Regional and local factors affecting patterns of *E. coli* distribution in southern Lake Michigan

Submitted by

Richard Whitman
Meredith Nevers

To

Great Lakes National Program Office, EPA
David Rockwell, Technical Representative

Department of Environment, City of Chicago
Renante Marante, Project Manager

In cooperation with

Sharyl Rabinovici, USGS
Jim Meyer, City of Gary
David Schwab, Great Lakes Environmental Research Laboratory, NOAA
Renante Marante, Naren Prasad, City of Chicago
Julie Kinzelman, Robert Bagley, City of Racine
Mark Pfister, Lake County, IL
Mary Ellen Bruesch, Stephen Gradus, City of Milwaukee
Scott Hicks, Cheryl Burdett, National Park Service
Alex DaSilva, Indiana Dept. of Environment

Funding for the project was supplied in part by US EPA grant # GL-98500001

Table of Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

Predictive modeling has been suggested as an alternative to the current approach of monitoring recreational waters for fecal indicator bacteria. The traditional technique for monitoring *E. coli* is based on laboratory analysis of representative samples and involves a considerable amount of time for obtaining results. It has been suggested that predicting bacteria counts using statistical models may be a suitable alternative because results are timelier, less subject to local variation, and more explanatory. Presently, model development for beach management purposes has not been well explored. Relatively few beaches have been modeled and the results have varied widely; of these trials, few models have been conservatively validated. While some have been cross-validated using modern recursive techniques (e.g., jackknife, PRESS) on inherently biased subsets from the same collected data period, there has been a notable absence of validation using autonomously collected data sets. Autocorrelation, especially in time series analysis, conservatively requires that model validation is performed during a different time period (in our case, swimming season) while being consistent in all other sampling approaches. Another inadequacy of most models is that they are limited spatially which results in limited explanation of coastal variation, effects of fix factors (orientation, development, source contaminants), and how background levels vary among themselves. The general assumption is that beaches are essentially unique and subject to mostly local influences and pollution effects; this undermines a wider view that the lake is a complex system that interacts with local beaches in similar ways and that once accounted for, local and pollution effects can be more efficiently explained. Nonetheless, it is intuitive that traditional modeling will be more successful as one restricts space and time, but this is done at the expense of a more generalized and explanatory view of regional waters.

Restricting models to a local scale has major scientific and managerial disadvantages. First, it is well known that general seasonal, weather, and hydrological conditions greatly influence the physical, chemical, and biological characteristics of large water bodies such as the Great Lakes. Such factors may in turn affect the occurrence, distribution, and survival of microbiological contaminants in the water. How these hydrometeorological and biological conditions may affect the densities of indicator bacteria is poorly studied, aside from the obvious forcing of contaminated water by local currents. Regional models may provide clues as to how these factors interact and relate to beach contamination. Second, by restricting observations to only one beach, little can be said of adjoining waters and potential sources. One can see the effects but not the overall influence of outside factors, so they appear to be indeterminate forces acting on the local model. Since beach designations and their respective management are largely defined by political jurisdiction, selection of spatial scale and boundaries is essentially arbitrary.

This study reviews four years of data (2000-2003) from 55 beaches along southern Lake Michigan from Milwaukee, Wisconsin through Michigan City, Indiana. The dataset included 10,422 observations, with an overall mean log *E. coli* of 1.72 (SD=0.76), or 51.5 CFU/100 ml. Multi-dimensional spatial analysis showed that these beaches fell into seven relatively homogeneous groupings: Milwaukee and Racine, Wisconsin; Lake County, north Chicago and south Chicago, Illinois; and eastern and western Indiana. On the broad scale, beaches were generally correlated with one another. Most beaches were not significantly different in overall

mean *E. coli* densities; however, 63rd Street Beach, IL and South Shore Beach, WI were notable exceptions. At the local scale there was spatial correlation with beaches close to one another behaving similarly. A trend towards increasing *E. coli* density over the course of the summer sampling season was evident.

During the course of the study, 79 hydrometeorological factors were collected and screened to ascertain how these factors influence *E. coli* density at the beaches. The final candidate list of potential predictors included averages for: minimum temperature, rainfall, wave height, wave period, wind speed, wind direction, barometric pressure, lake stage at Calumet Harbor (measured inside the harbor mouth), and *E. coli* from the previous day. When all data were pooled, these prediction coefficients were all highly significant as was the overall model using multiple linear regressions. The overall, R^2 was 0.29; the R^2 using yesterday's *E. coli* reading (the presently accepted model) was only 0.19. Clustering beaches increased the R^2 to 0.32, but regression on individual zones did not improve the model variance explanation. Modeling of Indiana had limited success because of the low sampling intensity. Overall, the model accurately predicted whether the mean bacterial levels were over or under EPA criteria (235 colony forming units/100 ml) 78% of the time; analysis of data for individual zones did not increase the predictability of closures.

Regression tree statistics suggested that 3-day prior moving average *E. coli* density was the factor best separating present *E. coli* into subgroups. For the lower *E. coli* subgroup, rainfall was an important classification factor, while for higher levels of *E. coli* wave height best separated *E. coli* levels. Partitioning the data into zones or N-S wind vectors notably changed the classification tree structure. Analysis of groups as determined by hierarchichal clustering resulted in more significant regression models than individual beaches. Nonetheless, regression confidence intervals were wide.

Chicago's 63rd Street Beach was used as an example modeling exercise for a single location because there were several sets of independently collected data. *E. coli* densities from a USGS study conducted in 2000 were compared to concurrent data collected by the Chicago Park District. Morning *E. coli* was compared to afternoon data and knee-deep was compared to waist-deep collections. The R^2 was higher for afternoon samples taken at waist deep water, and knee-deep morning samples had the lowest coefficients of determinations. We conclude that weather and lake conditions have only marginal effect on *E. coli* readings at 0700 h collections, but deeper water integrates and smoothes sampling variation and allows predominant coastal conditions to exert an influence. We suggest that serious consideration be given to sampling in deeper water late in the morning or mid-day.

The performance of our derived predictive model was evaluated against the present monitoring approach (culturing water samples for *E. coli*) for all zones and years using Root Mean Square Error (RMSE). The RMSE behaves similarly to the standard deviation but expresses the variation between the predicted and the measured *E. coli.* Thus, a perfect model would have an RMSE of 0 (model performance) and an R^2 (model explanation) of 1. The model developed in this study performed better than the recommended state and EPA 'model' in all four years evaluated. The RMSE for the developed model was 14% lower for all pooled beaches from Illinois and Wisconsin than the currently recommended approach. This improvement must be

weighed against the increased cost and effort of collecting, assembling, and applying the hydrometeorological predictive factors. From a management perspective the use of a regional model is unlikely to improve predictability efficiently on a routine basis. The strength of these exercises was to differentiate local from general effects and fixed from random factors, describe the variation and relative concentrations of indicator bacteria, and develop an understanding of background levels of *E. coli* within the regions. The exercise goes far in explaining why beaches tend to vary together, illustrates the difference in the character of local regions, and challenges the sampling practices currently used.

INTRODUCTION

Beaches along southern Lake Michigan are closed to swimming when there are high concentrations of *E. coli* bacteria in the lake. To protect public health, several jurisdictions in this area have been monitoring their beaches for *E. coli*, for as long as 20 years, and closing them to swimming when counts exceed the US EPA recommended limit of 235 CFU/100 ml. Beaches are differently affected by this policy with some being closed frequently during the swim season and others rarely being closed. Factors that affect *E. coli* counts are varied and numerous, and patterns of beach closures have been difficult to perceive.

Along this highly populated length of beaches that include the cities of Milwaukee, Chicago, and Gary, beach closures are a nuisance to the public, resulting in limited recreational activities. Current monitoring protocols have been widely criticized because results of testing are not available until 18-24 hours after a sample is collected, a time period during which *E. coli* counts can change considerably (Whitman et al., 1999; Boehm et al., 2002). Furthermore, the variation in *E. coli* counts between adjoining and intervening beaches can be high, which leads to questions about management policies.

Alternatives to current monitoring protocols are being explored to improve the reliabity of management decisions regarding *E. coli* counts, and among these, predictive modeling has shown some promise. Predictive modeling relies on parameters that can be obtained immediately or within a short period of time rather than having to wait 18-24 hours. In predictive models, water and weather conditions are determined, and using a mathematical equation, probable *E. coli* count is determined. Other attempts at modeling have determined specifically whether *E. coli* count at a beach exceeds the 235 recommended level (Francy and Darner, 2002). Previous attempts at predictive modeling have included the variables rainfall in association with wind (Whitman et al., 1999; Haack et al., 2003), nearby discharge (Olyphant et al., 2003) and turbidity (Olyphant and Whitman, 2004; Nevers and Whitman, In Review). In order to develop a predictive model, however, a robust data set of *E. coli* counts is needed with coincident water and weather conditions. From this dataset, *E. coli* counts can be mathematically predicted using ambient water and weather data.

Predictive models are typically developed for single beaches, which can force beach managers to gather data and calculate results for dozens of beaches separately, if models are available for all of them. Typically, however, predictive models are only developed for the highest priority beaches, and the effort is not extended to all beaches within a given jurisdiction. Many beaches have well-established similarities with neighboring beaches, and combining the modeling effort to apply to several beaches could improve effectiveness of modeling and decrease associated costs.

Regional Forecast Model

Monitoring programs along southern Lake Michigan are among the most robust in the nation, with samples usually collected 5-7 days a week at many popular beaches. Furthermore, there are numerous weather and water monitoring stations in the region that have collected continuous data for many years. With the available database, a predictive model was designed

that could be used for the southern Lake Michigan region, extending from Milwaukee, Wisconsin to Michigan City, Indiana.

By incorporating historical monitoring data collected by the City of Milwaukee, City of Racine, Lake County Illinois Health Department, Chicago Park District, Gary Sanitary District, Indiana Dunes National Lakeshore, US Geological Survey, and US EPA from the beaches monitored from Milwaukee, Wisconsin to Michigan City, Indiana, the regional forecast model relies on ambient conditions measured concurrently with lake *E. coli* counts. Using these data, correlational relationships and multivariate regressions were developed with the outcome being a regional model of *E. coli* counts for use by beach managers.

DATA DESCRIPTION AND SOURCES

E. coli Data

E. coli monitoring data for the study years (2000-2003) were gathered from the several sources responsible for individual monitoring programs and included 55 beaches (Figure 1; Table 1). Sampling frequency was highly variable among jurisdictions, but the data were typically included in analyses with the given frequency of collection. Wisconsin beaches included in the analysis included two jurisdictions: Milwaukee and Racine. Milwaukee beaches were sampled generally four days a week in 2000 and seven days a week in 2001-2003. Racine beaches were sampled 5-7 days a week for 2000-2003. In Illinois, Lake County beaches were sampled seven days a week but only for years 2002 and 2003; prior to 2002, beaches were sampled for fecal coliform bacteria, which is not comparable to *E. coli* for data analysis. Chicago beaches were sampled five days a week for 2000-2003. Indiana beaches were sampled once a week for the period 2000-2003; in the event of a high *E. coli* count (above 235) a given beach was sampled again until *E. coli* count fell below 235 CFU. Prior to incorporation into the overall database, all data were checked for outliers and unusual values, and when data were determined to be likely in error, they were removed from the analyses, which included very few instances.

Figure 1. Southern Lake Michigan, indicating 55 beaches included in study.

Table 1. Beaches in the regional forecast model, listed from Milwaukee south to Michigan City, Indiana.

MILWAUKEE, WI	Bradford
	McKinley
	Wisconsin-South Shore
RACINE, WI	Zoo
	North
LAKE COUNTY, IL	North Point Marina
	Illinois Beach SP North
	Illinois Beach SP Sailing
	Illinois Beach SP South
	Waukegan North
	Waukegan South
	Lake Bluff
	Lake Forest
	Park Ave
	Rosewood
CHICAGO, IL	Juneway
	Rogers
	Howard
	Jarvis/Fargo
	Leone/Loyola/Greenleaf
	Pratt
	North Shore
	Albion
	Thorndale
	Hollywood
	Foster
	Montrose
	North Ave
	Oak
	Ohio
	12th
	31st
	49th
	57th
	63rd
	South Shore
	Rainbow
	Calumet
INDIANA	Lake Street
	Marquette Park
	Wells Street
	West Beach
	Ogden Dunes
	Dune Acres
	Porter
	State Park West
	State Park East
	Kemil
	Dunbar
	Lakeview
	Central
	Mount Baldy

Parameters considered for the model were collected from locations throughout the study region (Figure 2). Weather and water information were both collected, and averages were typically for the few hours prior to *E. coli* collection time (determined as 10 AM).

Weather Data

Data were collected from the National Climatic Center for several locations, including Milwaukee, Racine, and Kenosha, Wisconsin; Waukegan, Park Forest, Chicago and Midway Airport, Illinois; and Indiana Dunes and La Porte, Indiana. At each of these locations, the parameters collected included total daily precipitation (cm), daily maximum air temperature (°C), daily minimum air temperature (°C), and daily average temperature (°C).

More extensive data from Gary Regional Airport, Indiana were also considered in the model. Parameters collected included averages from 4-10 AM for air temperature (°C), wind direction (° from true north), wind speed (m/s), wind gust (m/s), atmospheric pressure (cm Hg), cloud cover (%), and cloud height (m).

Weather parameters collected by Great Lakes Environmental Laboratory, NOAA, were also included from several locations: Milwaukee, Chicago, and Michigan City, Indiana. These data were generated using some of their models, and included averages from 4-10 AM for air temperature (° C), minimum wind speed (m/s), maximum wind speed (m/s), average wind speed (m/s), wind direction (° from true north). Additionally, those data were used to calculate resulting vector magnitude and resulting vector direction for all three locations.

Insolation data were used from several locations and different lengths of exposure were considered for the models. From Wanatah, Indiana, insolation (MJ/m^2) was recorded for total daily insolation (24-hr), insolation from 4-10 AM (6-hr), and insolation from 8-10 AM (2-hr). From St Charles, Illinois, insolation was recorded as total daily (24-hr). Insolation for 4-10 AM (6-hr) was recorded for Chicago at Ohio Street, at Illinois Beach State Park, and at Gary Airport. In 2000 only, an onsite weather station was located at Chicago's 63rd Street Beach, and those data were also considered, using 4-10 AM (6-hr) data.

Water Data

Physical water conditions from monitoring stations located in Chicago and near Burns Ditch, Indiana and maintained by the US Army Corps of Engineers were also considered in the model. These data were averaged for the period 4-10 AM and included depth (m), wave height (m), wave period (seconds), and wave direction (° from true north). Water depth for two nearby outfalls, Calumet Harbor in Illinois and Burns Ditch in Indiana, were also considered.

Water conditions collected by NOAA at a Lake Michigan buoy were also examined. These included wind direction (° from true north), wind speed (m/s), wind gust (m/s), wave height (m), barometric pressure, air temperature (°C), and water temperature (°C).

Data Acquisition Sources

US Army Corps of Engineers Waterways Experiment Station
At *Chicago, Illinois*: wave height, wave period, wave direction, water depth
At *Burns Ditch, Indiana*: wave height, wave period, wave direction, water depth

National Climatic Data Center
At *Racine Wisconsin; Kenosha, Wisconsin; Waukegan, Illinois; Park Forest, Illinois; Midway airport Chicago, Illinois; Chicago Botanical Gardens, Chicago, Illinois; La Porte, Indiana, Indiana Dunes National Lakeshore* : precipitation, air temperature, weather events

National Climatic Data Center, At *Milwaukee, Wisconsin*: precipitation, air temperature, weather events, barometric pressure, wind direction, wind speed

Gary Regional Airport, *Gary, Indiana*: air temperature, wind direction, wind speed, wind gust speed, barometric pressure, cloud cover, cloud height

Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration,
At *Milwaukee, Wisconsin*: air temperature, wind speed, wind direction
At *Chicago, Illinois*: air temperature, wind speed, wind direction
At *Michigan City, Indiana*: air temperature, wind speed, wind direction

Water and Atmospheric Resources Program, At *St. Charles, Illinois*: solar insolation

Purdue University, Applied Meteorology Group. At *Wanatah, Indiana*: solar insolation

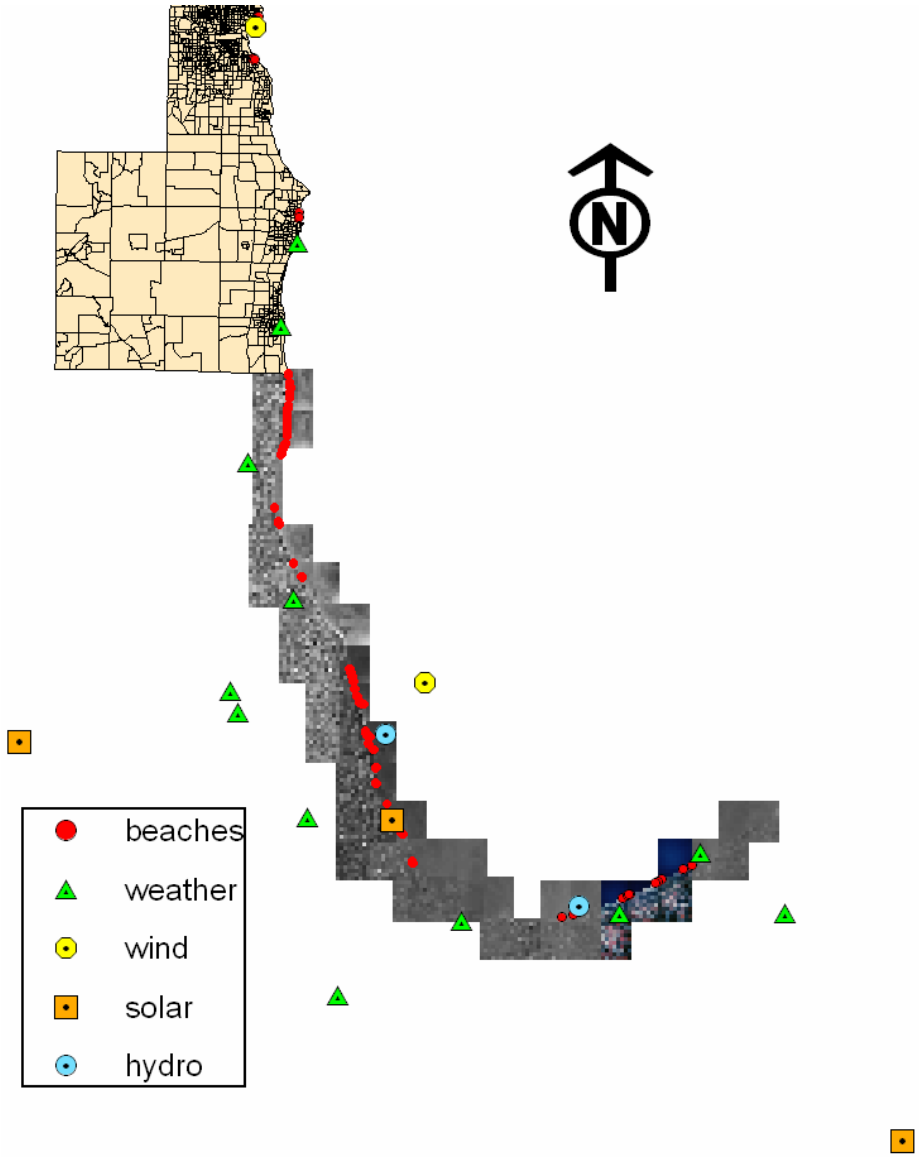


Figure 2. Locations for study beaches and stations from which hydrometeorological data were collected for consideration in the model.

Data Grouping

In order to maximize comparisons, site locations were often divided into groups based on geographic location. At the finest scale, individual beaches were considered separately. Next, zones of beaches were considered, which were identified primarily by municipality or managing entity, and these included Milwaukee, Wisconsin; Racine, Wisconsin; Lake County, Illinois; Chicago, Illinois (north); Chicago, Illinois (south); Gary District, Indiana; and Indiana Dunes National Lakeshore, Indiana. At the broadest level, state designations (Wisconsin, Illinois, Indiana) were considered.

In later analyses, several parameters were combined to get means that covered the entire study region. For minimum daily temperature (4-10AM) and rainfall (24-hr total), data collected at numerous locations were averaged: Milwaukee, Racine, Kenosha, Waukegan, Chicago, Midway, Indiana Dunes, and LaPorte. Wave height (4-10 AM) and wave period (4-10 AM) from stations in Chicago and Burns Harbor, Indiana were averaged. Average wind speed was calculated from data collected at Milwaukee, Chicago, Gary, and Michigan City. Also used in analyses, average pressure was calculated using data from Milwaukee and Gary, and insolation (4-10AM) was averaged from Illinois Beach State Park, Ohio Street, Gary, and Wanatah.

RESULTS

E. coli at Individual Beaches

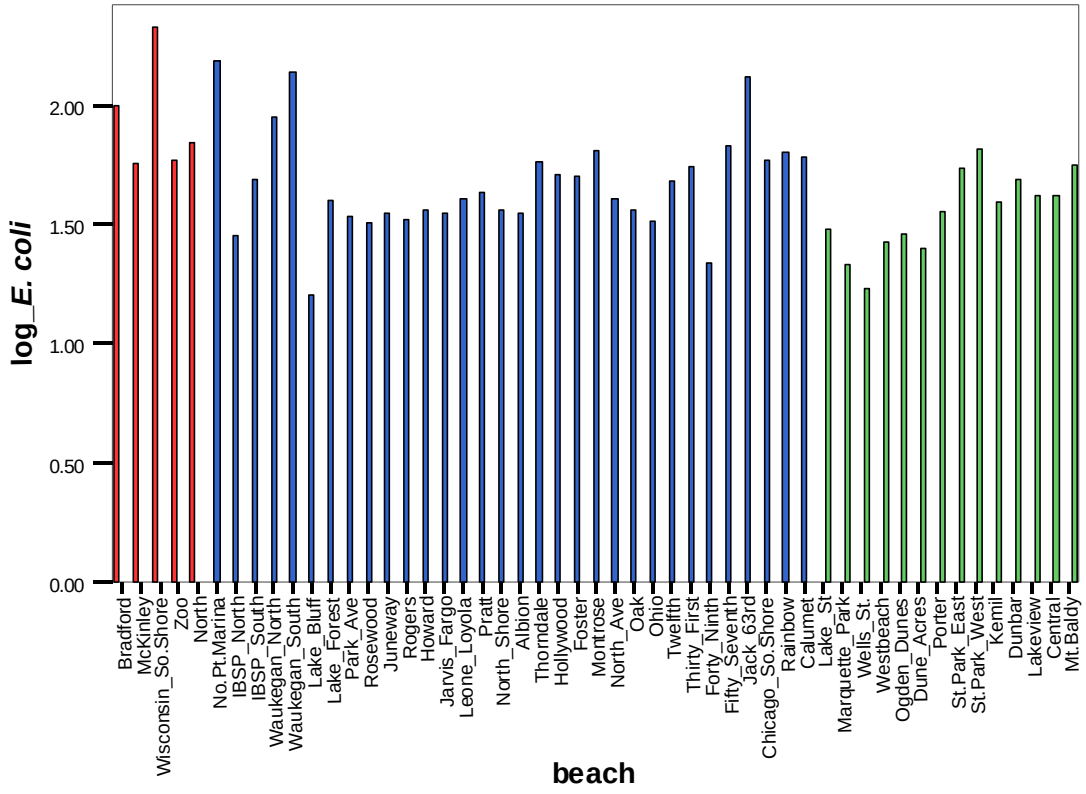


Figure 3. Mean log *E. coli* count for each beach over the 2000-2003 period. Different colors indicate state.

Mean *E. coli* counts over the four-year period were variable among beaches, with means ranging from 1.2-2.3 log *E. coli* CFU (Figure 3). Beaches with higher mean *E. coli* included South Shore in Wisconsin, North Point Marina, Waukegan South and 63rd Street in Illinois, and State Park West in Indiana. Highest counts overall were in Wisconsin and lowest in Indiana, but there was considerable variability between states in the frequency of sampling and number of beaches sampled.

Beaches within specific zones were significantly different from one another. In Milwaukee, *E. coli* counts at South Shore beach were significantly higher than at the other two beaches ($P<0.05$), with a mean log *E. coli* of 2.33. McKinley beach had the lowest mean *E. coli* counts. There were only two beaches for Racine, so no comparison of significance was made, but North Beach generally had higher mean *E. coli* counts. In Lake County, Illinois, North Point Marina Beach (log mean 2.19) and Waukegan South (log mean2.15) had significantly higher *E. coli* counts than all other Lake County beaches ($P<0.05$); Lake Bluff had the lowest mean (1.2). There were 24 Chicago beaches included in the study, so groupings were large; however, 63rd

Street Beach alone had significantly higher mean *E. coli* (2.13) than all other Chicago beaches ($P<0.05$), and 49th Street had the lowest mean *E. coli* count (1.34). In Gary Indiana, Lake Street had highest *E. coli* (1.58) and Wells Street the lowest (1.26), and for Indiana Dunes, groupings were large, but Dune Acres had the lowest mean *E. coli* (1.40) and State Park West had the highest (1.82).

When compared by zone, Milwaukee beaches had the significantly higher log mean *E. coli* (2.03) than other zones ($P<0.05$), and West Indiana (Gary) beaches had *E. coli* counts that were significantly lower (1.39) than all other beaches ($P<0.05$) (Table 2).

Table 2. Results of Duncan post-hoc test for mean log *E. coli* counts by zone.

Duncan

		Subset for alpha = .05				
Zone	N	1	2	3	4	5
West Indiana	460	1.3925				
North Chicago	4261		1.6178			
East Indiana	750		1.6491	1.6491		
Lake County, IL	1727			1.6988		
City of Racine	631				1.8061	
South Chicago	1235				1.8687	
City of Milwaukee	1155					2.0318
Sig.		1.000	.363	.149	.069	1.000

Means for groups in homogeneous subsets are displayed.
a Uses Harmonic Mean Sample Size = 923.327.
b The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

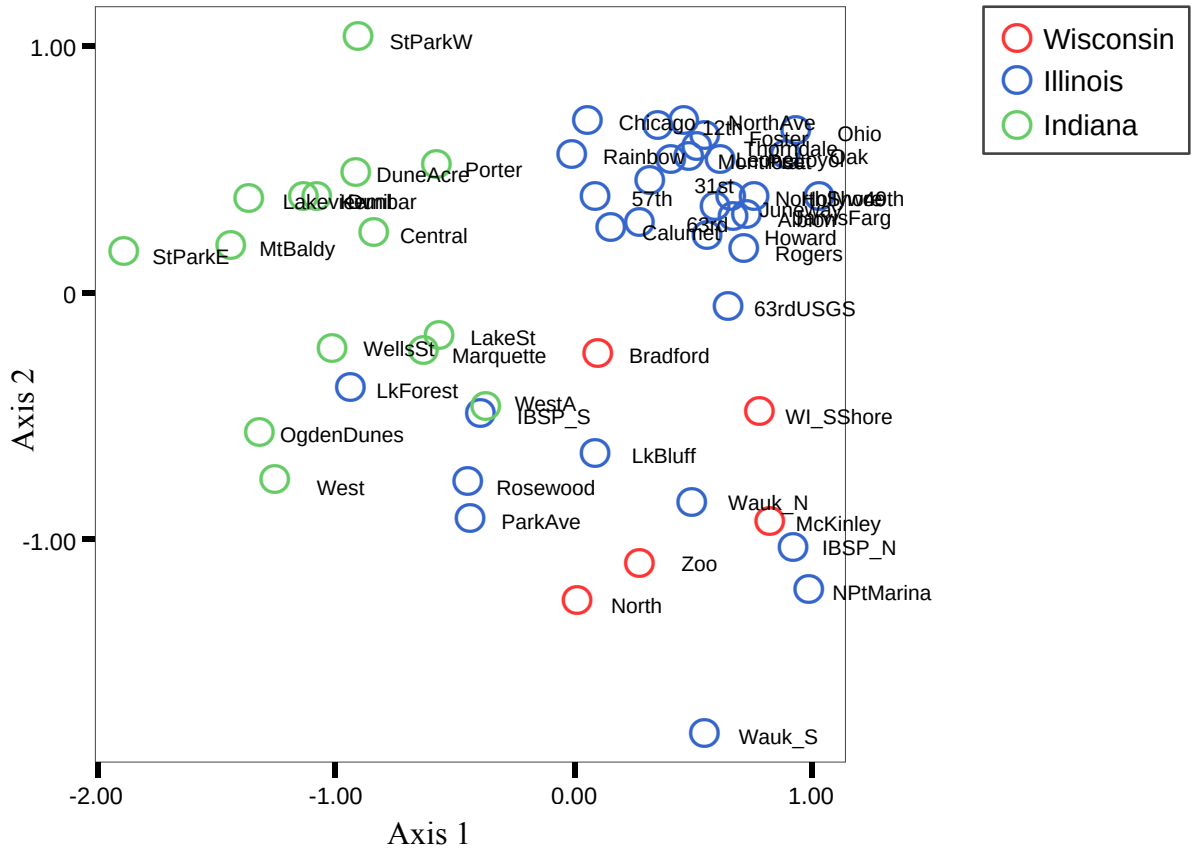


Figure 4. Multidimensional scaling results for all beaches included in the study. Different colors indicate states.

Results of multidimensional scaling showed that beaches grouped generally within states (Figure 4). Indiana beaches were grouped together somewhat loosely, as were Wisconsin beaches. Illinois beaches exhibited an interesting pattern, with Chicago beaches forming a tighter grouping together and the other Illinois beaches (in Lake County) appearing fairly widely distributed and separate from Chicago beaches.

A separate MDS of the Chicago beaches reveals a clearer picture of the relationship and results in an interesting pattern (Figure 5). The north Chicago beaches group with each other, and the south Chicago beaches appear widely distributed and separate from the north Chicago beaches. Chicago South Shore, Rainbow, Calumet, and 63rd Street are the furthest south in the city, and 57th lies just to the north of those.

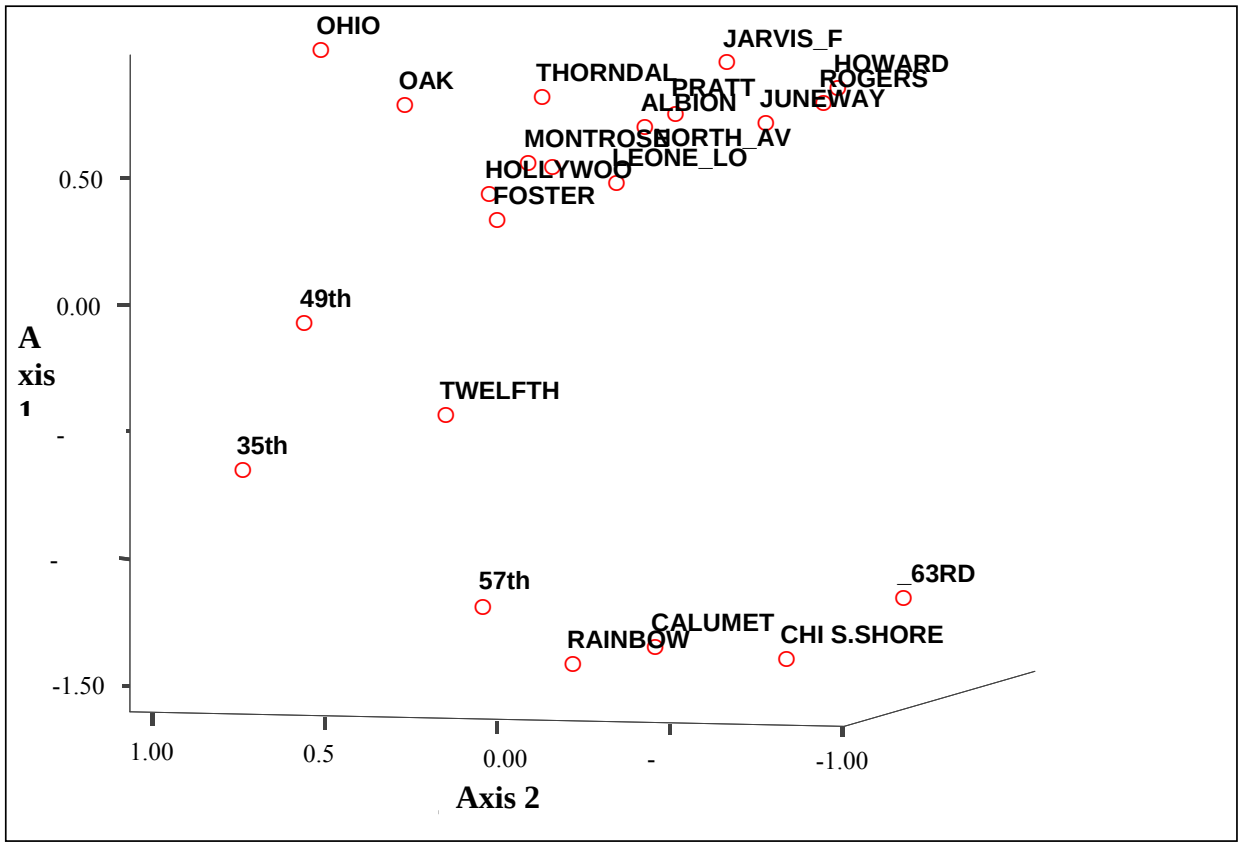


Figure 5. Multidimensional scaling results for Chicago beaches.

Although the Indiana beaches could not be used in many of the analyses due to the scarcity of available data (collected once a week), relationships between beaches could be highlighted using a correlation scatter graph matrix (Figure 6). Beaches close to each other geographically often show some relationship in overall *E. coli* counts, although numerous other factors can be more important than geographic proximity depending on the nature of the beach. The scatterplot matrix for these beaches shows the relationship between each pair of beaches, and general trends clearly show relationships among all beaches. Periodic high counts at either comparison beach reveal the high variation inherent in E. coli counts and influence the correlation between beaches overall.

Figure 6. Scatter matrix of Indiana beaches with best fit lines.

Milwaukee beaches were all highly correlated, with McKinley and Bradford with the highest Pearson R result (Table 3; Figure 7). These beaches lie adjacent to one another, so it was expected that *E. coli* counts would be similar and similarly affected by environmental influences. Clear boundaries in E. coli counts along each axis indicate the detection limit or the maximum count (missed dilution) for the method, in this case Colilert-18.

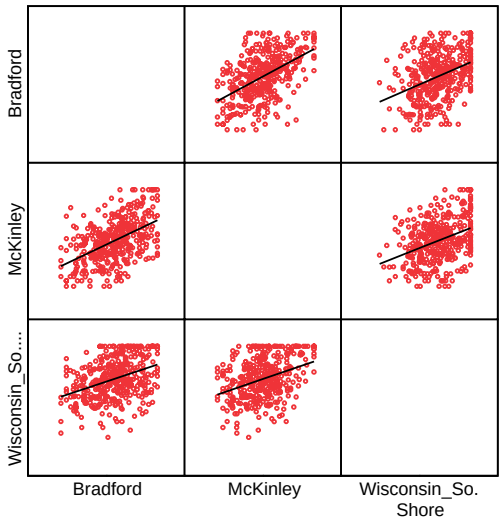
Table 3. Pearson correlation results for Milwaukee beaches. ** indicates a significant correlation at P<0.01.

Correlations

	Bradford	McKinley	Wisconsin_So. Shore
Bradford	1	.506(**)	.374(**)
McKinley	.506(**)	1	.364(**)
Wisconsin_So.Shore	.374(**)	.364(**)	1

** Correlation is significant at the 0.01 level (2-tailed).

Figure 7. Scatter matrix for Milwaukee beaches with best fit lines indicated.



In Racine, *E. coli* counts at the two beaches were highly correlated (Table 4; Figure 8). These beaches are similarly contiguous with a storm drainage dividing the two named areas from each other, so a high correlation would be expected.

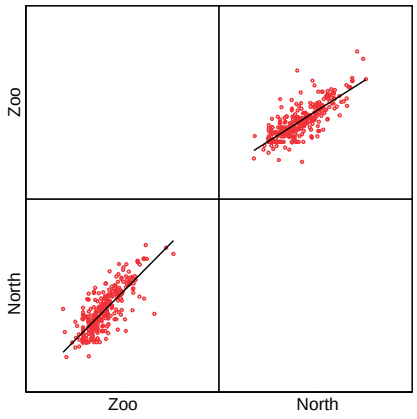
Table 4. Pearson correlation results for Racine beaches. ** indicates a significant correlation at P<0.01.

Correlations

	Zoo	North
Zoo	1	.799(**)
North	.799(**)	1

** Correlation is significant at the 0.01 level (2-tailed).

Figure 8. Scatter matrix for Racine beaches with best fit lines indicated.



Beaches in Lake County, Illinois are spread over a longer length of Lake Michigan than Milwaukee and Racine. All of the beaches had *E. coli* counts that were highly correlated with one another, but the Pearson R values were lower (Table 5; Figure 9). The highest R value was between Park Avenue and Rosewood (0.661), two contiguous beaches, and the lowest R value was between Rosewood and Waukegan South (0.154). Overall, the patterns can be seen in the scatterplot matrix, with lower correlations exhibiting almost a horizontal fit line.

Table 5. Pearson correlation results for Lake County, Illinois beaches. ** indicates a significant correlation at P<0.01.

	No.Pt.Marina	IBSP_N	IBSP_S	Waukegan_N	Waukegan_S	Lake_Bluff	Lake_Forest	Park_Ave	Rosewood
No.Pt.Marina	1	.467(**)	.380(**)	.415(**)	.487(**)	.240(**)	.352(**)	.344(**)	.224(**)
IBSP_North	.467(**)	1	.525(**)	.360(**)	.371(**)	.406(**)	.378(**)	.447(**)	.422(**)
IBSP_South	.380(**)	.525(**)	1	.476(**)	.382(**)	.605(**)	.599(**)	.642(**)	.561(**)
Waukegan_North	.415(**)	.360(**)	.476(**)	1	.456(**)	.403(**)	.427(**)	.446(**)	.390(**)
Waukegan_South	.487(**)	.371(**)	.382(**)	.456(**)	1	.262(**)	.339(**)	.314(**)	.154(*)
Lake_Bluff	.240(**)	.406(**)	.605(**)	.403(**)	.262(**)	1	.594(**)	.576(**)	.600(**)
Lake_Forest	.352(**)	.378(**)	.599(**)	.427(**)	.339(**)	.594(**)	1	.520(**)	.596(**)
Park_Ave	.344(**)	.447(**)	.642(**)	.446(**)	.314(**)	.576(**)	.520(**)	1	.661(**)
Rosewood	.224(**)	.422(**)	.561(**)	.390(**)	.154(*)	.600(**)	.596(**)	.661(**)	1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

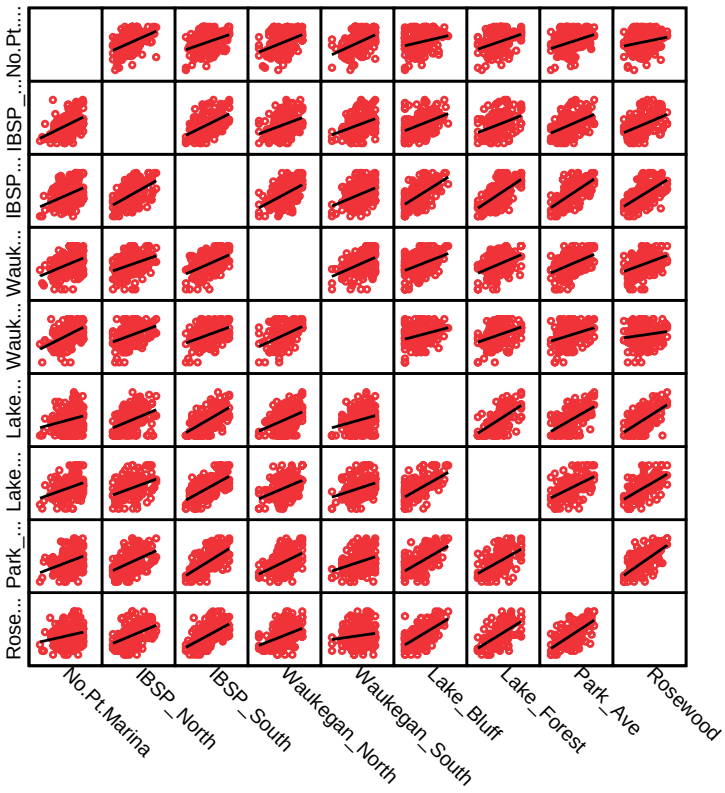


Figure 9. Scatter matrix for Lake County, Illinois beaches with best fit lines indicated.

E. coli counts for beaches in North Chicago were all highly correlated with one another (Table 6; Figure 10). Albion and North Shore beaches had a very high Pearson R (R=0.910); these beaches are in close proximity to one another. The lowest value was between Jarvis-Fargo and 31st Street Beaches (R=0.427), two beaches a great distance from one another. Generally, however, the beaches were more closely correlated with one another than the beaches of Lake County were to each other. The correlations seen in the scatterplot matrix show the similarity among beaches, with a general appearance of close correlations in all comparisons.

Table 6. Pearson correlation results for North Chicago, Illinois beaches. ** indicates a significant correlation at P<0.01.

	Juneway	Rogers	Jarvis_Fargo	Pratt	Leone_Loyola	Albion	North_Shore	Thorndale	Hollywood	Foster	Montrose	North_Ave	Oak	Ohio	Twelfth	Thirty_First	Forty_Ninth
Juneway	1	.878(**)	.867(**)	.777(**)	.845(**)	.820(**)	.791(**)	.706(**)	.711(**)	.699(**)	.704(**)	.719(**)	.711(**)	.696(**)	.556(**)	.456(**)	.523(**)
Rogers	.878(**)	1	.802(**)	.742(**)	.813(**)	.798(**)	.758(**)	.663(**)	.660(**)	.680(**)	.677(**)	.713(**)	.689(**)	.666(**)	.577(**)	.464(**)	.499(**)
Jarvis_Fargo	.867(**)	.802(**)	1	.752(**)	.841(**)	.773(**)	.755(**)	.681(**)	.695(**)	.653(**)	.684(**)	.693(**)	.701(**)	.667(**)	.545(**)	.427(**)	.547(**)
Pratt	.777(**)	.742(**)	.752(**)	1	.831(**)	.803(**)	.822(**)	.729(**)	.778(**)	.717(**)	.735(**)	.755(**)	.733(**)	.688(**)	.599(**)	.512(**)	.565(**)
Leone_Loyola	.845(**)	.813(**)	.841(**)	.831(**)	1	.818(**)	.803(**)	.777(**)	.782(**)	.761(**)	.798(**)	.814(**)	.787(**)	.733(**)	.640(**)	.522(**)	.559(**)
Albion	.820(**)	.798(**)	.773(**)	.803(**)	.818(**)	1	.910(**)	.768(**)	.752(**)	.744(**)	.757(**)	.777(**)	.738(**)	.713(**)	.622(**)	.540(**)	.566(**)
North_Shore	.791(**)	.758(**)	.755(**)	.822(**)	.803(**)	.910(**)	1	.756(**)	.759(**)	.712(**)	.733(**)	.768(**)	.749(**)	.723(**)	.584(**)	.526(**)	.586(**)
Thorndale	.706(**)	.663(**)	.681(**)	.729(**)	.777(**)	.768(**)	.756(**)	1	.848(**)	.737(**)	.773(**)	.745(**)	.706(**)	.663(**)	.588(**)	.495(**)	.542(**)
Hollywood	.711(**)	.660(**)	.695(**)	.778(**)	.782(**)	.752(**)	.759(**)	.848(**)	1	.793(**)	.802(**)	.750(**)	.759(**)	.712(**)	.644(**)	.564(**)	.583(**)
Foster	.699(**)	.680(**)	.653(**)	.717(**)	.761(**)	.744(**)	.712(**)	.737(**)	.793(**)	1	.801(**)	.781(**)	.726(**)	.718(**)	.662(**)	.526(**)	.548(**)
Montrose	.704(**)	.677(**)	.684(**)	.735(**)	.798(**)	.757(**)	.733(**)	.773(**)	.802(**)	.801(**)	1	.786(**)	.724(**)	.680(**)	.616(**)	.511(**)	.555(**)
North_Ave	.719(**)	.713(**)	.693(**)	.755(**)	.814(**)	.777(**)	.768(**)	.745(**)	.750(**)	.781(**)	.786(**)	1	.819(**)	.750(**)	.643(**)	.549(**)	.562(**)
Oak	.711(**)	.689(**)	.701(**)	.733(**)	.787(**)	.738(**)	.749(**)	.706(**)	.759(**)	.726(**)	.724(**)	.819(**)	1	.818(**)	.604(**)	.495(**)	.574(**)
Ohio	.696(**)	.666(**)	.667(**)	.688(**)	.733(**)	.713(**)	.723(**)	.663(**)	.712(**)	.718(**)	.680(**)	.750(**)	.818(**)	1	.546(**)	.430(**)	.567(**)
Twelfth	.556(**)	.577(**)	.545(**)	.599(**)	.640(**)	.622(**)	.584(**)	.588(**)	.644(**)	.662(**)	.616(**)	.643(**)	.604(**)	.546(**)	1	.716(**)	.641(**)
Thirty_First	.456(**)	.464(**)	.427(**)	.512(**)	.522(**)	.540(**)	.526(**)	.495(**)	.564(**)	.526(**)	.511(**)	.549(**)	.495(**)	.430(**)	.716(**)	1	.685(**)
Forty_Ninth	.523(**)	.499(**)	.547(**)	.565(**)	.559(**)	.566(**)	.586(**)	.542(**)	.583(**)	.548(**)	.555(**)	.562(**)	.574(**)	.567(**)	.641(**)	.685(**)	1

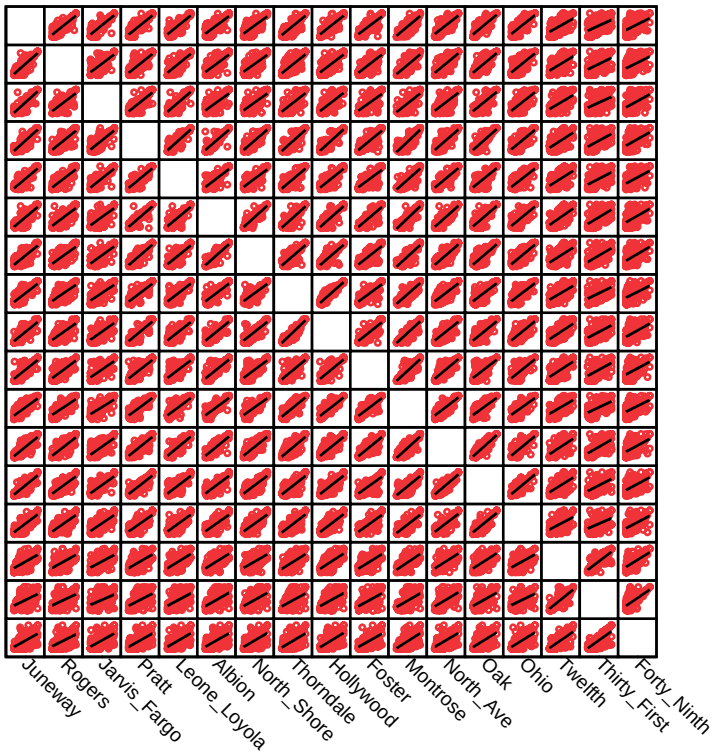


Figure 10. Scatter matrix for North Chicago, Illinois beaches with best fit lines indicated.

Beaches in southern Chicago were also highly correlated with one another (Table 7; Figure 11). Far fewer beaches were included in this zone, and the highest correlation was between Calumet and Rainbow (R=0.743), two beaches close to one another, and the lowest was between Calumet and 63rd Street (R=0.582). As a group of beaches, other than Racine, these beaches were the most closely correlated with one another. These high correlations are apparent in the scatterplot matrix, with few singular high counts at individual beaches; high counts at one beach tend to be associated with high counts at all other beaches.

Table 7. Pearson correlation results for South Chicago, Illinois beaches. ** indicates a significant correlation at P<0.01.

	Fifty_Seventh	Jack_63rd	Chicago_So.Shore	Rainbow	Calumet
Fifty_Seventh	1	.717(**)	.627(**)	.618(**)	.620(**)
Jack_63rd	.717(**)	1	.690(**)	.627(**)	.582(**)
Chicago_So.Shore	.627(**)	.690(**)	1	.699(**)	.602(**)
Rainbow	.618(**)	.627(**)	.699(**)	1	.743(**)
Calumet	.620(**)	.582(**)	.602(**)	.743(**)	1

** Correlation is significant at the 0.01 level (2-tailed).

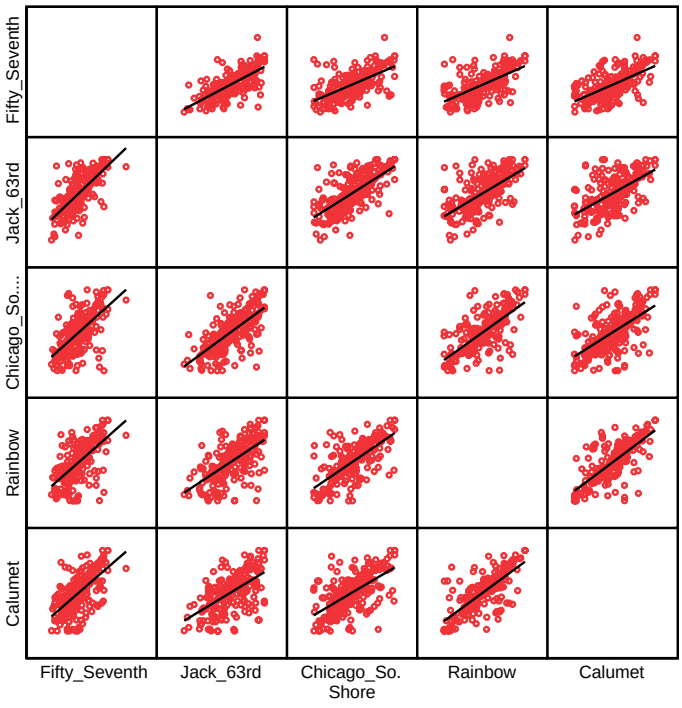


Figure 11. Scatter matrix for South Chicago, Illinois beaches with best fit lines indicated.

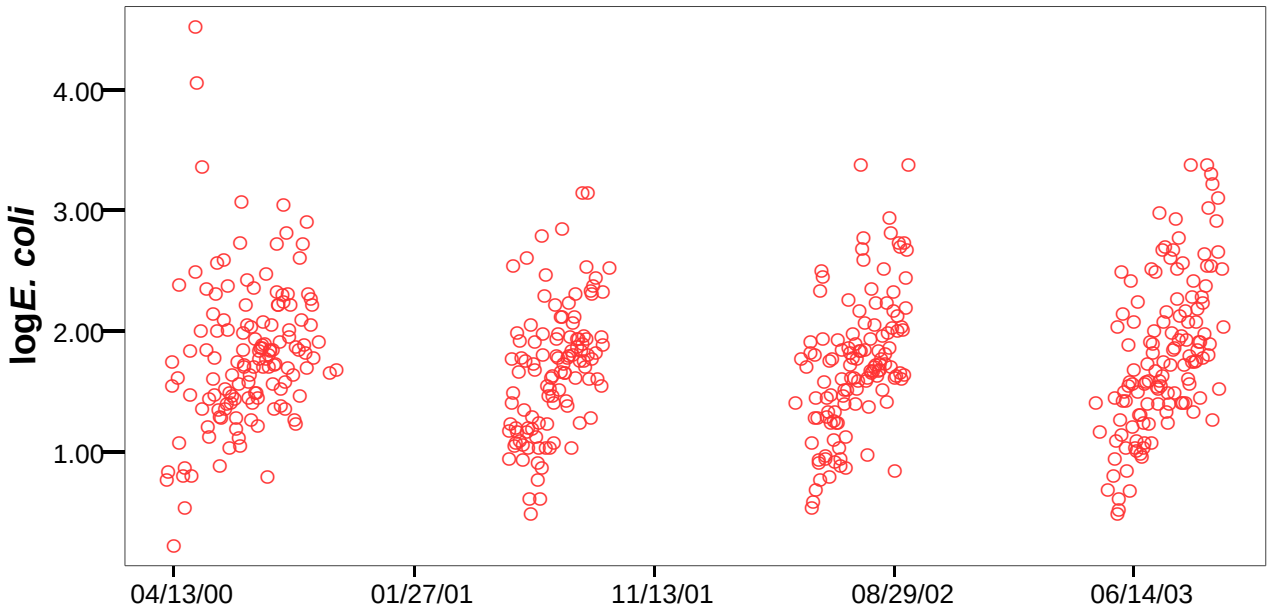


Figure 12. Scatter graph of mean log *E. coli* counts each year showing seasonal increase in *E. coli* counts

E. coli showed a seasonal pattern in all four years analyzed (Figure 12). Generally, *E. coli* increases throughout the swimming season, with many beaches reporting the greatest number of closures in August. This seasonal trend in *E. coli* has been seen at other times (Whitman et al., 1999), and could be the result of increasing water and air temperatures or else concentration, persistence, and resuspension.

Modeling

The currently used model for *E. coli* sampling involves collecting a water sample, incubating it for 18-24 hours and then making a management decision on whether to close the beach based on the number of colonies present. Comparing day of sampling with day of results reporting, this approach results in a R^2 of 0.192 for the entire dataset. The results by zone for this approach were variable (Table 8).

Table 8. Result of regression model using *E. coli* count from previous day to predict current day's count (currently used approach).

Zone	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
City of Milwaukee	1	.465(a)	.216	.215	.66102
City of Racine	1	.360(a)	.129	.127	.55913
Lake County, IL	1	.385(a)	.149	.148	.74810
North Chicago	1	.441(a)	.195	.194	.67472
South Chicago	1	.351(a)	.124	.123	.65521
West Indiana	1	.323(a)	.104	.096	.48823
East Indiana	1	.081(a)	.007	-.007	.52105

a Predictors: (Constant), lec_mn_1

Component Matrix(a)

Principal component analysis (factor analysis) was performed on the independent variables used in the models to understand better the relationship between predictors and to explore whether reducing the number of factors would be advantageous. Table 9 and Figure 13 show the rotated component matrix using varimax with Kaiser normalization. Ten independent factors were reduced to four factor components. Cloud cover, sunlight, rainfall contributed most to Factor 1; wave height, wave period and wind speed Factor 2; lake depth and temperature, Factor 3; while Factor 4 was best described by $E.\ coli_{t-1}$. Regression analysis was then performed using these four factors; all were coefficients, were significant at $p<0.0001$, and for the overall model R^2 was 0.245. Since there were a sufficient number of observations relative to the number of independent variables available and the amount of variation was explained by the original variable, the extracted factors were not used further in regression analysis.

Table 9 Relationships between parameters used in the predictive model using principal component analysis.

Rotated Component Matrix(a)

	Component			
	1	2	3	4
Zcloudcover	-.791	.096	-.059	.003
Zlakedepth	.066	-.042	.723	.237
Zsunlight	.853	.067	.157	-.076
Zrain	-.633	-.113	-.069	.020
Zpressure	.556	.190	-.202	.351
Ztemperture	.083	.050	.822	-.094
Zwaveheight	.136	.868	-.071	.296
Zperiod	.249	.810	.028	.157
Zwindspeed	-.321	.695	.077	-.380
ZEcoli_lag	-.081	.152	.152	.829

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 6 iterations.

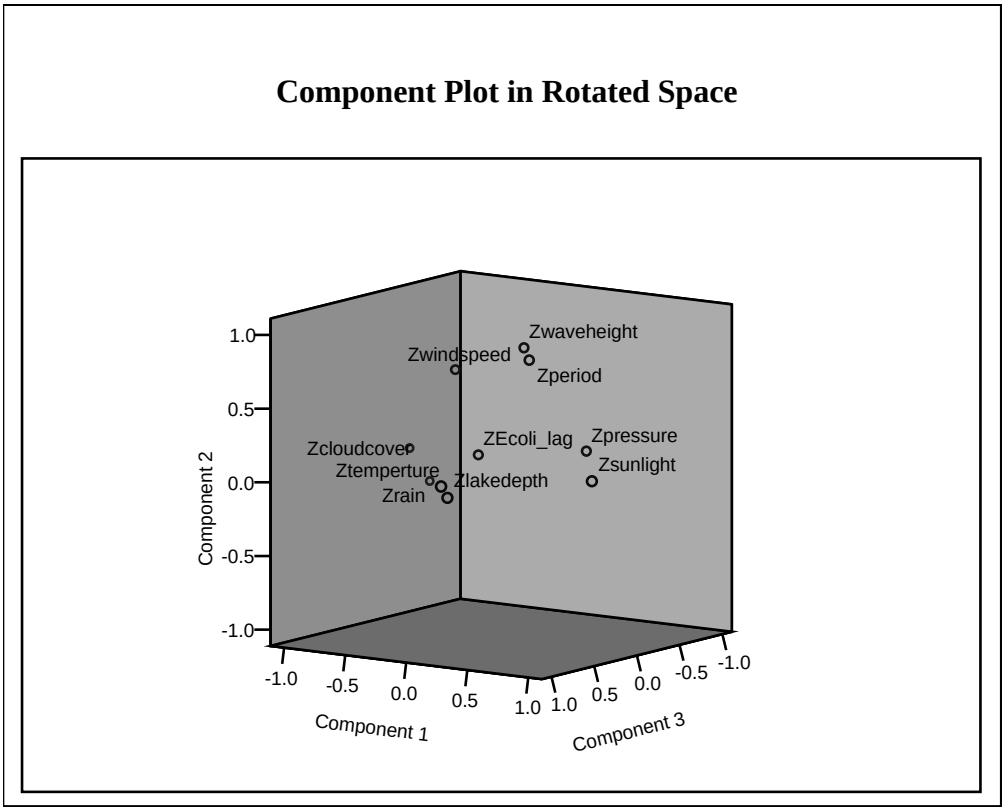


Figure 13 Resulting configuration of principal component analysis of variables used in the model.

In a model using univariate analysis of variance, *E. coli* counts were included for separate zones. The variables used in the model were previous day's *E. coli*, average rainfall, average wave height and period, depth of Calumet Harbor, minimum air temperature, and an interaction term of wind speed and wind direction (Table 10). The interaction term used a weighted wind direction term that accounts for wind vector and average wind speed. All parameters were log-transformed averages from several collection locations. The resulting model had an R^2 of 0.285.

Table 10. Parameters used in the best model for entire regional dataset.

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	699.494(a)	13	53.807	130.330	.000
Intercept	6.620	1	6.620	16.035	.000
lec_mn_1	191.443	1	191.443	463.706	.000
laverain	62.062	1	62.062	150.323	.000
lavewave	53.788	1	53.788	130.284	.000
laveperiod	11.781	1	11.781	28.534	.000
lavecaldepth	6.591	1	6.591	15.965	.000
lavemintemp	73.513	1	73.513	178.060	.000
avewdspXwdcode	25.140	1	25.140	60.894	.000
Zone	59.733	6	9.956	24.114	.000
Error	1752.153	4244	.413		
Total	16271.731	4258			
Corrected Total	2451.647	4257			

a R Squared = .285 (Adjusted R Squared = .283)

Variations on the model were explored, including changing the wind direction increments to onshore and offshore components. The beach zone was also included as a parameter, and the resulting model improved the model only slightly to 0.291. A second variation had zone and wind direction as interacting variables, which also increased the model slightly, to 0.293.

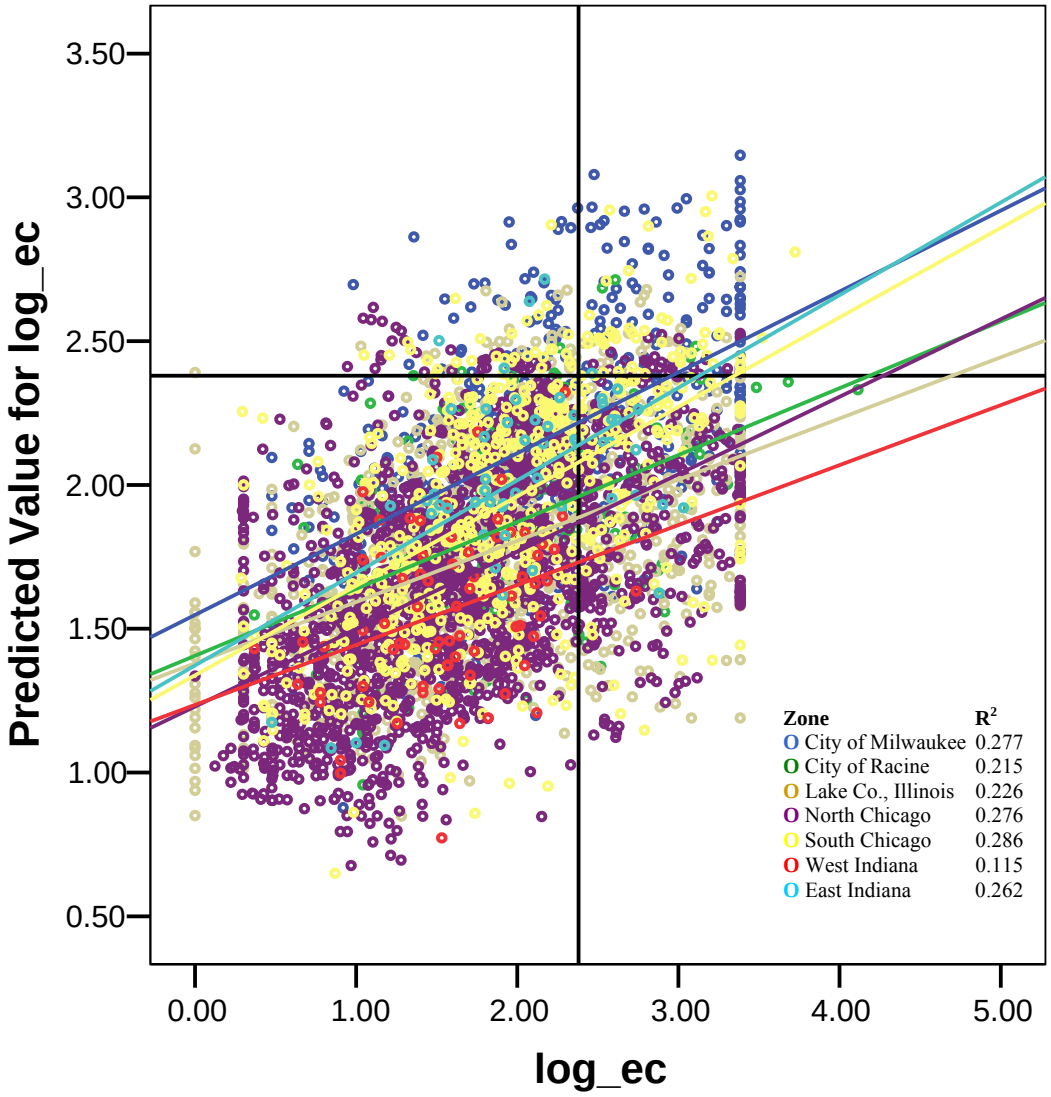


Figure 14. Actual *E. coli* count measured vs. count predicted using the best model developed. Colors indicate different zones, and lines indicate performance of zones within the overall model.

The resulting model produces lower R^2 values when the predictions are examined for individual zones (Figure 14). The Indiana beaches were sampled far less frequently, and as a result, the *E. coli* counts are more difficult to predict. It is apparent that the model still produces numerous false positives and false negatives. Of the 4258 cases included in the prediction, 3149 occasions (74% of samples) when *E. coli* counts were below log 2.38 CFU were correctly predicted, and 188 occasions (4% of samples) when *E. coli* counts were above log 2.38 CFU were correctly predicted. On 148 occasions (3% of samples), *E. coli* was below log 2.38 CFU, but the model predicted the count was above (false positive), and on 773 occasions (18% of samples), *E. coli* was above log 2.38 CFU, and the model predicted the count was below (false negative). It should be noted that counts below log 2.38 are far more common, but 22% of the samples had counts above the 2.38 limit. Overall, for 78% of the samples, *E. coli* level was

accurately predicted as above or below 235, which is the goal of current recommended monitoring programs by the EPA.

The vertical lines of data points indicate either detection limits (low) or maximum count possible using analysis method (high). Generally, it is noticeable that the Milwaukee beaches have higher counts and that many low counts were recorded for Chicago beaches. Chicago beaches dominated the overall number of cases included in the analysis, and Indiana beaches had far fewer samples than all other zone locations.

Running the same model on individual zones yielded less successful models, overall (Table 11). In this exercise, data from zones were input in the same model to examine how the parameters interact with *E. coli* data from each zone. There were not enough data points for the input parameters to complete model analyses for the Indiana beaches. There was no improvement when the model was applied to individual zones, and many parameters were no longer significant in the analyses.

Table 11. Results of model as applied to different regional zones.

Tests of Between-Subjects Effects

Dependent Variable: log_ec

Zone	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Milwaukee	Corrected Model	84.544(a)	8	10.568	28.315	.000
	Intercept	10.612	1	10.612	28.433	.000
	windcode_gyy_onoffshore	.243	1	.243	.652	.420
	lec_mn_1	25.313	1	25.313	67.821	.000
	laverain	10.888	1	10.888	29.174	.000
	lavewave	3.577	1	3.577	9.585	.002
	laveperiod	1.492	1	1.492	3.997	.046
	lavecaldepth	10.618	1	10.618	28.449	.000
	lavemintemp	6.344	1	6.344	16.997	.000
	avewdspXwdcode	2.117	1	2.117	5.672	.018
	Error	188.481	505	.373		
	Total	2656.891	514			
	Corrected Total	273.025	513			
Racine	Corrected Model	29.919(b)	8	3.740	14.699	.000
	Intercept	.510	1	.510	2.005	.158
	windcode_gyy_onoffshore	2.551	1	2.551	10.027	.002
	lec_mn_1	5.043	1	5.043	19.822	.000
	laverain	15.109	1	15.109	59.384	.000
	lavewave	.326	1	.326	1.281	.259
	laveperiod	.096	1	.096	.379	.539
	lavecaldepth	.511	1	.511	2.008	.158
	lavemintemp	.043	1	.043	.170	.681
	avewdspXwdcode	2.771	1	2.771	10.890	.001
	Error	58.263	229	.254		
	Total	887.462	238			

	Corrected Total	88.182	237			
Lake Co.	Corrected Model	166.651(c)	8	20.831	42.457	.000
	Intercept	10.040	1	10.040	20.463	.000
	windcode_gyy_onoffshore	1.734	1	1.734	3.535	.060
	lec_mn_1	59.996	1	59.996	122.282	.000
	laverain	19.030	1	19.030	38.786	.000
	lavewave	9.804	1	9.804	19.981	.000
	laveperiod	6.759	1	6.759	13.775	.000
	lavecaldepth	10.124	1	10.124	20.634	.000
	lavemintemp	.004	1	.004	.007	.931
	avewdspXwdcode	7.727	1	7.727	15.749	.000
	Error	435.690	888	.491		
	Total	3360.017	897			
	Corrected Total	602.341	896			
N Chicago	Corrected Model	313.170(d)	8	39.146	98.495	.000
	Intercept	2.606	1	2.606	6.556	.011
	windcode_gyy_onoffshore	2.765	1	2.765	6.956	.008
	lec_mn_1	36.232	1	36.232	91.163	.000
	laverain	5.795	1	5.795	14.580	.000
	lavewave	42.890	1	42.890	107.914	.000
	laveperiod	10.304	1	10.304	25.925	.000
	lavecaldepth	2.645	1	2.645	6.656	.010
	lavemintemp	86.537	1	86.537	217.733	.000
	avewdspXwdcode	.001	1	.001	.004	.952
	Error	745.608	1876	.397		
	Total	6395.986	1885			
	Corrected Total	1058.779	1884			
S Chicago	Corrected Model	77.707(e)	8	9.713	27.255	.000
	Intercept	.032	1	.032	.090	.765
	windcode_gyy_onoffshore	.218	1	.218	.612	.434
	lec_mn_1	15.075	1	15.075	42.300	.000
	laverain	11.768	1	11.768	33.020	.000
	lavewave	4.877	1	4.877	13.685	.000
	laveperiod	.618	1	.618	1.733	.189
	lavecaldepth	.031	1	.031	.086	.770
	lavemintemp	11.971	1	11.971	33.591	.000
	avewdspXwdcode	3.694	1	3.694	10.365	.001
	Error	208.487	585	.356		
	Total	2528.523	594			
	Corrected Total	286.195	593			
W Indiana	Corrected Model	2.409(f)	7	.344	1.709	.122
	Intercept	.000	0	.	.	.
	windcode_gyy_onoffshore	.000	0	.	.	.
	lec_mn_1	.000	0	.	.	.

	laverain	.000	0	.	.	.
	lavewave	.000	0	.	.	.
	laveperiod	.000	0	.	.	.
	lavecaldepth	.000	0	.	.	.
	lavemintemp	.000	0	.	.	.
	avewdspXwdcode	.000	0	.	.	.
	Error	13.093	65	.201		
	Total	193.307	73			
	Corrected Total	15.503	72			
E Indiana	Corrected Model	5.423(g)	7	.775	3.570	.004
	Intercept	.000	0	.	.	.
	windcode_gyy_onoffshore	.000	0	.	.	.
	lec_mn_1	.000	0	.	.	.
	laverain	.000	0	.	.	.
	lavewave	.000	0	.	.	.
	laveperiod	.000	0	.	.	.
	lavecaldepth	.000	0	.	.	.
	lavemintemp	.000	0	.	.	.
	avewdspXwdcode	.000	0	.	.	.
	Error	10.633	49	.217		
	Total	249.545	57			
	Corrected Total	16.055	56			

a R Squared = .310 (Adjusted R Squared = .299)
b R Squared = .339 (Adjusted R Squared = .316)
c R Squared = .277 (Adjusted R Squared = .270)
d R Squared = .296 (Adjusted R Squared = .293)
e R Squared = .272 (Adjusted R Squared = .262)
f R Squared = .155 (Adjusted R Squared = .064)
g R Squared = .338 (Adjusted R Squared = .243)

The model was then tested on individual beaches to see if there was an improvement in model capability when beaches were examined on an even smaller scale. Indiana beaches were not analyzed because the dataset was not sufficient. The model results were highly variable, with R^2 ranging from 0.094 to 0.474 (Figure 15); models were significant for all beaches except North Point Marina.

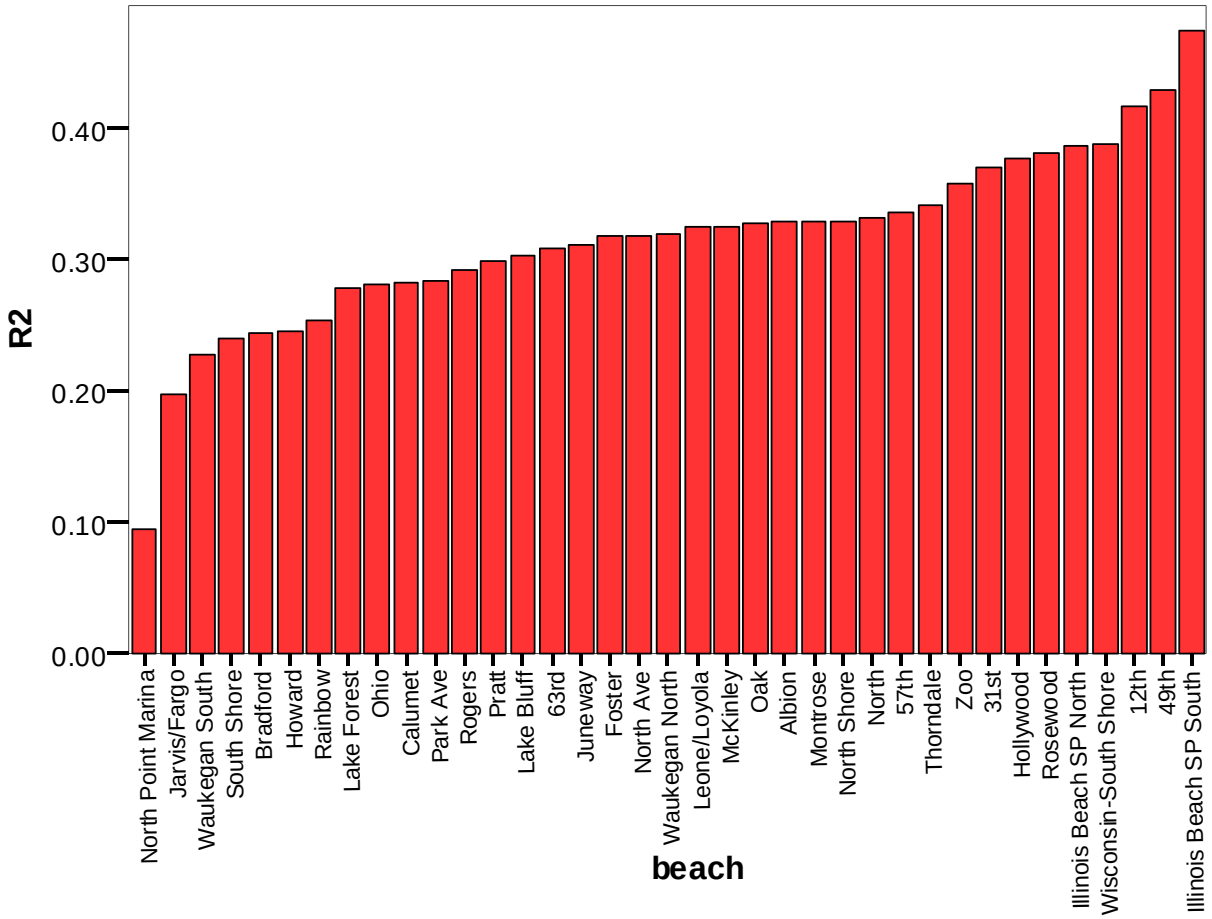


Figure 15. R^2 values for individual study beaches when best model applied to separate beaches. Beaches are ranked in order of R^2 value.

Classification and Regression Trees

Regression trees were used to indicate which parameters would subdivide the *E. coli* dataset, indicating the parameters having the greatest effect on counts. In an analysis of the dataset (N=10348), previous *E. coli* count, as 3-day prior moving average, was the first to subdivide the data, with counts lower than 1.68 log CFU resulting in a lower overall mean of *E. coli* (Figure 16). The lower *E. coli* group was further subdivided by rainfall, with lower rainfall associated with lower *E. coli* counts. The higher 3-day prior moving average group was then subdivided by wave height, with higher waves associated with higher *E. coli* counts. Improvements with branching were limited. Importance of each of the input parameters is also calculated as part of the analysis (Table 12).

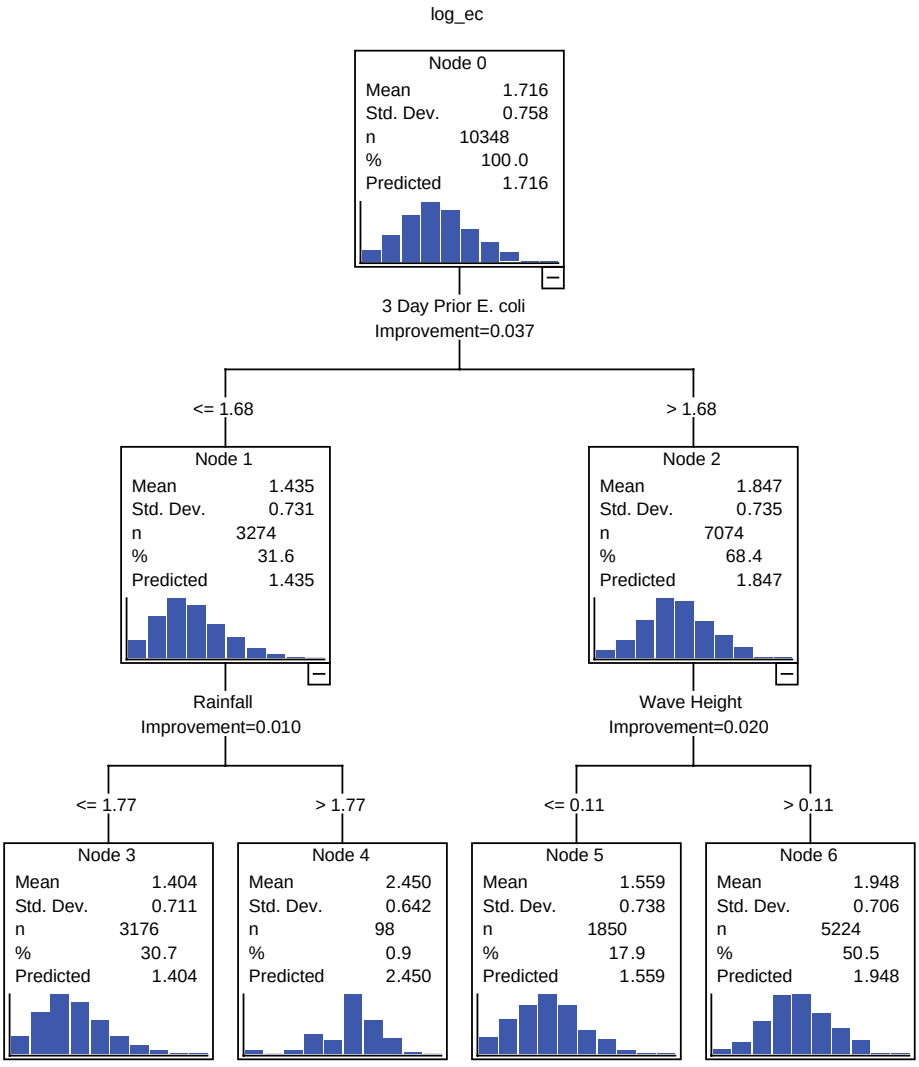


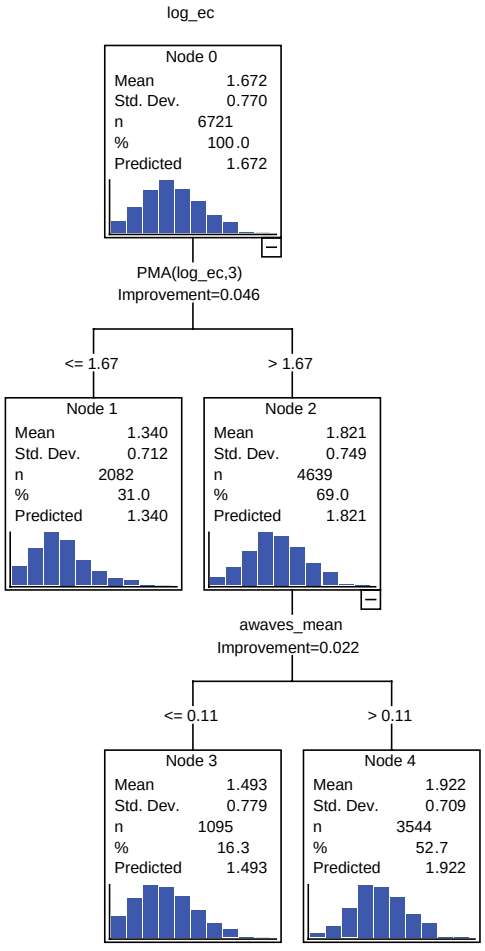
Figure 16. Regression tree for entire dataset considered in the study.

Table 12 Independent variable importance as determined by regression tree for the entire dataset.

Independent Variable	Importance	Normalized Importance
3 Day Prior *E. coli*	.037	100.0%
Wave Height	.032	85.7%
Temperature	.014	38.3%
Rainfall	.014	38.1%
Pressure	.010	27.7%
Sunlight	.001	1.5%

Growing Method: CRT
Dependent Variable: log_ec

Wind direction was also considered with its effect on the classification tree, and a model with south winds had similar results, with 3-day prior moving average having the most importance in the model, followed by wave height and air temperature (Figure 17). Rainfall was far less important when only south winds were considered, with normalized importance resulting as less than 0.1%.



Independent Variable Importance

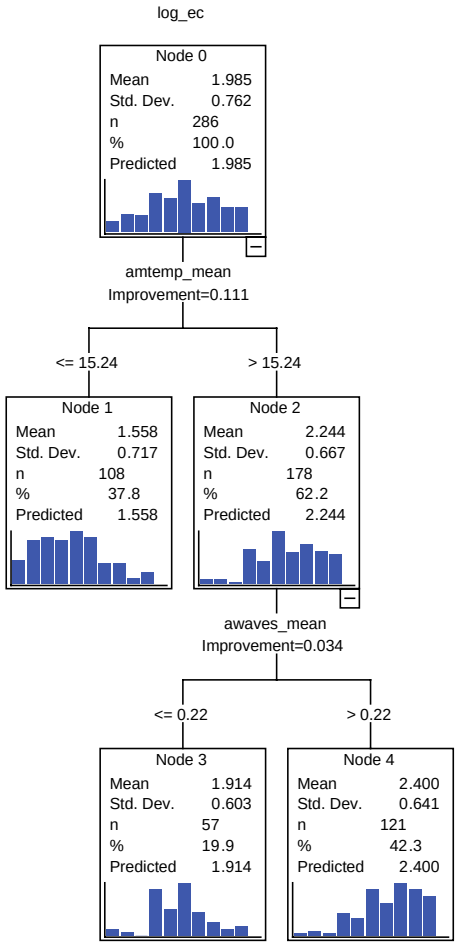
Independent Variable	Importance	Normalized Importance
log_ec_3dayPMA	.046	100.0%
awaves_mean	.041	88.4%
amtemp_mean	.027	59.6%
asolar_mean	.000	.7%
awspeed	.000	.3%
averain_mean	5.78E-006	.0%

Growing Method: CRT
Dependent Variable: log_ec

Figure 17. Regression tree for all data collected on days of prevailing south winds. Table indicates independent variable importance, as derived from the regression tree.

Regression trees were also developed for north and south winds for two select zones: south Chicago and Milwaukee.

Results for south Chicago beaches during north winds indicated that air temperature was the first variable to split the dataset (Figure 18). Air temperature is often a surrogate factor for seasonality because *E. coli* and air temperature both increase through the course of the beach season. The *E. coli* group with a higher mean count was then split by wave height, with higher waves associated with an *E. coli* group whose mean was 2.4 log CFU. This is significant because the mean is above the 2.38 log CFU threshold for beach advisory as recommended by the EPA.



Independent Variable Importance

Independent Variable	Importance	Normalized Importance
amtemp_mean	.113	100.0%
log_ec_3dayPMA	.046	40.8%
awspeed	.043	38.0%
awaves_mean	.034	29.7%
asolar_mean	.001	.7%

Growing Method: CRT
Dependent Variable: log_ec

Figure 18. Regression tree for south Chicago beaches on days of prevailing north winds. Table indicates independent variable importance, as derived from the regression tree.

The classification tree for south winds for South Chicago similarly was first divided by air temperature, and the group with lower air temperature was then subdivided by amount of rainfall (Figure 19). The subset with higher air temperature was secondarily subdivided by 3-day prior moving average of *E. coli*. The dataset divided several more times with different subgroups, but the important parameters were wave height, air temperature, solar insolation, and 3-day prior moving average (Table 13).

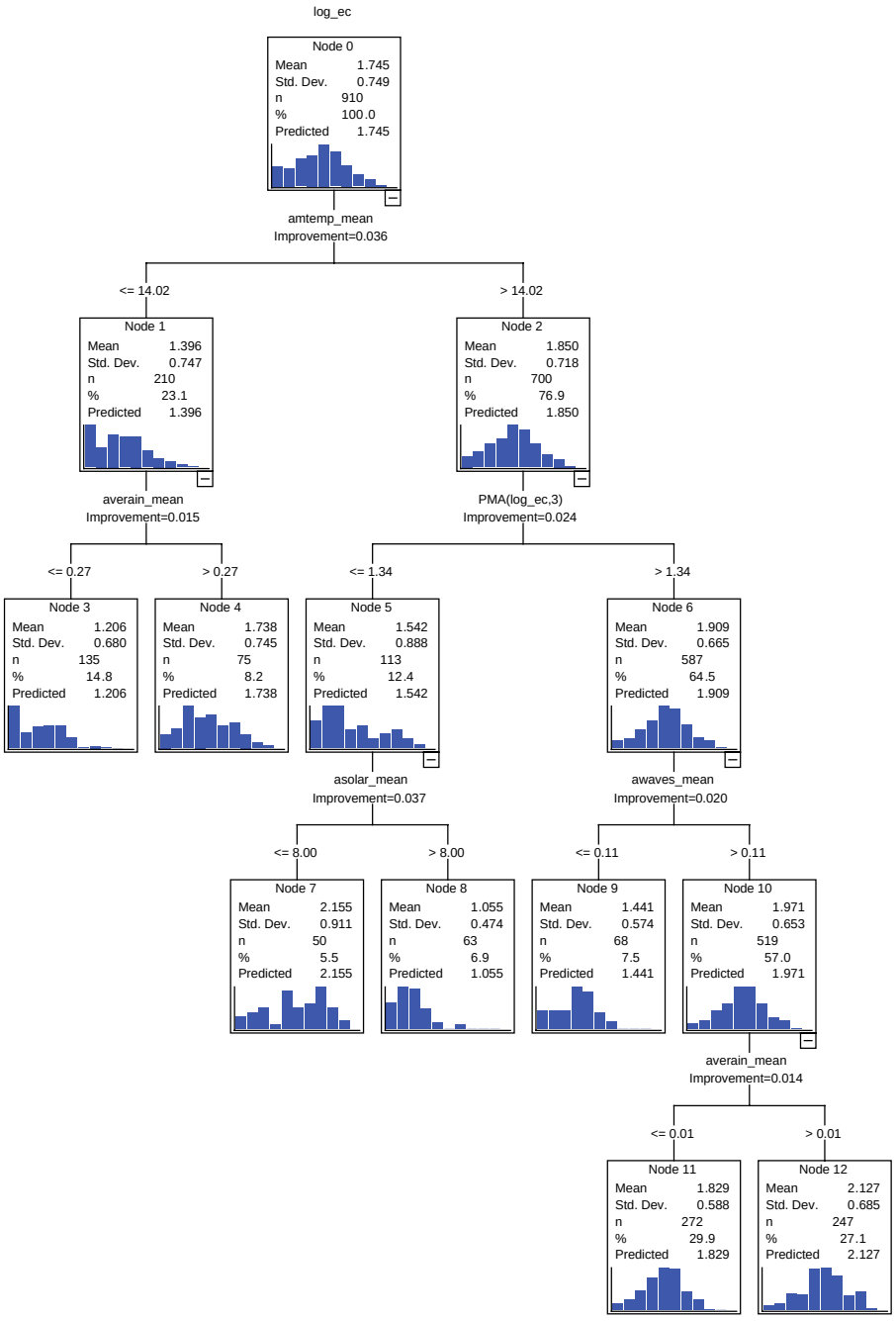


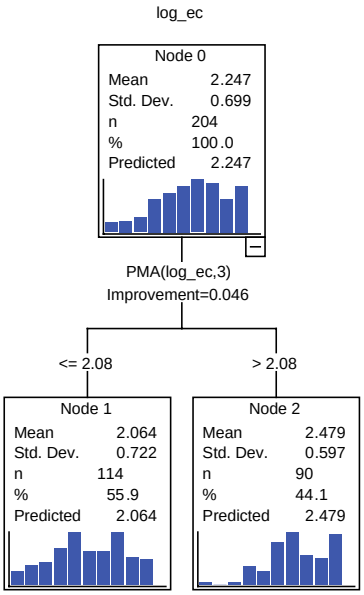
Figure 19 Regression tree for south Chicago beaches on days of prevailing south winds.

Table 13. Independent variable importance for south Chicago during prevailing south winds, as derived from the regression tree.

Independent Variable	Importance	Normalized Importance
Awaves_mean	.059	100.0%
amtemp_mean	.051	86.4%
asolar_mean	.039	66.2%
log_ec_3dayPMA	.039	65.7%
Averain_mean	.038	63.5%
awspeed	.008	12.9%

Growing method: CRT
Dependent Variable: log_ec

For the Milwaukee beaches, north winds resulted in a tree with 3-day prior moving average of *E. coli* causing the primary split of the dataset (Figure 20). It was the most important variable in the model, and it caused the only split, given the input definition of the model.

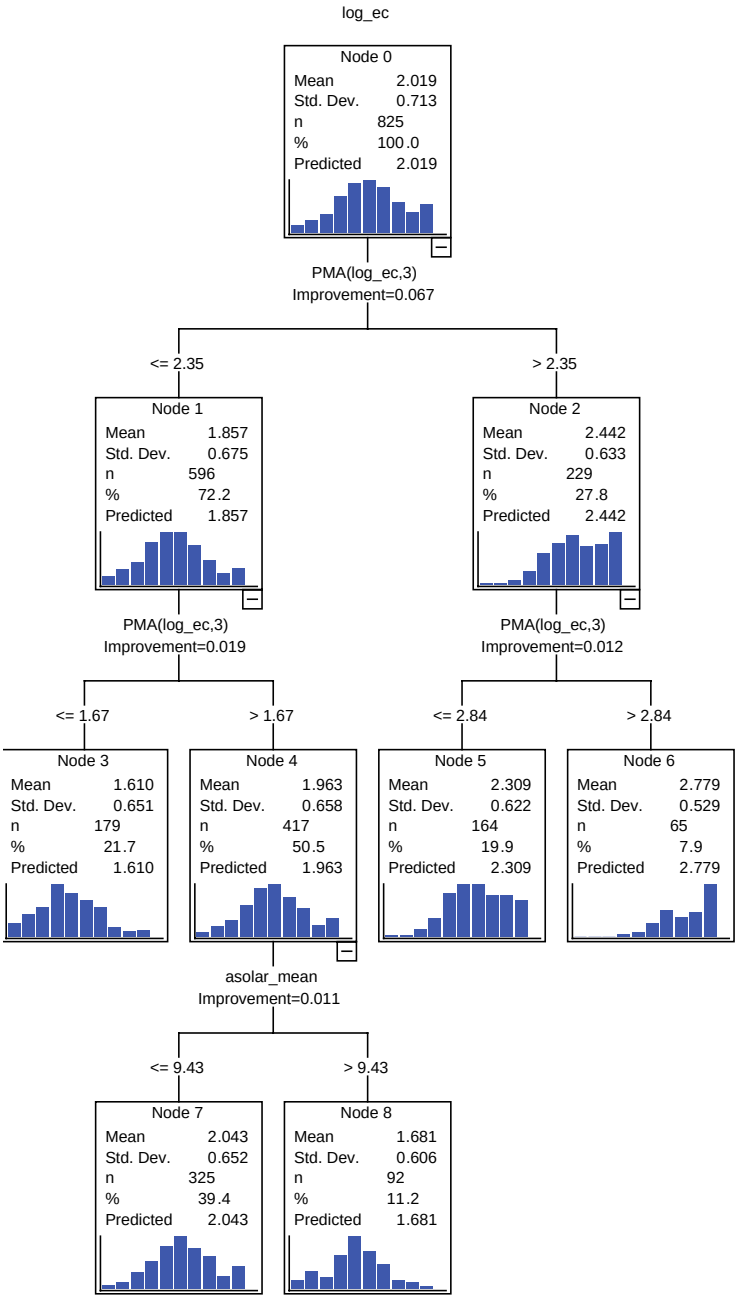


Independent Variable	Importance	Normalized Importance
log_ec_3dayPMA	.046	100.0%
Awaves_mean	.015	31.6%
amtemp_mean	.008	16.4%
asolar_mean	.004	8.9%
Averain_mean	.003	5.8%
awspeed	.000	.6%

Growing Method: CRT
Dependent Variable: log_ec

Figure 20. Regression tree for Milwaukee beaches on days of prevailing north winds. Table indicates independent variable importance, as derived from the regression tree.

Similarly, 3-day prior moving average of *E. coli* first subdivided the dataset for Milwaukee during south winds (Figure 21). The division was for cases with *E. coli* counts lower than 2.35 log CFU and cases with *E. coli* counts higher, which is close to the EPA advisory limit of 2.38 log CFU. Further subdividing was based on different 3-day prior moving average levels, which was identified as the most important variable; all other variables had far lower importance values.



Independent Variable	Importance	Normalized Importance
log_ec_3dayPMA	.099	100.0%
asolar_mean	.017	17.1%
amtemp_mean	.009	9.2%
awaves_mean	.004	3.6%
awspeed	.002	1.6%
averain_mean	.001	1.5%

Growing Method: CRT
Dependent Variable: log_ec

Figure 21. Regression tree for Milwaukee beaches on days of prevailing south winds. Table indicates independent variable importance, as derived from the regression tree.

Hierarchical Cluster Analysis

In order to determine which beaches could be modeled together in smaller groups, hierarchical cluster analysis was used. A cluster of all beaches in Wisconsin and Illinois was created; Indiana beaches were omitted due to the scarcity of available data relative to other beaches studied. The resulting cluster showed numerous groupings, with some expected clustering of beaches that are in close proximity to one another (Figure 22).

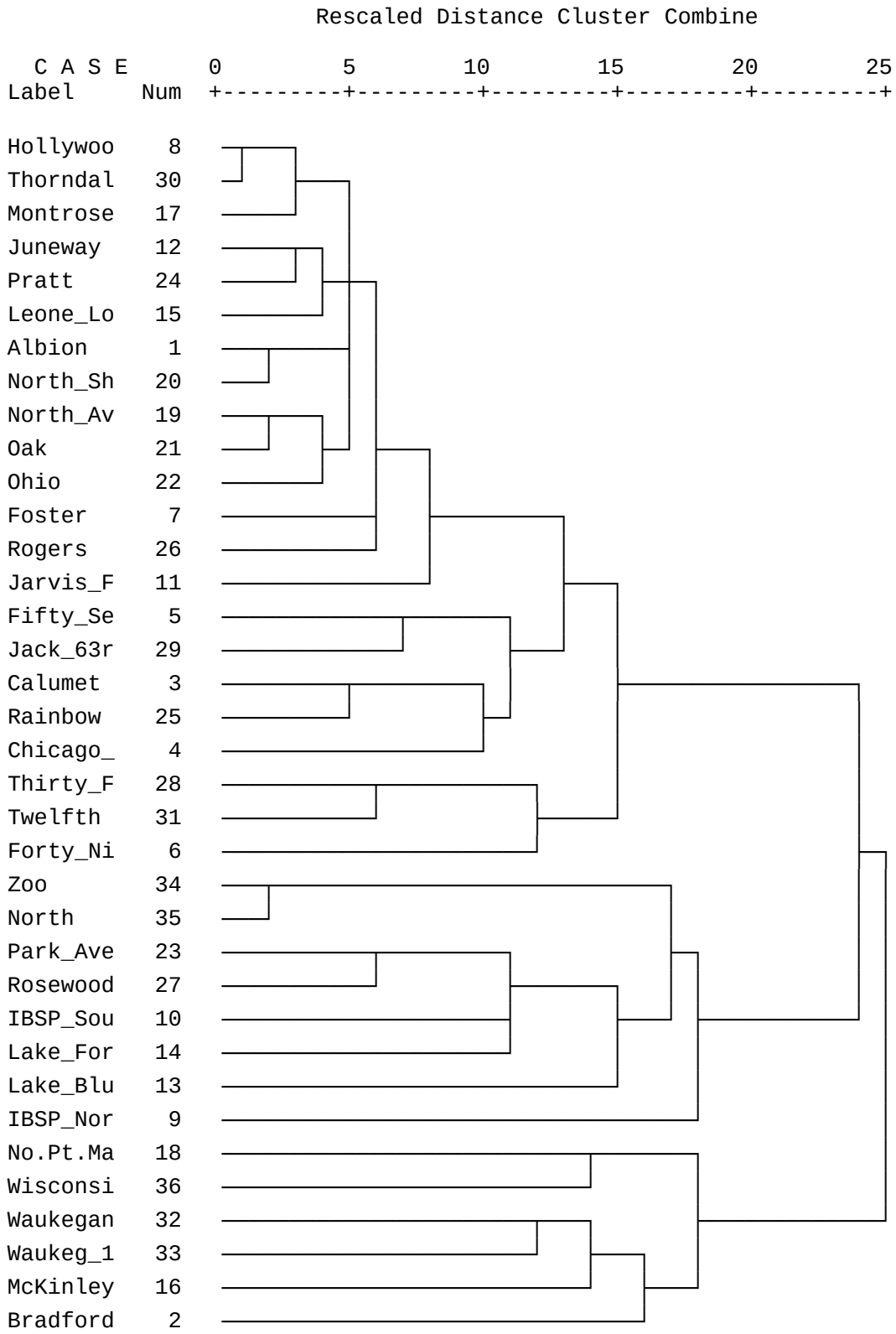


Figure 22. Dendrogram of results from hierarchical clustering analysis for all beaches but Indiana beaches.

The beaches were clustered again, and specific numbers of clusters were forced: 2, 3, and 4 clusters (Table 14). The groupings based on three clusters were then used as smaller groups for modeling.

Table 14. Results of hierarchical clustering analysis with forced number of clusters. Numbers indicate cluster membership. All beaches but Indiana beaches included.

Case	4 Clusters	3 Clusters	2 Clusters
Albion	1	1	1
Bradford	2	2	2
Calumet	1	1	1
Chicago_So.Shore	1	1	1
Fifty_Seventh	1	1	1
Forty_Ninth	1	1	1
Foster	1	1	1
Hollywood	1	1	1
IBSP_North	3	3	1
IBSP_South	4	3	1
Jarvis_Fargo	1	1	1
Juneway	1	1	1
Lake_Bluff	4	3	1
Lake_Forest	4	3	1
Leone_Loyola	1	1	1
McKinley	2	2	2
Montrose	1	1	1
No.Pt.Marina	2	2	2
North_Ave	1	1	1
North_Shore	1	1	1
Oak	1	1	1
Ohio	1	1	1
Park_Ave	4	3	1
Pratt	1	1	1
Rainbow	1	1	1
Rogers	1	1	1
Rosewood	4	3	1
Thirty_First	1	1	1
Jack_63rd	1	1	1
Thorndale	1	1	1
Twelfth	1	1	1
Waukegan_North	2	2	2
Waukegan_South	2	2	2
Zoo	4	3	1
North	4	3	1
Wisconsin_So.Shore	2	2	2

Modeling Groups of Beaches

From hierarchical clustering, cluster 1 was the largest group of beaches, with 23 Chicago beaches. If the same model is used for this subgroup, each of the parameters remains significant but for Calumet Harbor depth; however, with this included or removed from the model, the result is R^2=0.315 (Table 15; Figure 23).

Table 15. Results of best model applied to one of three clusters determined through hierarchical clustering analysis. The cluster included Chicago beaches.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	349.617(a)	27	12.949	38.355	.000
Intercept	6.117	1	6.117	18.118	.000
lec_mn_1	38.902	1	38.902	115.228	.000
laverain	15.709	1	15.709	46.531	.000
lavewave	43.427	1	43.427	128.632	.000
avewdspXwdcode	5.408	1	5.408	16.018	.000
laveperiod	8.596	1	8.596	25.462	.000
lavemintemp	87.444	1	87.444	259.012	.000
code	31.447	21	1.497	4.436	.000
Error	761.302	2255	.338		
Total	7653.034	2283			
Corrected Total	1110.919	2282			

a R Squared = .315 (Adjusted R Squared = .307)

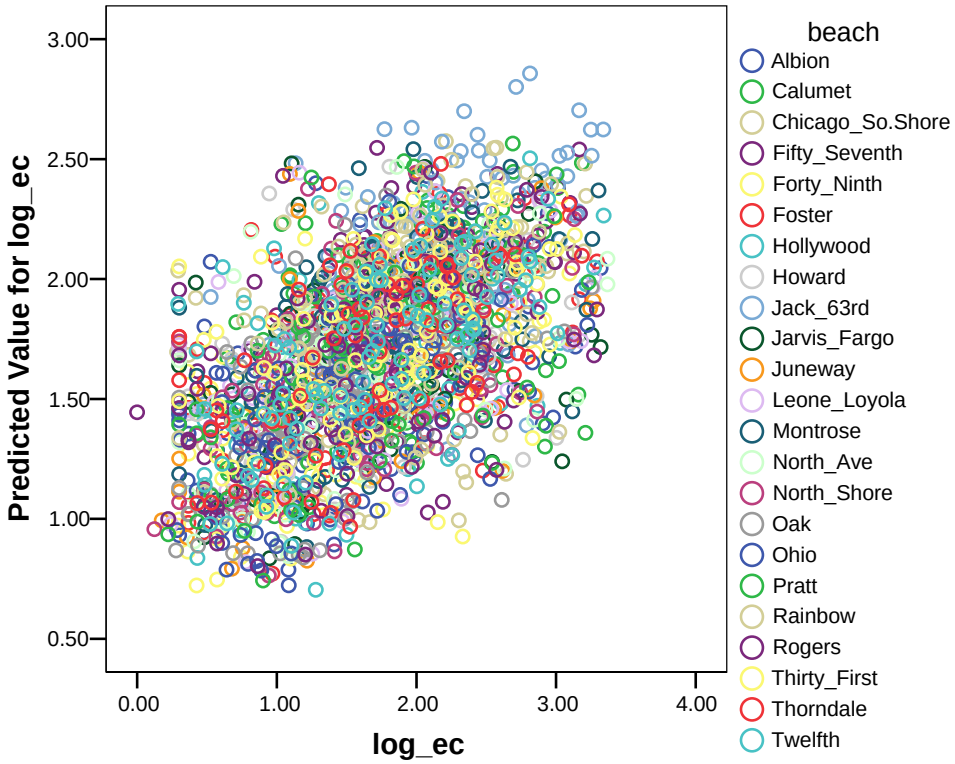


Figure 23. Scatter of measured mean log *E. coli* count and predicted count using best developed model. Colors indicate individual beaches; this group includes Chicago beaches.

Cluster 2 included six beaches: the 3 Milwaukee beaches (Bradford, McKinley, and Wisconsin South Shore) and 3 beaches located in Lake County Illinois (North Point Marina, Waukegan North, and Waukegan South). In the model, wave period was no longer significant, and the model resulted in an R^2 of 0.301; values at the maximum count level were eliminated for the analysis (Table 16; Figure 24).

Table 16. Results of best model applied to one of three clusters determined through hierarchical clustering analysis. The cluster included Milwaukee and Lake County Illinois beaches.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	97.848(a)	12	8.154	25.518	.000
Intercept	14.820	1	14.820	46.380	.000
lec_mn_1	8.685	1	8.685	27.181	.000
laverain	9.650	1	9.650	30.201	.000
lavewave	6.425	1	6.425	20.106	.000
avewdspXwdcode	1.463	1	1.463	4.577	.033
laveperiod	1.005	1	1.005	3.144	.077
lavemintemp	5.500	1	5.500	17.213	.000
lavecaldepth	14.837	1	14.837	46.432	.000
code	20.824	5	4.165	13.034	.000
Error	227.193	711	.320		
Total	3520.311	724			
Corrected Total	325.040	723			

a R Squared = .301 (Adjusted R Squared = .289)

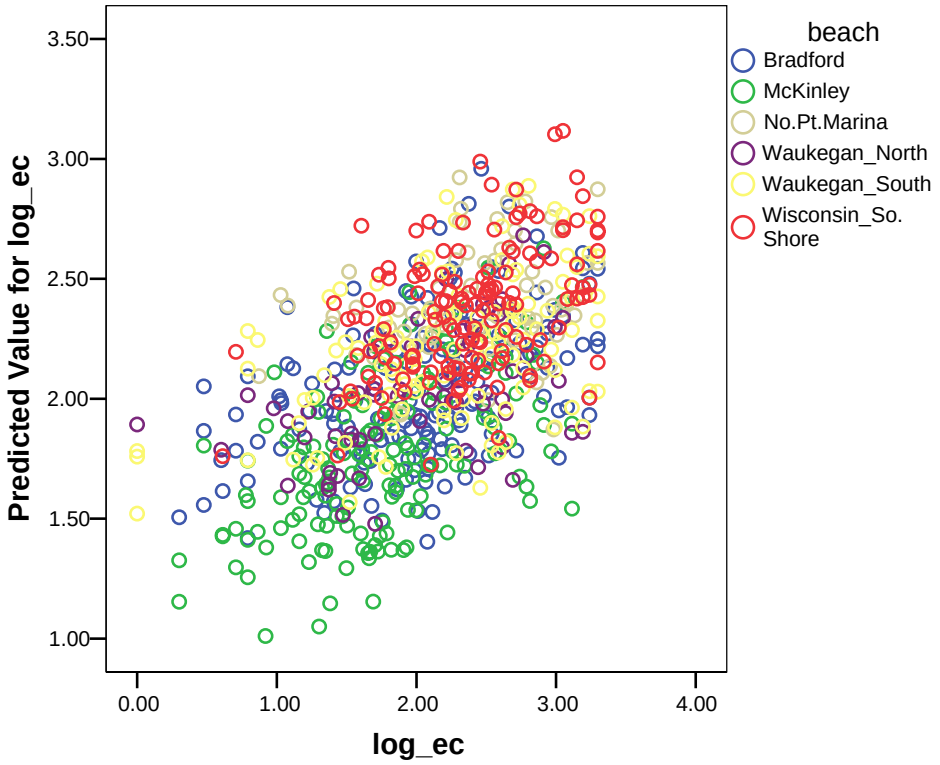


Figure 24. Scatter of measured mean log *E. coli* count and predicted count using best developed model. Colors indicate individual beaches; this group includes Milwaukee and Lake County, Illinois beaches.

The third cluster included 6 Lake County, Illinois beaches (IBSP North, IBSP South, Lake Bluff, Lake Forest, Park Ave., Rosewood) and the 2 Racine beaches (North, Zoo). If the model is applied to this group of beaches, neither minimum temperature nor Calumet Harbor depth remains a significant parameter, and removing them from the model results in an R^2 of 0.292 (Table 17; Figure 25).

Table 17. Results of best model applied to one of three clusters determined through hierarchical clustering analysis. The cluster included Racine, Wisconsin and Lake County, Illinois beaches.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	122.734(a)	12	10.228	28.886	.000
Intercept	11.670	1	11.670	32.958	.000
lec_mn_1	11.310	1	11.310	31.942	.000
laverain	27.579	1	27.579	77.891	.000
lavewave	13.181	1	13.181	37.226	.000
laveperiod	6.171	1	6.171	17.428	.000
avewdspXwdcode	11.089	1	11.089	31.317	.000
code	17.451	7	2.493	7.041	.000
Error	298.132	842	.354		
Total	2574.262	855			
Corrected Total	420.866	854			

a R Squared = .292 (Adjusted R Squared = .282)

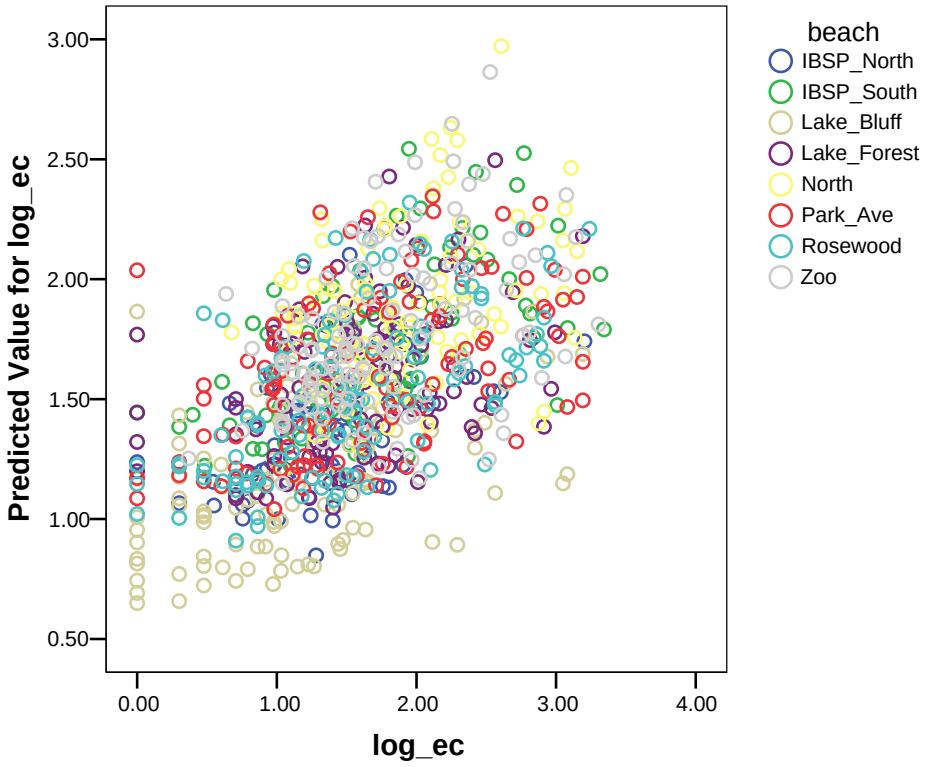


Figure 25. Scatter of measured mean log *E. coli* count and predicted count using best developed model. Colors indicate individual beaches; this group includes Racine, Wisconsin and Lake County, Illinois beaches.

Case Study for a Single Beach

Intensive sampling at 63rd Street Beach in Chicago in 2000 provides a large dataset from which numerous models can be explored. In these examples, the data are examined to find the best model and best approach at modeling a single beach.

Concentrations of *E. coli* measured in 90 cm of water at 13:00 h were used as the dependent variable, while wave height, wind speed, sunlight intensity, rainfall, barometric pressure, wave period, and temperature were available as independent factors. Data were collected from local weather stations, a buoy operated by NOAA, wave measurements from the Army Corps of Engineers and hydrometeorological information from instruments placed at the beach. The data set consisted of 42 measurements made between May and September 2000.

Multiple Regression Model

A multiple regression model, similar to those used for the regional and zone area models, was developed using three major influencing factors: wind speed, sunlight, and wave height. The model explains 60% (R^2=0.597) of the variation log-*E. coli* of the data set. This model resulted in an RMSE=0.429 with 2 (5%) type 1 errors and 0 (0%) type 2 errors. It had the form:

$$y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$$

where B_0 is a constant; y is log-*E. coli* where *E. coli* has units of CFU/100 mL; B_1, B_2, and B_3 are regression coefficients for the predictors wave height, wind speed, and sunlight, respectively; X_1, X_2, and X_3 are the value of each predictor wave height, wind speed, and sunlight, respectively; and e is the residual error of the model. The three factors were chosen using a step-wise multiple regression that maximized the model success as measured by an r-square. The analysis yielded the following equation:

$$y=2.310+0.029(X_1)-0.083(X_2)-2.287(X_3)$$

The positive coefficient for wave height suggests that larger waves give rise to higher *E. coli* concentrations. The negative coefficients for wind speed and sunlight indicate that high winds and high sunlight intensity lead to reduced concentrations of *E. coli*. The model explains 60% (R^2=0.597) of the variation log- *E. coli* of the training data set.

When constructing a multiple regression model, care must be taken not to violate any of the assumptions used in formulating the model. One such assumption requires the dependent variable to not be autocorrelated. The Durbin-Watson statistic can be used to assess this. In the case of the model presented above, the Durbin-Watson statistic is 1.8, suggesting that log-*E. coli* are not autocorrelated. Another assumption requires that the independent variables used to formulate the model are not correlated. The wind speed and sunlight are correlated (r= 0.795, $p<0.05$) suggesting that this assumption has been violated.

Models should be validated with independent data before they are used with confidence. Most empirically derived models tend to be biased because they are based on the data that they seek to predict. Even cross-validation by techniques such as jack-knife procedures are biased since the generated subsets are taken from the raw data itself. Validation is best used on independently collected data. In the case of 63rd Street Beach, the Chicago Park District (CPD) collected *E. coli* at the same transects at 1000 h at 45 cm on our sampling days. Our model was developed for *E. coli* at 1300 h in 90 cm of water. Our 45 cm morning water was higher than comparable CPD samples but both were higher than 90 cm afternoon water. All three were significantly correlated but our samples were more closely correlated to one another than to CPD samples. There were no afternoon CPD samples, so we attempted to validate using CPD 1000 h *E. coli* concentrations using the same independent variables developed for our deeper, later samples.

The coefficient of determination (R square) for CPD morning samples using wind speed, wave height and sunlight was only 0.124 and the model was not significant ($p=0.098$). When rainfall, air temperature and wave height were used, the model improved markedly (R square =0.351, $p < 0.001$). Comparable analysis of our own morning 45 cm data suggested equally poor results ($p=0.078$). Thus, it appears that our attempts to validate the model were incomplete.

The lack of relationship between hydrometeorological conditions and morning *E. coli* concentration is understandable; there was no reason to expect immediate relationships between bacteria counts and ambient conditions. Deeper water had a smoothing effect on *E. coli* variation, which increased the performance of the equation. We conclude that afternoon *E. coli* samples would need to be taken to validate the current model; alternately lagged data or data taken from earlier in the day might yield better morning predictions. Unfortunately, the effects of sunlight on *E. coli* would not be factored into these pre-dawn hours. We argue that predicting the *E. coli* at 1300 h is justified based on the increased water contact at this time and that 90 cm water while not the most conservative estimate is still protective for many of the active swimmers. The inability to extrapolate from 90-cm, afternoon to 45-cm, morning bathing water suggests that routine monitoring data are satisfy validation requirements. That is, the same model is required on completely independent but comparable samples. To this end, either improvement in morning, shallow models are necessary or if early afternoon models are acceptable then new equivalent data are required.

Regression Tree Model

A regression tree algorithm (SYSTAT, Point Richmond, CA) was used to create a classification tree for 63rd Street Beach using the three independent variables used in the multiple regression modeling. The data were sorted into homogeneous subsets using recursive partitioning. At the end of each branch were leaf nodes.

The initial population of 42 observations had a log-mean *E. coli* 1.749 CFU/100 mL. The first branching occured based on wave height. When wave height was below 31 cm, the log-mean was 1.52 CFU/100 mL while above this criteria, the log-mean was 2.33 CFU/100 mL (thus, many of the readings in this subset will be in exceedance of the *E. coli* water quality standard log-*E. coli* equivalent to 2.37 CFU/100 mL). The next branching used a wind speed criteria of 14.4 m/s. When wind speed was below this value and wave height above 31 cm, closures were

predicted as common. Apparently higher offshore winds had a moderating effect on *E. coli* concentrations while slower winds (during increased waves) allowed contaminants to accumulate in the nearshore water. At low wave heights, sunlight was important, but both leaves of the tree were well below closing criteria. This is apparently because of the negative effect of sunlight on the bacteria coupled with enhanced exposure by decreased turbulence, turbidity, and surface conditions. Overall, the proportion reduction in error (equivalent to the R square) for this tree was 47 % and approaches that delivered by traditional linear regression models.

Discriminant Analyses

Often the beach manager is interested not so much in the actual *E. coli* concentration but rather whether the beach is in or out-of compliance with the water quality criteria. For this purpose, a discriminant analysis, which seeks to predict categorical values by finding the best combination of continuous variables, is used. Whether a beach is open or closed/posted depends on the specific policy in place. At 63rd Street Beach, we assumed that a beach with *E. coli*<235 MPN (most probable number)/100 mL should be open, whereas if *E. coli*≥235 MPN/100 mL it should be closed.

A multiple analysis of variance was performed to determine the discriminant functions for the same data set used for the multiple-regression. First we tested the significance of the derived discriminant functions and then classified the data based on those functions. The following set of discriminant function coefficients were highly significant ($p<0.0001$, Wilk's lambda=0.497, df=3).

Using the discriminant function, we were able to predict correctly 34 of the 37 (91.9 %) *E. coli* observations that were within compliance with the single-sample standard and 5 of the 5 (100%) out-of-compliance observations. This model gave rise to 3 (7%) type 1 and 0 (0%) type 2 errors out of 42 testable outcomes. Note that these results are from comparing the model with the observations with which it was trained.

We validated the model by rebuilding with a smaller training data set and validating on the remaining observations. Thirty-three of 37 (89.2%) in-compliance and 5 of 5 (100%) out of compliance observations were correctly predicted. There were 4 (10%) type 1 and 0 (0%) type 2 out of 42 testable outcomes.

DISCUSSION

The regional forecast models developed here provide some insight into the southern Lake Michigan ecosystem for *E. coli*. Because of the high variability of many of the parameters considered due to the wide regional scale, predicting *E. coli* for all locations presented many challenges.

The different regional groupings presented an interesting pattern of beaches. Typically the Chicago beaches were grouped together, as were the Wisconsin and Lake County Illinois beaches and also the Indiana beaches. This type of grouping, in both multidimensional scaling and cluster analysis, can often be explained by regional factors related only to those beaches, which may be the result of nearby *E. coli* input sites, nearshore currents, beach orientation, or other physical or biological factors.

Numerous factors must be considered with *E. coli* counts. Primary source can influence abundance and persistence, and beaches can typically be divided into two types, those dominated by an effluent source and those not directly impacted by an effluent outfall. Those dominated by an effluent source can be periodically subject to inputs with high counts of *E. coli*. In Wisconsin, the Milwaukee River discharges in to Lake Michigan between McKinley and South Shore, and the river outflow carries water with high counts of *E. coli* (McLellan and Salmore, 2003). The Chicago River periodically discharges to Lake Michigan, flowing out near Oak Street, Ohio Street, and 12th Street beaches. Similarly, the Calumet River discharges south of Rainbow Beach and north of Calumet beach periodically, although both the Chicago River and Calumet River typically drain to the west. In Indiana, the Little Calumet River discharges to the east of Ogden Dunes, West Beach, and the Gary Beaches. Smaller natural creeks drain into beach areas, most notably Dunes Creek, which empties into Lake Michigan between State Park east and west sites. To the east, Trail Creek empties into the lake just east of Mount Baldy. The influence of these outfalls is variable, but typically, their influence would be stronger during rain events. Rain events rarely equally affect the entire region of study.

In situations in which there is no direct effluent or the effluent is not having an impact, predicting *E. coli* counts can be more difficult because the sources may be diffuse and differently affected by water and weather conditions. *E. coli* counts are less likely to fluctuate similarly at beaches far from a direct source, which makes it difficult to use the same model for numerous beaches. In the model presented here, all beaches were tested using the devised model, and for each of them at least one of the parameters had no significant impact on its model.

Parameters Used in the Model

Although current monitoring protocols typically rely on *E. coli* counts from the previous day for deciding whether to close a beach, this relationship has been widely criticized and undermined. However, when considered with other factors, it appears to have some predictive capability. Both previous day's *E. coli* and prior moving average of *E. coli* for three days were stronger parameters included in the models examined. Because of the highly variable nature of *E. coli* in natural water, taking a single sample for monitoring and applying it to the next day's beach closure decision results in poor reliability overall. However, with a larger and denser

dataset, patterns within the *E. coli* can be discriminated. The noticeable seasonal pattern of *E. coli* in these data indicates that some day to day relationship exists. Many monitoring programs do not collect daily samples, so the relationship is hidden in data gaps.

Rainfall was a major parameter in the models developed, indicating the importance of effluent, outfalls, and runoff in driving *E. coli* counts in southern Lake Michigan. Rainfall has been widely recognized as associated with high *E. coli* levels. This relationship can be attributed to both runoff and direct sewage input. During rainfall, *E. coli* can be washed from numerous surface sources and transported to draining waters and subsequently to Lake Michigan. Additionally, some sewage treatment plants that have combined sewage and stormwater systems (e.g., Milwaukee, much of Indiana) are permitted to bypass treatment during heavy rains, resulting in the release of untreated sewage into draining waters and ultimately Lake Michigan. There is often a delayed impact, and beach waters experience high counts 24-72 hours after initial rain event (Whitman et al., 1999; Haack et al., 2003).

Wave height is often associated with rainfall or wind direction but typically interacts with nonpoint sources of *E. coli* on the beaches. Wave height and wave period were effective predictors in the model presented. High waves are more associated with onshore winds and result in resuspension of nearshore and onshore sand and sediments. These sediments can harbor *E. coli* several orders of magnitude higher than what is in the beach water, and during high waves, the beach sand can act as a source of *E. coli* to the water (Whitman and Nevers, 2003).

An interactive term was used in the model that incorporated both wind speed and wind direction. Wind direction is commonly used to divide datasets because it has such a strong influence on *E. coli* counts in similar ways as wave height. Onshore winds are generally associated with higher *E. coli* counts because they increase the swash zone and therefore the amount of *E. coli* washed from the sand and sediment into the lake. In several approaches presented here, models were divided by wind direction, and the results showed different parameters being more closely related to *E. coli* counts depending on wind direction.

Other parameters considered in the models were depth of Calumet Harbor, which could be a direct correlate of rainfall or onshore winds, and minimum daily temperature, which may be directly related to seasonality in *E. coli* counts (i.e., higher *E. coli* later in the summer).

Grouping Beaches

For the entire region, the best model developed was capable of explaining the variance in *E. coli* 29% of the time. Using the same model on pre-determined zones, that number ranged from 12-29%, but when the beaches were divided based on general *E. coli* trends, that number increased to 31%, for the Chicago beaches. The beach groupings generally fell along geographical lines, with some mixing between Wisconsin and northern Illinois. These groupings were seen in both the multidimensional scaling and cluster analysis and may result from impacts and effects experienced on a smaller scale than the region examined here.

If beaches were considered individually, the R2 increased to as high as 0.474 (Illinois Beach State Park South), but low values were also abundant (e.g., R^2=0.09 for North Point Marina, R^2=0.20 for Jarvis/Fargo). More than half of the beaches included, however, had R2 values of 0.3 or higher, which makes them superior to the overall regional model. Using individual beaches, however, several parameters were no longer significant in contributing to the model; therefore models would have to be individually assessed for each of these beaches.

Model Effectiveness

In the current monitoring protocol, a water sample is collected one day, analyzed for *E. coli*, and when the results are read the following day, a determination is made whether to close the beach. Due to the high variation from day to day, the decision often results in closing a beach that has low counts or keeping a beach open that has high counts. More timely results are necessary to protect public health and to provide maximum recreation opportunities. Predictive modeling may offer a realistic solution, and a predictive model that can be used for numerous beaches would help expand the application.

In this study, the regional model developed could account for 29% of the variation in *E. coli* counts. As low as the predictor result is, it is still superior to the 19% of variance explained using the current monitoring approach of sampling the water on day 1 to determine whether to close it on day 2. This low result is due to the numerous types of beaches included in the model and also the wide extent over which the beaches are spread. Because weather and water conditions may vary highly from the northern to the southern portions of the study area, using data for a single parameter, collected in a single location, will rarely incorporate conditions throughout the study region.

Subdividing the beaches by type improved the model results for several locations, but in such cases, several parameters were no longer significant to the overall equation. Separate models could be developed for certain groups of beaches, but that would obviously limit the scope of the application.

Model Validation

We evaluated the performance of the overall regional forecast modeling effort by calculating respective Root Mean Square Error (RMSE) statistics. RMSE is simply the mean of the sum of the squared differences between the actual (measured) *E. coli* reading and the *E. coli* predicted by the model. The square root of the mean is calculated in order to yield the same units as the original data set. The RMSE is conceptually equivalent to the standard deviation of the residuals of each model. A larger value indicates an inferior performance by the model. While R^2 relates to the proportion of the variance that is explained by the model, the RMSE provides information on the performance or reliability of the model. RMSE is inversely proportional to R^2. When all 55 beaches across all four years are combined, the RMSE for the linear regression model developed in the present report (Table 10) is 0.709 (log *E. coli*/100ml); mean log *E. coli* for the same dataset was 1.73 (SD=0.77). For the same data set, the RMSE for conventional EPA model would be 0.843. The EPA model is merely that yesterday's *E. coli* equals that of today, or $EPA_{t-1} = EPA_t$. Thus, the overall regional model is about 14% better in performance than the EPA model, but that improvement is made only with the additional expenses and observations needed to run the model. Most of the parameters are weather-related and thus with advent of

internet portals, these hydrometeorological factors can be automated to an extent to minimize these costs.

The relative performance of the EPA and Regional models can be illustrated by a paired plot of residuals (Figure 26). Both residual plots should approximate a normal curve with a mean of 0. The narrower the curve around the mean, the better the performance (a perfect fit would have no variance and be represented by a vertical line at 0).

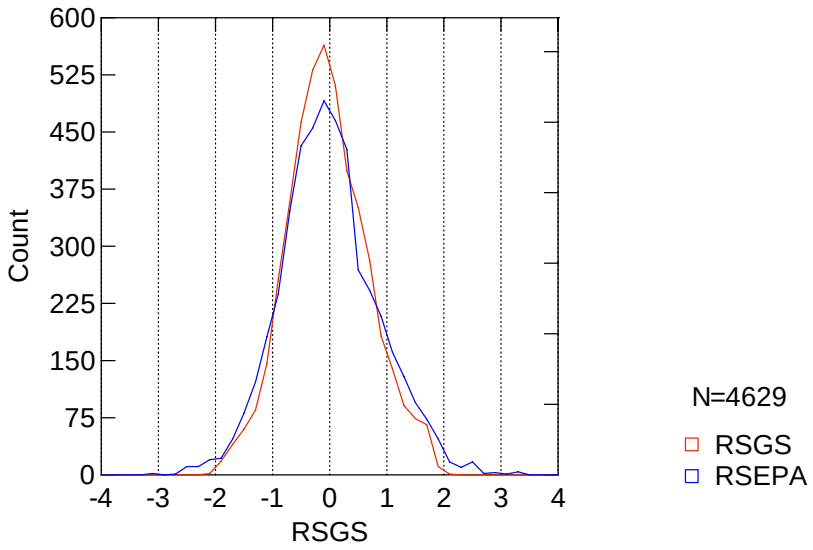


Figure 26 Residuals for EPA (RSEPA) and regional (RSGS) models for determining *E. coli* count.

Both the EPA and Regional plots are similar but it is apparent that the regional model has more values closer to the observed ('true') values and that the regional model hugs the origin slightly better than the conventional model.

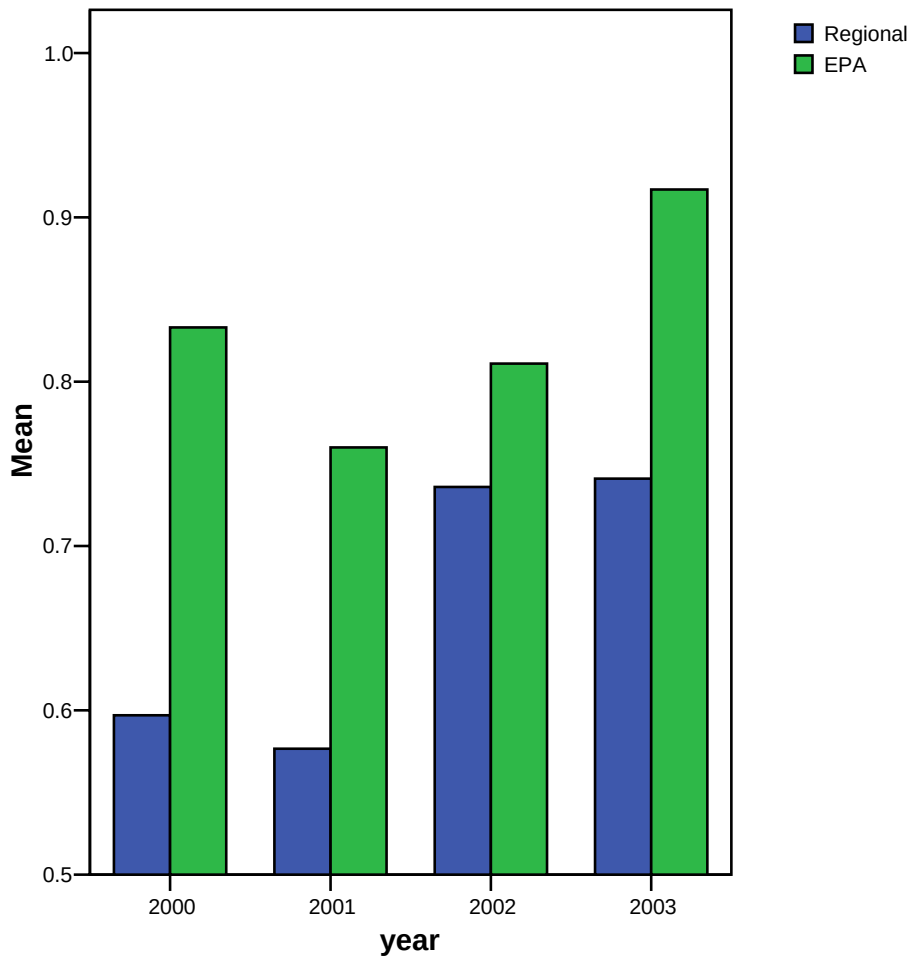


Figure 27 RMSE (root mean square of the error) values by year for EPA and regional (USGS) models of determining *E. coli* count.

Calculation of the RMSE is likely a less biased assessment of the performance of the respective models. For this exercise, we left out one year of observations, created our model on the remaining three years and used that formula to predict the outstanding year. This process was repeated to validate each year of data, so several permutations were calculated (i.e., leave out 2000; leave out 2001; etc.) The same parameters used in the regression equation (Table 10) were included in all cross-validated analyses. Figure 27 clearly shows lower error variance associated with each of the four years using the USGS over the EPA model. Cross-validated RMSE differences varied from 0.236 in 2000 to 0.075 log *E. coli* (CFU/100 ml) in 2002. The best (lowest) RMSE was in 2001 or 0.597, the mean log *E. coli* for that year was 1.55, N=1060, S.D.= 0.740.

The results of validation analyses indicate that the regional model approach is superior to the currently used EPA monitoring protocol. The incremental increase in performance using the regional model, however, must be weighed against the need for numerous parameters and the applicability across such a large region. The results indicate, however, that there are certain factors that contribute to the overall fluctuations in E. coli counts at the 55 beaches studied.

REFERENCES

Boehm, A.B., Grant, S.B., Kim, J.H., Mowbray, S.L., McGee, C.D., Clark, C.D. et al. (2002) Decadal and shorter period variability of surf zone water quality at Huntington Beach, California. Environmental Science and Technology **36**, 3885-3892.

Francy, D.S., and Darner, R.A. (2002) Forecasting bacteria levels at bathing beaches in Ohio. In: US Department of the Interior, US Geological Survey.

Haack, S.K., Fogarty, L.R., and Wright, C. (2003) *Escherichia coli* and enterococci at beaches in the Grand Traverse Bay, Lake Michigan: Sources, characteristics, and environmental pathways. Environmental Science and Technology **37**, 3275-3282.

McLellan, S.L., and Salmore, A.K. (2003) Evidence for localized bacterial loading as the cause of chronic beach closings in a freshwater marina. Water Research **37**, 2700-2708.

Nevers, M.B., and Whitman, R.L. (In Review) Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan.

Olyphant, G.A., and Whitman, R.L. (2004) Elements of a predictive model for determining beach closures on a real time basis: The case of 63rd Street Beach Chicago. Environmental Monitoring and Assessment **98**, 175-190.

Olyphant, G.A., Thomas, J., Whitman, R.L., and Harper, D. (2003) Characterization and statistical modeling of bacterial (*Escherichia coli*) outflows from watersheds that discharge into southern Lake Michigan. Environmental Monitoring and Assessment **81**, 289-300.

Whitman, R.L., and Nevers, M.B. (2003) Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach. Applied and Environmental Microbiology **69**, 5555-5562.

Whitman, R.L., Nevers, M.B., and Gerovac, P.J. (1999) Interaction of ambient conditions and fecal coliform bacteria in southern Lake Michigan waters: Monitoring program implications. Natural Areas Journal **19**, 166-171.