By fitting numerous decision trees to the data and taking a consensus, Random Forests often perform well with noisy data REFERENCE. Previous studies have shown that extensive parameter tuning is often unnecessary to obtain good results REFERENCE. Indeed, the parameters which can be meaningfully tuned are limited to the ‘mtry’, ‘ nodesize’ and arguably ‘ntree’. These parameters represent the number of variables selected for each tree, the number of instances in the smallest node before terminating tree generation and the number of trees generated respectively REFERENCE. For the yarn dataset, the default values for these parameters are 15, 5 and 500 REFERENCE.

The randomForest package was used to train Random Forest regression models on the dataset and tune the parameters. The mtry and nodesize parameters were tuned on the following grid:

Mtry: XX

Nodesize: XX

Experimentation with a subset of the data indicated that these parameter ranges gave the best results. Models were trained and evaluated on the full dataset using EC2, which required approximately 3 hours of computation time.

INSERT SURFACE PLOT OF PARAMETER TUNING PERFORMANCE