ArboMAP: Arbovirus Modeling and Prediction to Forecast Mosquito-Borne Disease Outbreaks

Variable Selection (v2.0)

Justin K. Davis and Michael C. Wimberly
(justinkdavis@ou.edu, mcwimberly@ou.edu)
Geography and Environmental Sustainability, University of Oklahoma

Updated January 16, 2019

Variable selection has been performed, with environmental variables chosen pairwise from the available environmental covariates. There were 15 models considered, and these are ordered below by the Akaike information criterion (AIC). A lower AIC in this case means the model fit the observed data better, so we generally assume that the first model on the list is the best model, and use these two variables for modeling and prediction.

The "area under the curve" (AUC) which is also known as the c-statistic, is also calculated for all models. This measure ranges from 0, indicating a model that is never able to correctly discriminate between positive and negative district-weeks, and 1, whenever the model is able to discriminate perfectly. Think of it roughly as an R2 for logistic regression models. It is calculated on all available data and may be artificially high, since it is usually easy, for example, to predict that there will be no human disease in cold months. This metric should probably correlate with the AIC, but should probably not be used for model selection. You should hope to see values above 0.60, and typically we would want to see models with 0.70 or higher.

Here, therefore, we recommend tmeanc and vpd as predictors in the `ArboMAP Main Code.Rmd` file.

var1	var2	AIC	AUC
tmeanc	vpd	6395.11	0.95
tmaxc	vpd	6406.59	0.95
tmeanc	rmean	6424.50	0.95
tminc	vpd	6428.23	0.95
tmaxc	rmean	6429.22	0.95
tminc	rmean	6447.87	0.95
tminc	pr	6504.59	0.95
tminc	tmeanc	6511.48	0.95
tminc	tmaxc	6511.48	0.95
tmeanc	tmaxc	6511.48	0.95
tmeanc	pr	6534.88	0.94
tmaxc	pr	6569.40	0.94
rmean	vpd	6617.04	0.94
pr	vpd	6697.74	0.94
pr	rmean	7164.97	0.93