A Random Forest for Predicting
Human Puumala Orthohantavirus Infections in North-Western Germany

Orestis Kazasidis, Joanna Dürger, Christian Imholt, Jens Jacob

Julius Kühn Institute (JKI), Institute for Plant Protection in Horticulture and Forests,
Toppheideweg 88, 48161 Münster, Germany

orestis.kazasidis@julius-kuehn.de

Human Puumala orthohantavirus (PUUV) infections in Germany fluctuate regularly in close correlation with the abundance of bank voles, which are the only PUUV-reservoir species. It has already been demonstrated that these fluctuations correlate with beech mast intensity of the previous year and weather parameters, which determine the food availability and affect the growth of the bank vole population. Such correlations have been used in the past to develop predictive models for the human PUUV-infections at the district level based on regression tree analysis.

A single regression tree can be both a transparent and a powerful tool for descriptive and inspection purposes. Nevertheless, it is fundamentally restricted when used for predictions, because it is prone to overfitting. Its greedy univariate nature can undermine predictors, whose importance may be overshadowed by more dominant ones. Furthermore, it exclusively produces axis-aligned hyper-rectangles in the predictor space, limiting the shape and the direction of the feasible partitioning. A tree ensemble relaxes many of the restrictions of a single regression tree, by combining the predictions of multiple trees. The two most popular methods are gradient boosted trees and random forests.

This contribution demonstrates the potential of a random forest for the prediction of human PUUV-infections. We selected the random forest over the gradient boosted trees algorithm, because the latter is biased to the training data and thus more prone to overfitting. The stochastic nature of a random forest allows out-of-sample prediction, especially relevant for datasets with different distributions of the predictors' values.

Our test case involved districts of the north-western part of Germany, specifically of the Federal States of Lower Saxony and North Rhine-Westphalia. We describe the selection of predictors from an extensive initial pool including multiple weather parameters, as well as tree phenology and land use. We report the increase in accuracy of a random forest regressor over a single regression tree with the same predictors. Finally, we apply the model for predicting human PUUV-infections for 2022.

This study was supported by the Federal Ministry of Education and Research (BMBF) for the “RoBoPub” consortium (grant number 01KI1721E), and the German Environment Agency (UBA) within the departmental research plan of the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (research code 3720 48 401 0).