

yaImpute: An R Package for k -NN Imputation

Nicholas L. Crookston
Rocky Mountain Research Station
USDA Forest Service
ncrookston@fs.fed.us

Andrew Finley
Department of Forest Resources
University of Minnesota
afinley@stat.umn.edu

Abstract: We present a new R package to support nearest neighbor (NN) imputation. The package provides functions for finding neighbors, accomplishing the imputations, and providing summaries and graphs of results. Five different methods for finding neighbors are supported, four of which include the major step of finding the k minimum sum-of-squared differences in common attributes between a target observation and the possible reference observations. All methods seek k -NNs, where nearness is measured by a distance. Measures of common attributes are called X 's. With method *Euclidean*, distance is computed in a normalized X space, with method *Mahalanobas* distance is computed in its namesakes space, with methods *msn* and *msn2*, distance is computed in projected canonical spaces, and with method *gmn* distance is computed using a projected ordination of X 's found using canonical correspondence analysis. In the last method, *randomForest*, observations are considered similar if they tend to end up in the same terminal nodes in a suitably constructed collection of classification and regression trees. The distance measure is one minus the proportion of trees where a target observation is in the same terminal node as a reference observation. Similarly to the other methods, k -NNs are the k minimum of these distances.

The first four methods are well defined in the literature—for these methods we herein present short definitions and citations. For method *randomForest* a more detailed explanation is included. Notable advantages of method *randomForest* are first that it is non-parametric and second that the attributes can be a mixture of continuous and categorical variables. The other methods require continuous measures where categorical variables transformed to some continuous space.

In all methods, finding the exact minimums involves time consuming searches between every target and all the reference observations. The package includes a much faster approximate nearest neighbor (ANN) search for use when appropriate and when the method includes a sum-of-squared difference as the distance measure, which is not the case for method *randomForest*.

Applications of k -NN include a data preparation step, a modeling and evaluation step, and lastly a production-oriented step for finding neighbors for many targets and making the imputations. The yaImpute package provides functions and options that support these individual tasks. An example is given illustrating the principal.

Our future plans for this package include providing additional methods as we discover them, adding new logic of computing imputed values when $k > 1$, providing variance estimators, and building linkages to mapping packages.