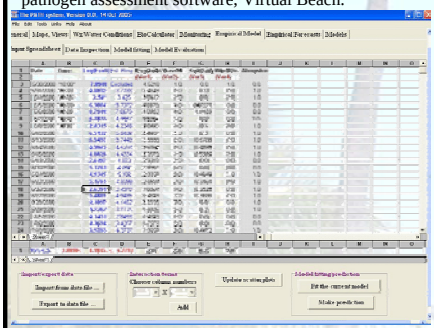
Development of the Virtual Beach Model, Phase 1: An Empirical Model

Zhongfu Ge[1], Walter E. Frick[2]: [1]National Research Council; [2]US EPA, NERL, ERD, Athens, GA

Abstract

With increasing attention focused on the use of multiple linear regression (MLR) modeling of beach fecal bacteria concentration, the validity of the entire statistical process should be carefully evaluated to assure satisfactory predictions. This work aims to identify pitfalls and misunderstandings of the statistical aspect of modeling. The importance of preliminary inspection of raw data, useful transformations, development of interaction terms, adjustment for time-series effects, identification of outliers, correlation studies, and model selection criteria are stressed. It is recommended the model selection process should be conducted using R^2 and Cp statistic as joint criteria. The methodology is illustrated with actual data from Huntington Beach, OH, in 2000-2004. Dynamic modeling, as a new concept, is advanced for prediction purposes, as beach bacteria MLR models are in fact beach specific and time varying. This work also serves as a statistical basis for US EPA's public domain pathogen assessment software, Virtual Beach.



Example Virtual Beach input screen. Rows and columns can be highlighted by the user to initiate actions, such as omitting a variable (last column), or by the program, for example, to show an identified outlier case (first row).

Objectives

- To demonstrate multiple linear regression modeling of *E coli* concentrations
- To clarify some misunderstandings and pitfalls of MLR modeling found in practice
- To promote the idea of dynamic modeling based on a growing data-base

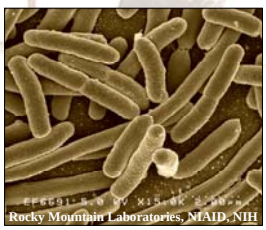
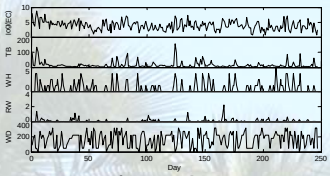
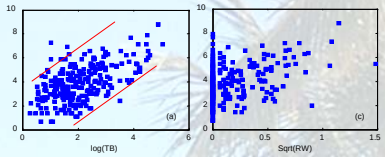


Image of E coli

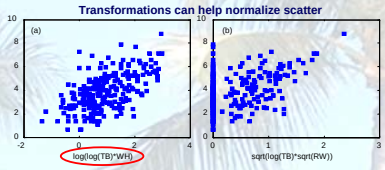
Example of modeling for Huntington Beach, Ohio, a Lake Erie beach



Data come from 247 days in 2000-2004. Four explanatory variables are available: turbidity (TB), wave height (WH), antecedent 24-hour rainfall (RW), and wind direction (WD). Cross-correlation with time delays shows that the data do not need to be synchronized. (USGS data)

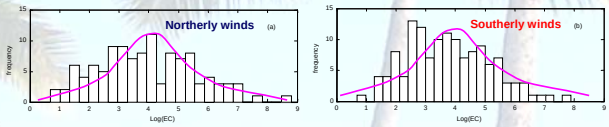


Inspection of the scatter plots shows that transformation will help equalize variances over the range of values. Here are two distributions after transformation.



Including interaction terms can substantially improve model performance.

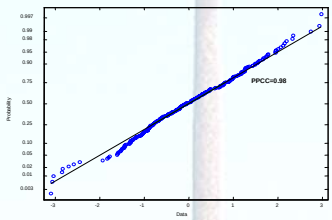
Wind direction was categorized into northerly (WD=0) and southerly (WD=1) winds. This change produces data histograms that exhibit equal spreads.



Other data inspection issues include multicollinearity. This table shows variance inflation factors (VIF) for each explanatory variable

	TB	WH	RW	WD	TB·WH	TB·RW	WH·RW	VIF
TB	1.00							8.710
WH	0.616	1.00						7.381
RW	0.188	0.230	1.00					8.278
WD	-0.237	-0.279	0.142	1.00				1.162
TB·WH	0.893	0.849	0.246	-0.252	1.00			17.45
TB·RW	0.377	0.305	0.915	0.074	0.420	1.00		28.43
WH·RW	0.323	0.456	0.911	0.060	0.431	0.968	1.00	31.51

Correlation coefficients

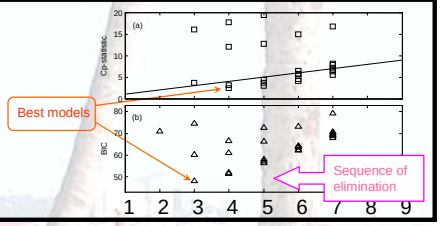


The full model before backwards elimination is given here. The graphic depicts the normality of the residuals.

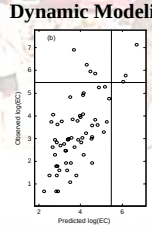
$$\mu\{\log(EC)\} = \beta_0 + \beta_1 TB + \beta_2 WH + \beta_3 RW + \beta_4 WD + \beta_5 TB \cdot WH + \beta_6 TB \cdot RW + \beta_7 WH \cdot RW$$

Model Selection

After outlier identification (not shown), model selection is accomplished using backward elimination, a process facilitated by Virtual Beach. Starting with the full model, elimination of the poorest predictor leads to variable Cp and BIC. (BIC = Bayesian Information Criterion.) The minima in these statistics help to identify the best models.



Dynamic Modeling



In the proposed dynamic modeling approach, models are updated as new observations are added to the data base. This graphic shows predictions for 60 days. (Note: measurements are not available for every day.)

Variable	Days #1 ~ #20	Day #21	Day #22	Days #23 ~ #60
TB		★		★
WH				
RW	★	★	★	★
WD				
TB·WH	★	★	★	★
TB·RW				
WH·RW				

★ The variable is in the model; R^2 is consistently around 48%

The table shows how the best models change with time (additional data). Although the variables that produce the best models change with time, the R^2 value remains steady, around 48%.

Conclusions

- Model selection should be based on Cp and R^2 as criteria; R^2 or t-statistic alone are found inadequate
- Transformations tend to improve results
- Interaction terms can improve model R^2 and are useful especially when variables are limited (herein 48% compared to 41% without interactions).
- Optimal models are both beach-specific and time-varying
- The idea of dynamic modeling based on a growing data-base is recommended

Acknowledgements: We thank Donna S. Francy and Robert A. Darner for their technical suggestions and guidance.

References:

Francy, D.S., Gifford, A.M., and Darner, R.A., 2003, Escherichia coli at Ohio bathing beaches—distribution, sources, wastewater indicators, and predictive modeling: Water Resources Investigations Report 02-4285, 120 p., accessed May 2006 at http://oh.water.usgs.gov/beaches/index.html.

Nevers, M.B. and R.I. Whitman 2004. Protecting visitor health in beach waters of Lake Michigan: problems and opportunities. The State of Lake Michigan. Ecology, Health and Management, Eds. T. Edsall & M. Munawar. Ecovision World Monograph Series, 2004 Aquatic Ecosystems Health and Management Society

Ramsey, F.L and D.W. Schafer 2002. The statistical sleuth: a course in methods of data analyssi, second edition. Duxbury Thomson Learning