New Statistical Techniques for Predictive Water Quality Modeling at Great Lakes Beaches

Michael N. Fienen¹
Wesley R. Brooks¹
Steven R. Corsi¹
Kurt Wolfe²
Rajbir Parmar²
Mike Galvin²
Mike Cyterski²

¹USGS Wisconsin Water Science Center, Middleton, WI
²USEPA Office of Research and Development, Athens, GA


**Introduction**

*E. coli* is used as a fecal indicator bacterium

Without rapid methods, 24 hours are required to get an answer.

Environmental predictors (weather, waves, etc.) are available in real time.

Our ultimate goal is to predict *E. coli* from environmental predictors.
Modeling Approaches

*getting the most information from all data available*

**persistence model**
- using yesterday’s *E. coli* to predict today’s conditions

**ordinary least squares regression (OLS)**
- well-established technique
- requires decisions on how to handle interaction and correlation

**partial least squares (PLS)**
- newer technique used extensively in spectroscopy
- introduced to beach community by Hou, Rabinovici, and Boehm (2006)
- useful when variables are correlated or insensitive
- overfitting prevented through cross validation and component trimming
- algorithm replaces trial-and-error interaction terms & variable selection
Partial Least Squares: Example

Dimension 2

Dimension 1

Prediction Sum of Squared Errors

number of components used

USGS
Partial Least Squares

regression built on principal directions relating variables to predictions

\[ y = X\beta + \varepsilon \quad \text{error term} \]

principal directions of covariance between \( X \) and \( y \) define components

using all components is equivalent to OLS with all variables

the number of components retained is chosen balancing lower PRESS against overfitting

each component includes information from all base variables
Decision and Regulatory Thresholds

log $E. \text{coli}$

Time

regulatory threshold

observed
modeled
Tradeoffs for Model Performance

robust methods and improved data integration

building a PLS model

managers decide the tradeoff between protective and permissive

sensitivity: proportion of true positives
high sensitivity means increasing true positives

specificity: proportion of true negatives
low specificity means increasing false positives

specificity limit (vertical blue line) is the dial controlling this tradeoff
Great Lakes Beaches Modeled

Wisconsin Department of Natural Resources

USGS Ohio Water Science Center
Building a PLS model – example for Edgewater, Ohio

2005-2010: split training data

Build and test models: predict each fold using data from the other four

2011: prediction year

divide the training data randomly into five equally-sized “folds”
select a few candidate specificity limits (tuning)

compare model performance on the test folds
pick the specificity limit that had the best performance, and train a new model over all five folds

make and record predictions to manage beach and provide data going forward
2010 Model Performance

Exceedance proportion *correct*

Exceedance proportion *incorrect*
Next Steps

Data acquisition and connectedness
Greater data acquisition efficiency and model-building efficiency leads to rapid application to many beaches
Integration with USEPA and Virtual Beach

beach managers on-site measurements

GLRI beach data effort

Beach Modeling Framework

GeoData Portal

Virtual Beach

USGS
Next Steps

Virtual Beach

Well-known EPA software for predicting bacterial concentration
PLS regression to be included in version 3 (coming 2012)
Acknowledgements

**OLS model results and collaboration:**
Rob Darner  
Donna Francy  
Amie Brady  
Dan Ziegler  
Adam Mednick

**Data provided by:**
Cuyahoga County Board of Health  
Northeast Ohio Regional Sewer District  
Ozaukee County, Wisconsin

**Funding provided by:**
Ocean Research Priorities Plan (ORPP)  
USEPA Great Lakes Restoration Initiative (GLRI)