Journal of Environmental Management 206 (2018) 910–919

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Evaluating statistical model performance in water quality prediction

Rodelyn Avila [a, b, *], Beverley Horn [b], Elaine Moriarty [b], Roger Hodson [c], Elena Moltchanova [a]

[a] School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[b] Institute of Environmental Science and Research, ESR, PO Box 29181, Christchurch 8540, New Zealand
[c] Environment Southland, Private Bag 90116, Invercargill 9840, New Zealand

ARTICLE INFO

Article history:
Received 29 June 2017
Received in revised form
19 October 2017
Accepted 19 November 2017

Keywords:
Water quality prediction
E. coli
Statistical models
Bayesian networks

ABSTRACT

Exposure to contaminated water while swimming or boating or participating in other recreational activities can cause gastrointestinal and respiratory disease. It is not uncommon for water bodies to experience rapid fluctuations in water quality, and it is therefore vital to be able to predict them accurately and in time so as to minimise population's exposure to pathogenic organisms. *E. coli* is commonly used as an indicator to measure water quality in freshwater, and higher counts of *E. coli* are associated with increased risk to illness. In this case study, we compare the performance of a wide range of statistical models in prediction of water quality via *E. coli* levels for the weekly data collected over the summer months from 2006 to 2014 at the recreational site on the Oreti river in Wallacetown, New Zealand. The models include naive model, multiple linear regression, dynamic regression, regression tree, Markov chain, classification tree, random forests, multinomial logistic regression, discriminant analysis and Bayesian network. The results show that Bayesian network was superior to all the other models. Overall, it had a leave-one-out and *k*-fold cross validation error rate of 21%, while predicting the majority of instances of *E. coli* levels classified as unsafe by the Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas 2003, New Zealand. Because Bayesian networks are also flexible in handling missing data and outliers and allow for continuous updating in real time, we have found them to be a promising tool, and in the future, plan to extend the analysis beyond the current case study site.

1. Introduction

Degraded water quality can be harmful to human health. Moreover, exposure to contaminated water via recreational use including swimming can result in individual illness and community outbreaks of gastrointestinal and respiratory disease (Fewtrell and Kay, 2015; Bridle, 2014; Soller et al., 2010; Yoder et al., 2008; Prüss, 1998). A consequence of these outbreaks can put unwanted pressure on health services and lead to financial losses both to the individual households, the regional and national economy (Bridle, 2014; Hunter et al., 2009; Given et al., 2006; Gleick, 2002). For these reasons, regulatory authorities manage risk by establishing guidelines for water quality to be monitored by responsible authorities.

The microbiological quality of recreational water is monitored via the presence of indicator bacteria. Annette Pruss reviewed 37 epidemiological studies on health effects from exposure to recreational water, and found that most studies reported a positive statistically significant association between the indicator-bacteria count in recreational waters and health risk in swimmers (Prüss, 1998). For freshwater, the indicator microorganisms that correlate best with health outcomes were *Escherichia coli* (*E. coli*), which is a type of fecal coliform that is used to measure the level of pollution (Odonkor and Ampofo, 2013). The presence of *E. coli* in recreational waters indicates fecal contamination which coincides with the presence of pathogenic microorganisms. Another systematic review of over 900 studies by (Wade et al., 2003) found that *E. coli* was a more consistent predictor of gastrointestinal illness than enterococci and other bacterial indicators. Although the result was not statistically significant, they found that a log (base 10) unit increase in *E. coli* count was associated with an average 2.12 (95% CI, 0.925, 4.85) increase in relative risk in fresh water. Since *E. coli* is

* Corresponding author. School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand.
E-mail address: rodelyn.avila@pg.canterbury.ac.nz (R. Avila).

https://doi.org/10.1016/j.jenvman.2017.11.049
0301-4797/

found in all mammal and bird faeces, higher concentrations mean an increased risk of presence of other pathogens (Sampson et al., 2006; Winfield and Groisman, 2003; Edberg et al., 2000).

To ensure the risk from recreational water is minimised for the public, many governments and groups have implemented water quality standards, such as the WHO Guidelines for Safe Recreational Water Environments (World Health Organization, 2003) and the revised European Union Bathing Water Directive 2006. These regulatory tools require recreational sites to be monitored with a minimum of one monthly sample taken during the bathing season with the results of the monitoring then disclosed to the public. The responsible government must then describe their risk management measures in relation to predictable short term pollution or abnormal events (European Parliament, 2006).

Freshwater management units (FMUs) are fresh water catchments that have been set up by New Zealand regional councils in order to set freshwater objectives and limits for freshwater quality. FMUs can be grouped according to their physical characteristics as well as their social significance, i.e. who are their main users and what purpose are they used for (Ministry for the Environment, 2015). In New Zealand, the Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas 2003 outlines the acceptable water quality for locations (FMU) designated for recreational use, where surveillance of water quality is carried out on a regular basis. These guidelines state the degree of surveillance required and if public disclosure of the water quality is required to be given based on a surveillance mode; Acceptable, Alert and Action (Green, Amber and Red). These modes are assigned to each location based on the reported *E. coli* concentration, see Table 2 (Ministry for the Environment, 2002). Acceptable/Green is defined to be generally safe for activities such as swimming and to continue routine surveillance. Alert/Amber means an increase in *E. coli* levels and sampling to be done on a daily basis and to refer to the Catchment Assessment Checklist (CAC), which is included in the aforementioned guide, to assist in identifying possible location(s) of sources of fecal contamination. Action/Red means that high levels of *E. coli* have been found and there is an increased risk to infection. The associated action plan for mode Alert/Red required to be undertaken follow the same steps as Alert/Amber with the addition of a sanitary survey with a report on sources of contamination, warning signs erected and public disclosure of a public health problem. Hence, it is especially important to distinguish Red days from the others.

Given the importance of recreational water quality, it is important not only to monitor it, but also to predict it. This is to ensure that the public can be given a timely warning of the possible contamination and the ensuing disease burden and economical loss can be avoided. This task is complicated by the fact that the water quality is influenced by a variety of factors such as seasonal changes, land-use, human activities, and extreme weather events (Kang et al., 2010; McDowell and Wilcock, 2008; Muirhead et al., 2004, 2006). It is also somewhat complicated by defining the optimal decision, and looking for a balance between false positives (warning of contamination when there is none) and false negatives (failing to spot contamination). The cost of misclassifying mode Green into Amber or Amber into Green is not as severe as these modes allow for recreational activities to be carried out. However, the misclassification of Red into Amber or Red into Green etc. should be treated seriously as it can result in severe illness.

In the past, a variety of statistical models have been used to predict water quality. Regression trees have been used to predict bathing suitability throughout Scotland (Stidson et al., 2012), and by Džeroski et al. (2000) for water quality prediction in Slovenian rivers. Discriminant analysis has been used to evaluate the spatial and temporal variations of water quality in the Gomti River, India Singh et al., 2004, and similarly in the Fuji River Basin (Shrestha and Kazama, 2007). Bayesian networks have also been used in water quality management: Ha and Stenstrom 2003 used a Bayesian network to identify the origins of storm water based on land use; and by Donald et al. (2009) to determine the risk of gastroenteritis from recycled water. The use of multiple regression models have also shown that heavy rainfall increases pollutant load (Maniquiz et al., 2010) and urban areas tend to decrease downstream water quality (Mallin et al., 2016). Moreover, Thoe et al., 2014 wanted a model to predict water quality at Santa Monica Beach that would perform better than the naive model that was used at the time. They compared model performance between five statistical models; multiple linear regression, logistic regression, partial least squares regression, artificial neural networks and classification tree and found that the all the statistical models performed better than the existing method.

The objective of this study was to find a model that could predict future *E. coli* counts or water quality modes based on preceding data in the same season or year. This prediction would be based on past values of *E. coli* counts, accumulated rainfall of a monitored upstream site in the past 48 h and river flow. The results of this study provides a basis for model suitability for real time prediction for bathing sites across Southland, New Zealand. The proposed model should be able to correctly identify mode Red days or predict higher levels of *E. coli* concentrations. An additional benefit would ideally show how the inputs and their varying levels affect water quality. This could aid in policy decisions and allow the public to better asses the level of risk in regards to recreational water use. In this case study, we apply a variety of statistical models, including log-linear regression model, logistic regression model, discriminant analysis, regression trees, random forests and Bayesian networks to predict water quality for the summers 2005–2014 for the Oreti river in Wallacetown, which is a recreational water site situated in Southland, New Zealand. The response variable, *E. coli* concentration, is treated both, as continuous counts and as categorical variable with modes Green, Amber and Red. The predictive power of each model is assessed using cross-validation and conclusions are drawn about the best practice.

2. Study site and data

The study site is situated on the Oreti River in Wallacetown, Southland New Zealand (see Fig. 1). The Oreti river in Wallacetown is a location which is identified as being of value for recreational use and is known to experience degraded water quality (Environment Southland, 2010; Environment Southland and Te Ao Marama Inc, 2010). The land use surrounding the area consists of dry stock (42%), natural state (32%), dairy farming (18%), forestry (7%) and other uses (1%). In addition, the Winton WWTP processes wastewater from the small town of Winton, the discharge is into a tributary of the Oreti River, the Winton Stream which is approximately 6 km upstream of the confluence and 23 km up stream of the Wallacetown monitoring site (Pearson and Couldrey, 2016).

These observations are for the summer months between December and April when recreational use is expected to occur see Table 1. There is variation in sample size (n) between years due to occasional missing weekly measurements. As water quality mode is derived directly from the *E. coli* counts, we can either model the reported *E. coli* concentration or the corresponding mode. These modes and their cut-off points are given in Table 2.

The data set consists of weekly measurements of *E. coli* MPN counts based on a single sample, water quality mode which is derived from *E. coli*, river flow (m^3/s) and rainfall data (mm). The *E. coli* counts were calculated using the Quantitray MPN method