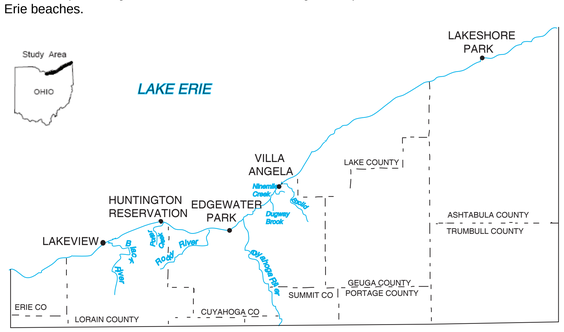
Using Models To Predict Recreational Water Quality At Coastal Beaches

Donna S. Francy[1], Robert A. Darner[1], Erin E. Berke[1], and Jill Lis[2]

[1]U.S. Geological Survey (USGS) Ohio Water Science Center, Columbus, OH
[2]Cuyahoga County Board of Health, Parma, OH

Predictive models have been recognized as tools to improve assessments of recreational water quality at coastal beaches. Multiple linear regression (MLR) models use easily measured environmental and water-quality factors to estimate bacterial indicator concentrations or the probability of exceeding target concentrations. In Ohio, the USGS is working with local and state agencies to develop and test models to predict *E. coli* concentrations at five Lake Erie beaches—Lakeview (Lorain), Huntington Reservation (Bay Village), Edgewater Park (Cleveland), Villa Angela (Cleveland), and Lakeshore Park (Ashtabula). For Huntington, where investigations are further along than at other beaches, 6 years of data (2000–2005) were used to compare the model's performance to performance of the currently used method for assessing recreational water quality.

Models for predicting *E. coli* concentrations are being developed for five Lake Erie beaches.



Predictive models are beach specific and can be developed by following the steps listed below.

1. Collect daily data on *E. coli* concentrations and environmental and water-quality variables. Samples were collected at 1 m depth in the swimming area on weekdays between 8 a.m. and noon. Samples were collected at two or three locations at each beach, and resulting concentrations were averaged for analysis. Samples were analyzed for *E. coli* by means of the mTEC or modified mTEC membrane-filtration methods (USEPA, 2000 and 2002). Associated environmental and water-quality data were collected by field crews or compiled from a variety of sources. Because a high-quality dataset is important for model development and testing, strict quality-assurance and quality-control procedures were followed.

2. Conduct exploratory data analysis and compute statistics to determine relations between *E. coli* concentrations and possible explanatory variables. Correlation coefficients were calculated and scatterplots constructed to determine the strength of the association between *E. coli* and possible explanatory variables. Box plots, analysis of variance, and Tukey's multiple comparison test (Helsel and Hirsch, 1992) were used to examine the distribution of *E. coli* concentrations as a function of variables grouped by categories.

Considerations for selecting explanatory variables. Which explanatory variables had strong associations with *E. coli*? Was the relation consistent from year to year? Was the relation influenced by one or two extreme values, and if so, were those extreme values valid points? Is the relation between the explanatory variable and *E. coli* linear? If not, can the data be transformed to yield a linear relation? Were two explanatory variables strongly related to each other? How difficult or costly is it to obtain the data? Does use of the explanatory variable make sense?

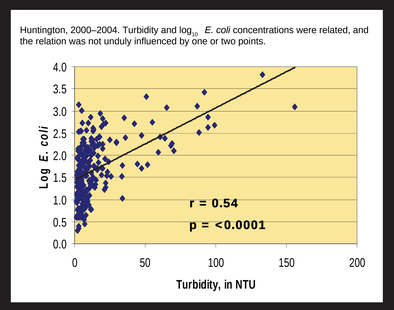
Huntington 2000–2004. Pearson's correlation (*r*) between $\log_{10}$ *Escherichia coli* concentrations in water and explanatory variables were examined to help select variables for inclusion in model development. [Relations that were significant at $\alpha = 0.05$ are in red.]

Factor	2000	2001	2002	2003	2004	Combined
Birds, number on beach at time of sampling	-0.28	0.12	-0.26	-0.16	-0.09	-0.10
Julian Day	-0.38	0.36	0.36	0.02	0.13	0.09
Rainfall last 24 hours (Rain 24)	0.47	0.28	0.30	0.36	0.50	0.34
Rainfall 2 days ago	0.28	-0.05	0.02	0.20	0.27	0.20
Rainfall 3 days ago	0.03	-0.09	-0.21	0.10	-0.09	0.08
Rainfall weighted 48 hours (Rainw48)	0.50	0.24	0.28	0.38	0.55	0.37
Turbidity	0.60	0.67	0.49	0.35	0.51	0.51
Lake level	-0.12	-0.16	-0.30	0.05	0.09	-0.12

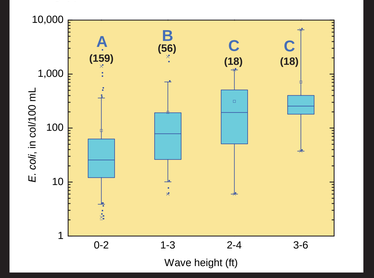
Other beaches, 2004 and 2005 combined. Pearson's correlation (*r*) between $\log_{10}$ *Escherichia coli* concentrations and explanatory variables indicate that at different beaches, explanatory variables differ in their level of association. [Relations that were significant at $\alpha = 0.05$ are in red.]

Factor	Lakeview	Edgewater	Villa Angela	Lakeshore
Birds, number at time of sampling	0.33	0.15	0.21	-0.18
Julian Day	0.43	0.15	-0.13	0.26
Airport rainfall last 24 hours	0.29	0.30	0.36	0.35
Airport rainfall 2 days ago	-0.02	0.22	0.09	0.20
Airport rainfall 3 days ago	-0.04	0.10	0.09	0.31
Airport rainfall weighted 48 hours	0.24	0.37	0.37	0.40
Local Rainfall last 24 hours	0.22	--	--	--
Local Rainfall 2 days ago	0.30	--	--	--
Local Rainfall 3 days ago	0.05	--	--	--
Turbidity	0.18	0.40	0.38	0.37
Lake level	-0.30	-0.25	-0.16	-0.12
Water temperature	0.49	0.09	0.34	0.36

Huntington, 2000–2004. Turbidity and $\log_{10}$ *E. coli* concentrations were related, and the relation was not unduly influenced by one or two points.



Huntington, 2000–2004. *E. coli* concentrations increase with increasing wave heights. Wind direction (another categorical variable) was not related to *E. coli* (data not shown). [Results of Tukey's test are shown as letters, and median concentrations with different letters were significantly different. Numbers in parenthesis are the number of samples in each category.]



3. Select explanatory variables and generate a list of models. The variables selected were used to generate a list of possible MLR models ordered to minimize the Mallow's Cp (Mallows, 1973) and maximize the R^2 statistic. The Cp statistic is a measure of the error variance and bias introduced by not including important variables in a model. The R^2 is the fraction of the variation in *E. coli* that can be explained by a combination of explanatory variables.

Huntington, 2000–2004. A list of models was generated using Mallow's Cp. Because Rain24 and Rainw48 were strongly related, models that included both variables were not considered. The highlighted model was selected for further testing.

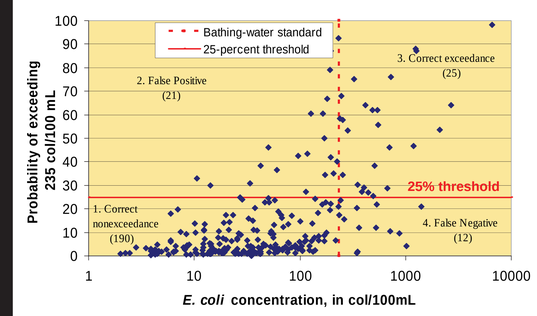
Number in Model	C(p)	R-Square	Adjusted R-Square	Variables in Model
4	4.4524	0.4809	0.4723	WaveHt, Date, Rainw48, logturb
4	7.7707	0.4738	0.4651	WaveHt, Rain24, Date, logturb
3	11.1264	0.4623	0.4557	Date, Rainw48, logturb
3	12.7050	0.4589	0.4523	Rain24, Date, logturb
4	23.6891	0.4397	0.4305	Turbidity, WaveHt, Date, Rainw48
3	26.0062	0.4304	0.4234	WaveHt, Rainw48, logturb
4	27.0380	0.4325	0.4232	WaveHt, Rain24, Rainw48, logturb
4	27.6786	0.4311	0.4218	Turbidity, WaveHt, Rain24, Date
3	28.2167	0.4257	0.4187	WaveHt, Rain24, logturb
3	36.0452	0.409	0.4017	Turbidity, WaveHt, Rainw48

4. Assess the models. Hypothesis tests and regression diagnostics were done to confirm that the model provided the best linear unbiased estimator of *E. coli* concentrations (Helsel and Hirsch, 1992). These included analysis of partial residual plots, plots of actual versus predicted concentrations, plots of residuals versus predicted concentrations or explanatory variables, and examination of outliers.

Assessing the model: Are the parameter estimates reasonable and significant? Do the partial residual plots indicate that each variable influences the regression? Are any of the outliers due to data errors? Are the residuals evenly distributed above and below the zero line? Is there a visual pattern to the residuals?

5. Generate model output values, determine the threshold probability, and validate the model. In earlier studies, prediction intervals were shown to be too wide to estimate *E. coli* concentrations with reasonable confidence (Francy and Darner, 1998; Francy and others, 2003). Consequently, the probability of exceeding the single-sample bathing-water standard of 235 col/100 mL was used as the model output variable. A threshold probability associated with too great a risk of exceeding the bathing-water standard was determined by maximizing the number of correct responses while minimizing the number of false negative responses. Data were collected during an independent year (2005) to compare the model's performance with the current method for assessing water-quality (using yesterday's *E. coli* concentration). The model resulted in increased sensitivity and specificity and a greater percentage of predicted correct responses than using yesterday's *E. coli* concentration.

Huntington 2000-2004. A threshold probability of 25% was chosen. Estimated probabilities less than the threshold indicate that bacterial water quality is most likely acceptable for swimming. Estimated probabilities equal to or above the threshold indicate that a water-quality advisory should be issued.



Huntington, 2005. Comparison of the model's performance (WaveHt, Rainw48, logturb in table 4) with the current method (yesterday's *E. coli* concentration) for assessing recreational water-quality.

	Percent of predictions based on	
	Model	Yesterday's E. coli
Sensitivity [a]	50.0	0
Specificity [b]	90.2	84.6
Correct	84.7	74.6
False + [c]	55.5	100
False - [d]	8.0	13.7

[a] Exceedances that were correctly predicted.
[b] Nonexceedances that were correctly predicted.
[c] Predicted exceedances observed as nonexceedances.
[d] Predicted nonexceedances observed as exceedances.

Conclusions and next steps

- MLR techniques can be used to generate beach-specific models for coastal recreational waters.
- A model for Huntington based on rainfall, wave height, and turbidity performed better than using yesterday's *E. coli* during a validation period.
- An Internet-based "nowcasting" system will be developed for Huntington in 2006 to inform the public of water-quality conditions based on predictive models.
- For the other beaches, the models developed with data from 2004–05 will be validated in 2006.