Available online at www.sciencedirect.com

ScienceDirect

Environmental Research 103 (2007) 358–364

Environmental Research

www.elsevier.com/locate/envres

Some statistical issues related to multiple linear regression modeling of beach bacteria concentrations[☆]

Zhongfu Ge[a,*], Walter E. Frick[b]

[a]National Research Council, U.S. Environmental Protection Agency, 960 College Station Road, Athens, GA 30605-2720, USA
[b]U.S. Environmental Protection Agency, Ecosystems Research Division, 960 College Station Road, Athens, GA 30605-2720, USA

Received 7 June 2006; received in revised form 4 November 2006; accepted 13 November 2006
Available online 26 December 2006

Abstract

As a fast and effective technique, the multiple linear regression (MLR) method has been widely used in modeling and prediction of beach bacteria concentrations. Among previous works on this subject, however, several issues were insufficiently or inconsistently addressed. Those issues include the value and use of interaction terms, the serial correlation, the criteria for model selection, and model assessment. The present work shows that serial correlations, as often present in sequentially observed data records, deserve full attention from the modeler. The testing and adjustment for the time-series effect should be implemented in a statistically rigorous framework. The R^2 and Cp-statistic as joint criteria are recommended for the model selection process, while using the t-statistics associated with the full model is erroneous. During model selection, using interaction terms can often help to decrease the bias in reduced models, although the resulting improvement in the numerical performance may be limited. For the assessment of the model predictive capacity, which is different from testing the goodness of fit, a comprehensive set of statistics are advocated to allow for an objective evaluation of different models. Results obtained from the data at Huntington Beach, OH, show that erroneous conclusions could be drawn if only the model R^2 and the count of type I and type II errors are considered. In this sense, several previous works deserve further investigation.

Keywords: Multiple linear regression; Empirical; Model selection; Model evaluation; Prediction

1. Introduction

Fecal pollution and microbial contamination pose a health threat to bathers at recreational beaches every year. To protect public health, beach personnel are responsible for measuring fecal indicator bacteria (FIB) regularly to assess water quality. Current methods for assessing recreational water quality required by the United States Environmental Protection Agency (USEPA) are based on concentrations of *Escherichia coli* (*E. coli*) or enterococci. It typically takes at least 18 to 24 h to analyze the samples for FIB. This means that, even if the bacteria concentration is excessive in the water samples, the problem is not identified until the next day. In contrast to the processing time, beach conditions such as sunlight and wind direction often change over smaller time scales (e.g. Leecaster and Weisberg, 2001). The current biological method of analysis alone cannot capture the trend of beach bacteria concentrations to help give timely advice on beach closures.

Another problem lies in the complexity of the physics. Deterministic modeling of beach bacteria concentrations requires comprehensive understanding of the complex physical, biological, hydrodynamical, and meteorological factors. Despite previous successful efforts in numerical simulation of the Great Lakes (Beletsky et al., 1997; Chen et al., 2004), many physical mechanisms remain unclear. Even after the methodology is sufficiently developed, its application will demand extensive resources, ranging from

[☆]This work was partially supported by an interagency agreement between the NOAA Ocean and Human Health program and USGS, Columbus, Ohio through the USGS Michigan Water Science Center, Lansing, Michigan.

*Corresponding author. Fax: +1 706 3558302.
E-mail addresses: ge.zhongfu@epa.gov (Z. Ge), frick.walter@epa.gov (W.E. Frick).

0013-9351/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.envres.2006.11.006

real-time meteorological input to detailed beach geometry and bathymetry. Customizing a model to a particular beach may be as challenging as the modeling itself. Therefore, highly accurate and practical deterministic models are not expected to emerge in the near future.

Because of these problems, statistical (empirical) approaches have received more attention during the past decade, and most of those are based on the multiple linear regression (MLR) method. From early attempts of ordinary MLR to more sophisticated recent applications, MLR has proven to be an effective tool for quick predictions of bacteria concentrations. Nevertheless, different authors tend to understand and treat some critical statistical issues in different ways, some of which seem to need further explanation and justification. For example, issues such as the necessity of considering interaction terms of explanatory variables, the testing and adjustment for serial correlations, the criteria for model selection, and model assessment are most inconsistently addressed and treated. Some inappropriate treatment might result in loss of robustness of the models. This work aims to give comments on several recently published papers concerning MLR modeling of beach bacteria concentrations, and to recommend a more rigorous framework. Discussions are illustrated using the data collected in the swimming seasons from 2000 to 2004 at Huntington Beach, Ohio.

2. Study site and explanatory variables

Used as a case study, the data of *E. coli* concentrations and explanatory variables for Huntington Beach, Bay Village, Ohio, were obtained from the U.S. Geological Survey (USGS) as public information. These data were collected 3–5 days a week during the recreational seasons (mid May through early September) of 2000–2004 by the Cuyahoga County Board of Health as part of their beach monitoring program. Detailed data collection procedures are described in Francy et al. (2003).

The four measured ambient variables used in this work include turbidity (TB), wave height (WH), antecedent 24-h rainfall (RF) and wind direction (WD). Categorized wave heights were obtained from the Great Lakes Forecasting System (GLFS), Ohio State University, on the Rocky River, which is about 6 mi to the east of the beach. Wind direction and rain data were measured at the weather station at Hopkins International Airport in Cleveland, Ohio, about 8 mi to the southeast of the beach. Further details may be found in the USGS report (Francy et al., 2003). In the following discussion, wind direction is further categorized as 0 or 1 to roughly represent northerly (onshore) and southerly (offshore) winds. (Categorization based on alongshore directions was also attempted, but was found to be less effective.)

The set of variables in the Huntington Beach data cannot be expected to encompass all major agents that determine the local bacteria concentrations. Based on previous studies, it is obvious that a wide variety of variables can influence the bacteria concentration at a beach. They include, but are not limited to, precipitation, wave height, wind, turbidity, sunlight, weather events, tidal currents, zooplankton, birds, and human activities. Ideally, the modeler should be familiar with the potential impacts of candidate variables on the transport and fate of the bacteria concentration, and select a balanced set of them to avoid misrepresenting the variation of the bacteria concentration. For the present case, for example, sunlight might be a complementary variable that would help to further improve models. However, for the purpose of demonstrating the statistical issues, the authors believe that the data available are reasonably satisfactory, although the resulting models are not intended to be used for future prediction on Huntington Beach.

3. Some statistical issues

An MLR model is typically expressed as

$$E[\log(\mathrm{EC})] = \beta_0 + \sum_{i=1}^{p} \beta_i x_i, \tag{1}$$

where $E[\cdots]$ represents the mean of a random variable, x_i and β_i denote the explanatory variables and the regression coefficients for the constant and the variables, respectively, and p is the number of explanatory variables. To achieve uniform (at least visually uniform) variances for all explanatory variables, as required by the assumptions of the least-squares method, the variables should be properly transformed when necessary. In case of unequal spreads observed from the scatter-plot of the response, log(EC), vs. a particular variable, the Box–Cox transformation, defined as $f(x) = (x^{\lambda} - 1)/\lambda$, can be employed for adjusting different patterns of data distribution. Particularly, when $\lambda \to 0$ and $\lambda = \frac{1}{2}$, the transform approaches natural-logarithmic and square-root functions, respectively (Box and Cox, 1964). One may also try other functions by tuning the parameter λ, although natural-log and square-root functions are the most familiar and widely used ones. In the four primary variables for the Huntington Beach case, TB is thus natural-log transformed and the square-root of RF is used. For a more comprehensive full model, additional interaction terms are added as derived explanatory variables: TB · WH, TB · RF and WH · RF. (To avoid further loss of numerical precision, interactions with WD, which has been categorized, are not considered.) After proper transformations, seven variables are finally identified to constitute the full model: log(TB), WH, sqrt(RF), WD, log(log(TB)·WH), sqrt(log(TB)·sqrt(RF)), and sqrt(WH·sqrt(RF)). They are still referred to as TB, WH, RF, WD, TB·WH, TB·RF, and WH·RF, respectively, in the following discussions. The full MLR model for the entire data set (2000–2004, sample size $N = 247$) and the associated statistics are listed in Table 1. The results are discussed in detail in Section 3.2.