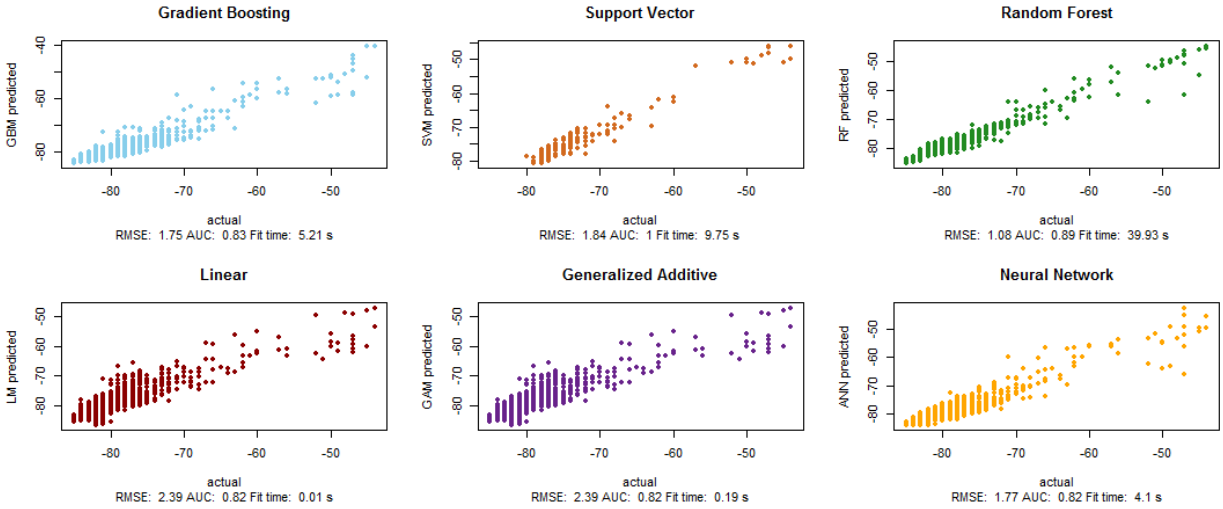

Figure 2. Comparison of linear and machine-learning models

All machine learning methods improved the prediction accuracy relative to the linear model. Although Random Forest regression was the most accurate method tested, Random Forest models cannot handle missing values in the predictor variables. Random Forest was also the most computationally intense method tested and does not have a robust method for calculating confidence intervals.

Gradient boosting regression and artificial neural network regression were the next most accurate methods and were able to predict the response variable even with missing values in the predictor variables. Of these methods, gradient boosting was also able to produce robust estimation of prediction confidence intervals and can provide an estimation of categorical variable importance. Therefore, gradient boosting regression was selected as the final method for sensor data imputation.

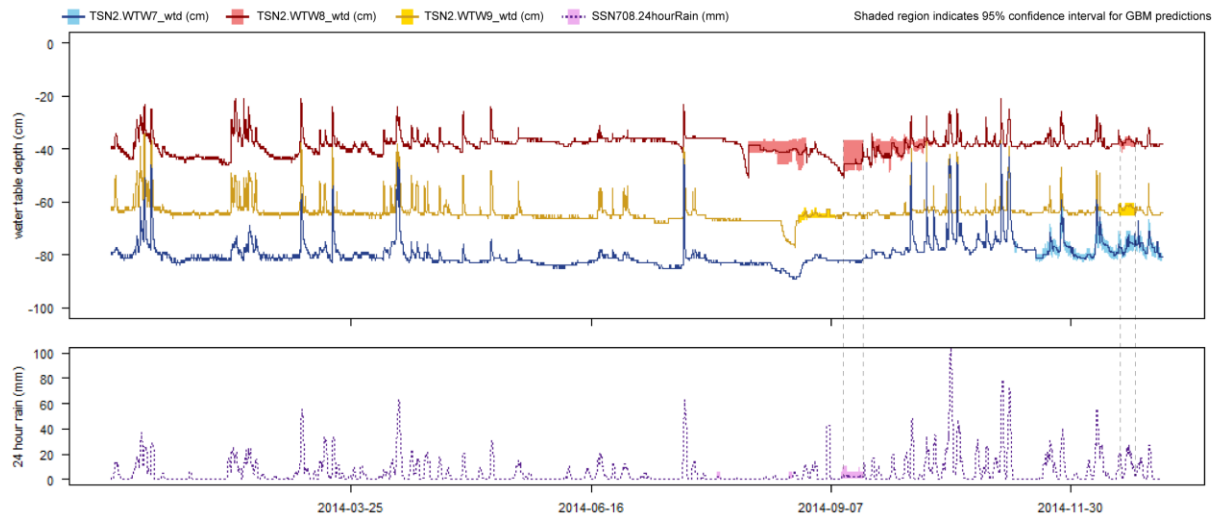

Figure 3. Imputed data using gradient boosting regression.

Gradient Boosting Regression Model

In addition to the prediction of specific water table values for TSN 2 well 7, GBM can be used to estimate the confidence interval of predicted values. Here, the 95% intervals are shown for all TSN 2 wells and for rain measurements at watershed 708. The 95% intervals for TSN2 well 7 (blue) fit tightly with the predicted data, showing a high level of confidence in these predictions. However, for TSN 2 well 8 (red), the intervals are much larger due to the lack of rain data in September of 2014. This shows the benefits and drawbacks of the GBM method.

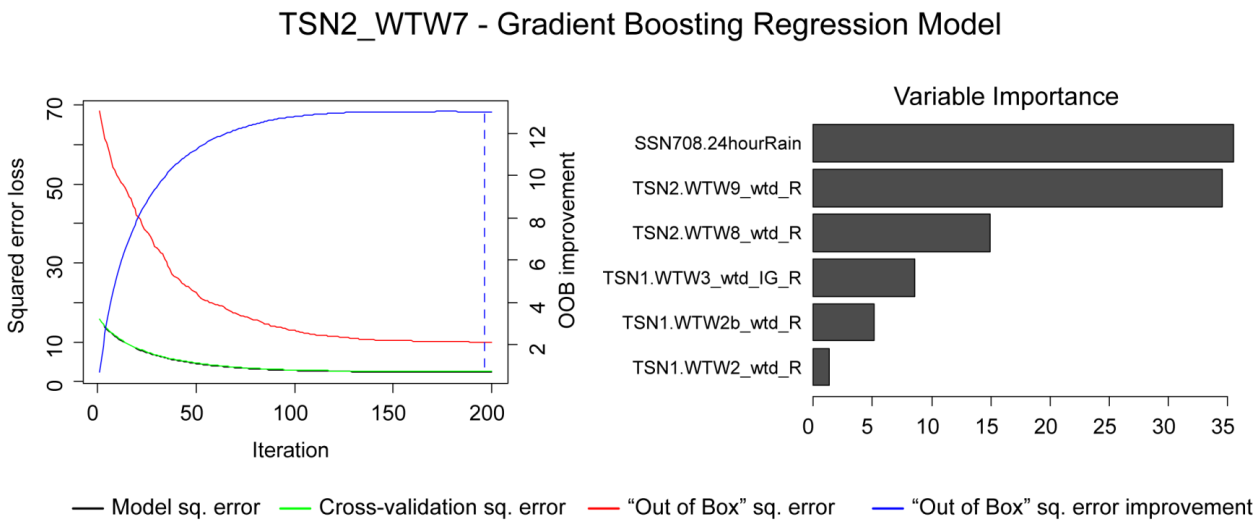

Figure 4. Gradient boosting regression model details

GBM is based on iterative reduction of error in either an out-of-box (OOB) sample (red line) or reducing the error in a cross-validation dataset (green line). In this model the OOB error reduction method was chosen. This GBM model ran for 200 iterations with 5-fold cross validation. Rain at watershed 708 and water table depths at the other wells in TSN 2 were the most important predictive variables, although these data were not always available.

Conclusion

Further investigation is required for methods for classification of sensor readings as being valid or invalid. Such methods could compare gradient boosting model predictions with actual data, and determine if observations are significantly different than expected. The results presented here are for a subset of methods, variables and available data, and therefore may not represent the most effective method of imputation. Any imputation method would require extensive testing and validation before use. However, these data indicate that GBM can accurately impute sensor network data. If GBM is selected as an imputation method of interest, the imputation function, currently written in R, can be re-written (e.g., in Python) and incorporated into data-handling pathways.