Proposal to BORG for implementing Cross-Validation of 4-D estimation GAMs.

Rebecca and Elgin
rmurphy@chesapeakebay.net
eperry@chesapeake.net
5/27/2022

Justification:

Up to this point the assessment of 4D model improvement has been based on standard tools such as r^2, AIC, F-values, and p-values. Some problems with these measures are becoming apparent. Because the trial data set is relatively large, ~ 120,000 observations, a prediction variable that makes a very small contribution to r^2, can have large F-values and small p-values. The AIC criterion has values in the hundreds of thousands and consequently the usual rule of thumb that a candidate prediction variable is important if it reduces AIC by four or more does not seem to apply. While AIC and F-values seem volatile, r^2 is almost immovable. We are concerned that a candidate prediction variable might lead to large improvement in a small region of the prediction space, but this will not be reflected in r^2. For these reasons we recommend implementing cross-validation for evaluating model improvement.

Concepts:

Cross-validation is a process of estimating parameters for a model using one set of data and then evaluating the goodness of fit of that model using an independent set of data. The data used for estimating parameters is call the training data. The data used for assessing goodness of fit is called the testing data. The time period used for our current trial data set is all DO data between 1990-2010 for selected stations in the central Bay. A simple strategy for cross-validation would be to use 1990-2010 as training data and a later period, say 2011-2015, as a testing data set. But this seems unsatisfactory for our purposes. Our goal is to assess the ability of our model to interpolate in space and time whereas this simple approach would address the ability of the model to forecast.

To assess interpolation, we propose segregating the data in the 1990-2010 period into training and testing data. We are currently thinking of doing this cross validation in two steps that address interpolation in time and space separately. Time interpolation will be tested by randomly splitting the observations from months with two observations between training data and test data. All observations that are one per month will be retained in the training data. This will result in some bias toward spring and summer observations in the testing data, but because these are the periods most likely to have DO violations, this is appropriate. In a separate cross-validation exercise, space interpolation will be tested by completely removing selected stations from the training data and reserving these for testing data. Because we have many fewer stations than dates, we may implement a jackknife procedure or repeated random selection of training data. Another source for testing data that is spatially independent of the training data is the validation observations made during data flow cruises where DO profiles are collected. Some research will be required to determine the degree of overlap in time space between the fixed station network and the data flow cruises.

Related Analyses

Commented [HJ1]: This seems like a real concern, along with the idea that a small but potentially important or meaningful region of the prediction space could have a poor fit, but this is masked by the agreement in so many other places.

This bears some resemblance to comparisons between simulation models and observations that I've seen. For example: a simulation model that predicts DO very well whenever DO is > 2 mg/L, but when DO is low, the model is less accurate. If DO is your endpoint of concern, this kind of lack of fit is unfortunate.

Commented [HJ2]: When you repeatedly select training data, then you are also repeatedly selecting data to be used fitting the model. This can provide a range of fitted values, so you can identify regions where the model provides higher-variability of predictions.

But, GAMs can already give confidence limits for prediction … I think. So, are these the same. IF they are different, what's the difference?

Commented [HJ3]: This makes sense to me. As a logical next step, it might be interesting to cross validate attainment of DO threshold or not. So for example, which variables provide the most improvement in correctly classifying observations as meeting or not meeting the threshold?

As part of this cross-validation process, we hope to develop methods for identifying regions in space and time where our model is performing poorly. This analysis may give some insights on how to improve the model. Failing that, it will give us insights on the challenges we will face in developing the simulation component of the full criteria assessment model.

Conclusion:

With this short proposal, we are inviting comments from BORG on whether cross validation should be our next priority and comments on the concepts and mechanics for implementation of cross validation.

Commented [HJ4]: This is important … is it possible to map cross validation error.

Commented [HJ5]: In summary, a major concern that I have about fitting models as an approach to interpolation is that the model can have stellar global fit statistics, but the fit to certain regions of the prediction space could be bad. Knowing where these regions are, and if certain model improvements could be LOCALLY important is useful. Also, if its possible to weigh errors are more or less important for certain policy objectives, is it possible to provide a weighted score that prioritizes improvements that matter most … such as correctly classifying DO as meeting or not meeting the seasonal and spatial-specified thresholds?

It seems like this problem is not different for a GAMS interpolation vs. performance evaluation for a mechanistic water quality simulation model.