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/© 2017 Elsevier Ltd. All rights reserved.

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

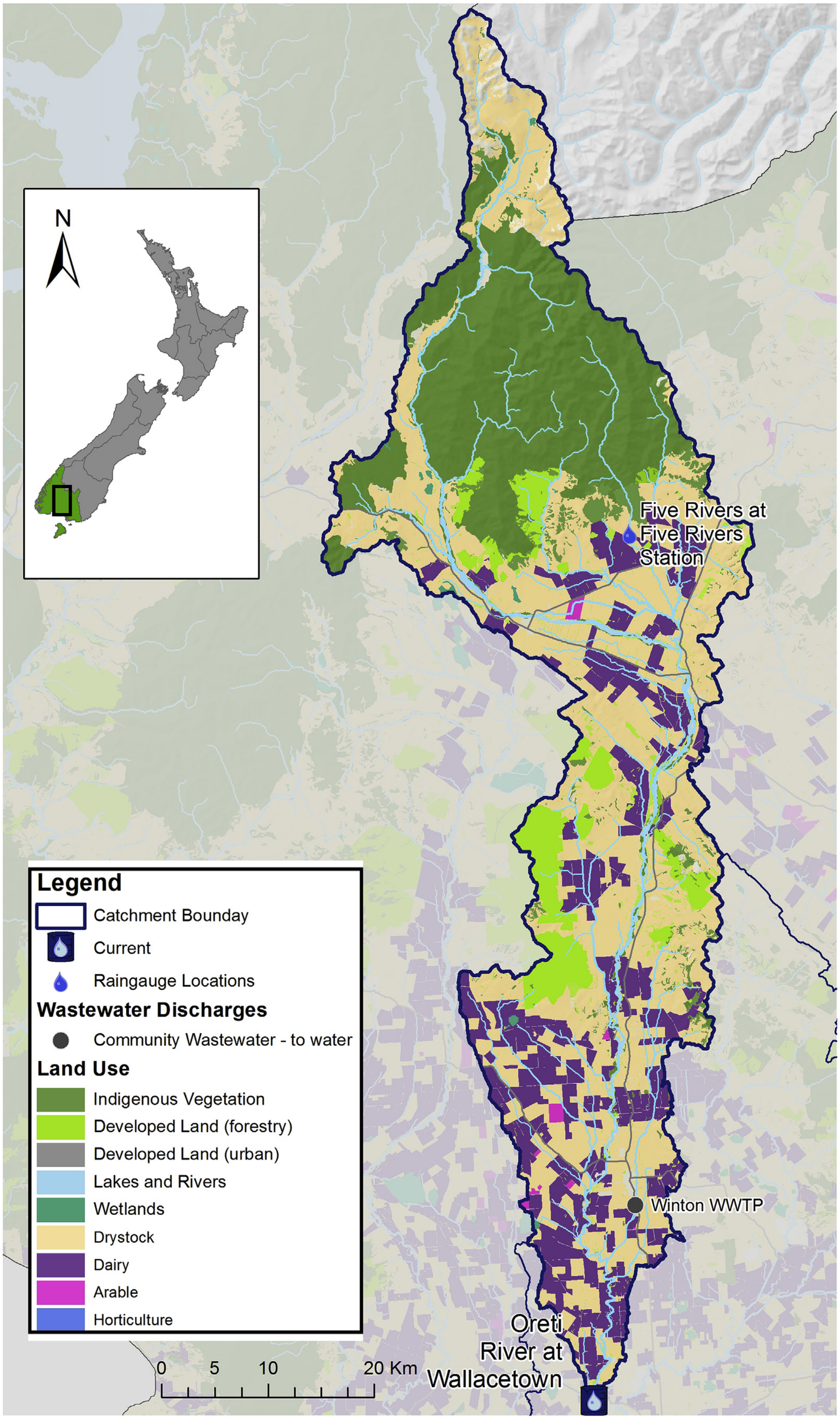

Fig. 1. A map of the Oreti River showing the study site of Wallacetown, rainfall gauge location (Five Rivers Station), waste water treatment plant (Winton) and surrounding land use.

Table 1
Frequency distribution of weekly observations of recreational water quality in the bathing seasons of 2005–2014. The modes range from Green for generally safe recreational use to Red for increased risk of infection.

Year	Mode Observed			n
	Green	Amber	Red	
2005–2006	10	4	4	18
2006–2007	10	3	2	15
2007–2008	13	4	1	18
2008–2009	16	2	1	19
2009–2010	14	2	1	17
2010–2011	11	0	5	16
2011–2012	13	2	1	16
2012–2013	13	2	0	15
2013–2014	14	2	1	17
Total	114	21	16	151

Table 2
Guidelines for water quality modes determined by *E. coli* concentrations as set by the Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas 2003, Ministry of Environment, New Zealand.

mode	*E. coli* MPN/100 mL
Green	≤ 260
Amber	>260 and ≤ 550
Red	>550

and river flow was taken at the Oreti River in Wallacetown with the water level sensor at 2 mm accuracy measured every 10 s. The rainfall data which consisted of three different types of measurements was taken from a rain gauge from the 5 Rivers Station which is a connected upstream water body. It consisted of rainfall, past 24 and 48 h rainfall. The rainfall variable was measured when 0.5 mm of rain accumulated in a tipping bucket gauge. The past 24 and 48 h rainfall is the cumulative rainfall in the 24 or 48 h prior to the time of *E. coli* sample collection.

2.1. Methodology

It is known that the normal distribution can be used to approximate the Poisson distribution for large values of λ, where λ is the mean of the Poisson distribution. Therefore, if $X \sim Poisson(\lambda)$, then for large values of λ, $X \sim N(\lambda, \lambda)$ approximately (Peizer and Pratt, 1968; Cheng, 1949). Therefore, for this analysis, the reported *E. coli* MPN counts will be treated as continuous. In order to clearly distinguish between the situations when the water quality is modelled as a continuous variable from those when a categorical response is used, Y_t is used to denote continuous *E. coli* counts and Z_t to denote the corresponding modes, $Z \in \{G, A, R\}$, where G is green, A is amber and R is red. The continuous response models include naive model, multiple linear regression, dynamic regression and regression tree. Categorical responses were modelled using Markov chain, classification tree, multinomial logistic regression, discriminant analysis and Bayesian network. Although the water quality modes are ordered categories, ordinal multinomial logistic regression was not used as the effect of the predictors varied across the modes, therefore violating the proportional odds assumption (Agresti, 1996). The details of these methods are further described in this section 2.1.1.

The analysis was carried out in R, where the following packages were used to fit the appropriate model. For the regression and classification tree **rpart** was used (Therneau et al., 2015), with the random forest fitted via the **randomForest** package (Liaw and Wiener, 2002). The Markov chain was fitted using the **markovchain** package (Spedicato, 2015), and to fit the multinomial logistic regression **nnet** was used (Venables and Ripley, 2002a). Discriminant analysis was carried out using **MASS**(Venables and Ripley, 2002b) and the rest were implemented in base **R** (R Core Team, 2015).

2.1.1. Continuous response

2.1.1.1. The naive model. The naive model uses the logic that tomorrow will be the same as today. Here, the previous week's *E. coli* measurement and mode is used as the current week's prediction.

$$E(Y_t|y_{t-1}) = y_{t-1}. \tag{1}$$

This model provides a benchmark to judge the other models against.

2.1.1.2. Multiple linear regression and dynamic regression. In the multiple linear regression model the response is linearly related to a set of independent variables (Draper and Smith, 1998).

$$E(Y_t|X_t) = X_t\beta, \tag{2}$$

where X_t is a matrix of the observed variables and β is a vector of regression coefficients.

To introduce a dynamic aspect into the model, the past *E. coli* and flow levels can be included as well (Fabaozzi et al., 2006).

2.1.1.3. Regression trees. Tree-based methods partition the variable space into a set of rectangles, and then fit a model (in this case a simple linear regression) in each one Hastie et al., 2009. While there are issues with their inherent instability and lack of smoothness, tree based models often provide a simple yet powerful tool for modelling and prediction.

2.1.1.4. Random forest: regression. To address the instability of a single regression tree random forest's can be used. For regression, the same regression tree is fitted many times to bootstrap sampled versions of the training data and averages the result (Hastie et al., 2009).

2.1.2. Categorical response

2.1.2.1. Markov chain. The Markov Chain is similar to the naive model in a sense that the expected value of a stochastic process depends on the immediate past. The probability of moving from mode i to mode j, from one day to the next, is called a transitional probability, denoted p_{ij}, and is estimated from the data (Freedman, 1971). For the first order Markov Process, only the previous observation matters, and the predicted state at time t, given the observation at the previous moment $t-1$ is given by the mode of the conditional distribution $P(Z_t|Z_{t-1} = i)$, i.e., j^* such that $p_{ij^*} = \max_j p_{ij}$.

2.1.2.2. Multinomial logistic regression. The multinomial logistic regression is an extension of logistic regression when more than two outcomes are possible (Hastie et al., 2009). In order to ensure that the probabilities of all the possible outcomes add to one, the link function takes the following form:

$$P(Z_t = z|X_t) = \frac{\exp(\eta_{tz})}{1 + \sum_{z=2}^{Z} \exp(\eta_{tz})}, \tag{3}$$

where $\eta_{tz} = X_t\beta_z$ with $z = 2, \ldots, Z$ and the predicted outcome is $\max P(Z_t = z|X_t)$. The first category, $z = 1$, is called the baseline category, and $\eta_{tz} = 0$. As in the Markov model, the predicted state at time t is given by the mode of the conditional distribution $P(Z_t = z|X_t)$, i.e. mode z^* such that $P(Z_t = z^*|X_t) = \max_z P(Z_t = z|X_t)$.

2.1.2.3. Discriminant analysis. Linear Discriminant Analysis (LDA) uses a linear combination of variables to distinguish between classes resulting in linear decision boundaries. The independent variables across the classes are assumed to be multivariate normal with a common variance-covariance. If the variance-covariance cannot be assumed equal, a modification known as quadratic discriminant analysis (QDA) is used instead (Hastie et al., 2009).

2.1.2.4. Classification trees. Classification trees are similar to regression trees: the variable space is partitioned into a set of rectangles and the most likely outcome is assigned to each. If the associated probability of an outcome assigned to a node is 1.0, the node is known as pure. Various statistics can be used to measure node purity, including misclassification error, Gini index, and cross-entropy of deviance (Hastie et al., 2009). The result can then be conveniently represented by a dendrogram.

2.1.2.5. Random forest: classification. The random forests for classification trees are similar to the regression random forest. However for classification problems, a committee of trees each cast a vote for the predicted class (Hastie et al., 2009).

2.1.2.6. Bayesian networks. A Bayesian network (BN) is a graphical model that encompasses probabilistic relationships amongst a set of variables (Fenton and Neil, 2012). A BN can be represented graphically by a directed acyclic graph (DAG), with the nodes corresponding to the variables of interest and arcs (directed edges) corresponding to the perceived relationships between them. The arcs thus represent the probabilistic relationship between the nodes and demonstrate the conditional dependence present in the network (Fenton and Neil, 2012). The Bayes' theorem is then applied to obtain probabilities of observing the class given a set of observed covariates. An example of the proposed model is given in Fig. 2 and illustrates how mode is affected by past rainfall in the last 48 h, river flow.

2.2. Model evaluation

All the fitted models were checked for compliance with the their corresponding assumptions. To assess predictive power, a leave-one-out and k-fold cross validation were used. The leave-one-out cross validation uses all but one observation in the data set to fit the model. The fitted model is then used to estimate the prediction error for the left out observation, and the step is then repeated for all observations (Hastie et al., 2009; Efron, 1983). The k-fold cross validation technique works similarly but removes k observations at a time (Hastie et al., 2009). In our case, the observations of each bathing season was removed to validate model performance with $k = 9$. The leave-one-out method gives an idea of how the model will perform in the long run while the k-fold cross validation tell us how different the bathing seasons are from one another.

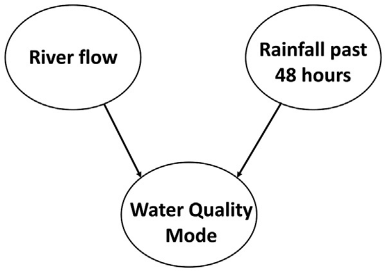

Fig. 2. Graph of the Bayesian Network used in modelling the modes observed in the Oreti River Wallacetown. The nodes are the variables and the edges show the conditional dependence between them. For example, mode is conditionally dependent on river flow.

If the continuous *E. coli* counts were being modelled, an observation was deemed predicted correctly if the estimate was within the mode boundaries of the observed mode. Model performance was evaluated via their respective cross validation error rates (CVER), i.e. the estimated proportion of misclassifications and the proportion of correct modes predicted given by the diagonal entries of the confusion matrices. The results for both leave-one-out and k-fold cross validation are reported.

Past values of weekly *E. coli* and weekly river flow were used in the dynamic regression model, with the previous two instances of each included in the model. The continuous counts of *E. coli* and river flow were log-transformed to improve compliance with various assumptions such as, for example, normality and homoscedasticity in the regression. A total of 3000 trees were constructed for the random forest (RF) for both the classification and regression. The Bayesian network model required the covariates to be discretized or split into groups. To aid in the decision of where to split the inputs, histograms of the past 48 h rainfall and river flow created to include the proportion of the observed modes at the corresponding bin. Scatter plots of river flow and past 48 h rainfall were also created with the points coloured to the corresponding mode. This visualisation allowed us to determine at which levels differentiated between modes. In addition it was found that splitting into two groups, i.e., dichotomisation, was found to be sufficient in obtaining high prediction rates. The variables were split by the following; 15.80 m^3/s and 2.00 mm for river flow and past rainfall 48 h respectively which corresponds to 60th percentile of the empirical data of these variables.

3. Results

The data contained observations of the summer months between 2005 and 2014 with an average of 17 weeks observed per year. The summary sample statistics of the variables used in the modelling process are reported in Table 3 and the predictors used for each model are given in Table 4.

The reported *E. coli* concentrations and their corresponding modes are shown in the top half of Fig. 3. The modes observed at Wallacetown were generally acceptable for recreational activities i.e. mode Amber and Green, and the poor water quality (Red) occurred only rarely. The observed modes were distributed as follows: Green 75.5%, Amber 13.9%, and Red 10.6%. Moreover, 11% of mode Green cases transitioned into Red the following week, while 80% of mode Red weeks became mode Green in the week that followed.

The 24 h accumulated rainfall is given in the bottom half of Fig. 3. It is evident that the 2005–2006 experienced more rain than other years which corresponded with a higher proportion of Amber and Red modes. The annual average rainfall for the surrounding Invercargill area is 1149 mm, with 33% falling between December to

Table 3
Table of sample statistics of variables used in modelling process, with $n = 151$.

Variable	Median	Mean	Std. dev	(2.5%, 97.5%)
E. coli MPN/100 mL	110.00	368.01	862.56	(10.00, 2832.50)
River Flow m^3/s	13.23	21.41	26.20	(4.79, 91.07)
Rainfall in mm	13.00	25.87	58.95	(0.00, 93.00)
Rainfall past 24 h in mm	0.00	3.08	6.03	(0.00, 23.12)
Rainfall past 48 h in mm	1.00	5.77	9.55	(0.00, 31.37)

Table 4
Predictors used for each model.

	Model	Predictors
Continuous Response	Dynamic Regression	River flow, rainfall past 48 Hours, *E. coli* (day previous), *E. coli* (two day previous) and river flow (day previous).
	Naive	Water quality mode (day previous).
	Regression	River flow, rainfall and rainfall past 48 h.
	Regression Tree	River flow and rainfall past 48 h.
	RF Regression	River flow, rainfall and rainfall past 48 h.
Categorical Response	Bayesian Network	River flow and rainfall past 48 h.
	Classification Tree	River flow, rainfall past 24 h and river flow (day previous).
	Linear Discriminant Analysis	River flow and rainfall past 48 h.
	Markov Chain	–
	Multinomial Logistic Regression	River flow and rainfall past 48 h.
	Quadratic Discriminant Analysis	River flow and rainfall past 48 h.
	RF Classification	River flow, rainfall and rainfall past 48 h.

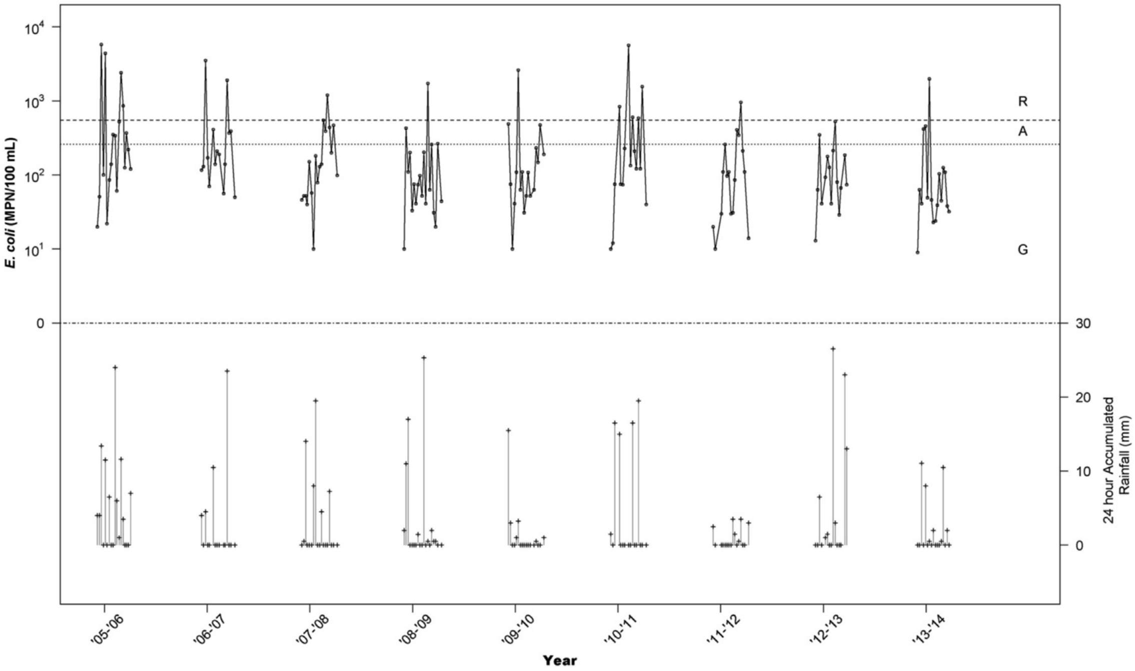

Fig. 3. Observed summer modes in the Oreti River at Wallacetown (top) and the corresponding 24 h accumulated rainfall at the Five Rivers (bottom). The boundaries of the modes are given by the horizontal lines and marked with their respective modes.

April. Rainfall is evenly distributed across the year in this area (Marcara, 2013). It can be noted that the 2005–2006 bathing season had above average rainfall was measured in the area (NIWA, 2005).

The leave-one-out cross validation results for the models are given in Table 5, with the proportion of correctly identified modes Green, Amber, Red and the cross validation error (CVER) reported. As mentioned in the methods, the naive model provides a benchmark for model performance. Overall, the naive model had the highest CVER as expected. The categorical-response models were generally better than then the continuous response models. All the models correctly predicted mode Green, with varying performance when predicting mode Red. This is with the exception of the Markov chain, as it could only correctly predict mode Green. However the prediction accuracy for the intermediate mode Amber was very poor for all models. In this study, much of the focus was to explore which model could best predict mode Red days. With this in mind the results show that the Bayesian network appeared to outperform all other models. The results for leave one out and the *k*-fold cross-validation are similar. The annual *k*-fold cross validation error rates are given in Table 6, and with the exception of the Markov chain are shown in Fig. 4. The high error rates observed for 2005–2006 summer can be explained by the fact that the proportion of Amber and Red modes were higher than other years with 44%, see Table 1. The above average rainfall that occurred at the time may account for the greater proportion of Amber and Red modes during those years (NIWA, 2005).

4. Discussion

In this study, various statistical models are used to predict water quality on a weekly basis, and their predictive accuracy is compared as well as assessing their suitability for prediction in real time. It was of particular importance to correctly predict mode Red, since it is associated with high risk of disease compared to the other modes. It was found that all models were able to accurately predict mode Green, but performed very poorly for mode Amber. For mode Red, the Bayesian network outperformed the other models, with 87%

Table 5
Model performance using leave-one-out cross validation. For proportion of correct mode predicted the closer the value is to 1 the better the performance and the value closer to zero for the cross validation error rate (CVER) indicates superior performance.

	Model	Proportion of Correct Mode Predicted for Succeeding Week			CVER
		Green	Amber	Red	
Continuous Response	Dynamic Regression	0.96	0.10	0.62	0.20
	Naive	0.77	0.24	0.06	0.38
	Regression	0.96	0.00	0.50	0.22
	Regression Tree	0.95	0.15	0.62	0.20
	RF Regression	0.95	0.14	0.62	0.20
Categorical Response	Bayesian Network	0.92	0.00	0.87	0.21
	Classification Tree	0.97	0.00	0.56	0.20
	Linear Discriminant Analysis	0.98	0.00	0.69	0.18
	Markov Chain	1.00	0.00	0.00	0.24
	Multinomial Logistic Regression	0.97	0.00	0.69	0.19
	Quadratic Discriminant Analysis	0.87	0.14	0.75	0.24
	RF Classification	0.93	0.00	0.62	0.23

Table 6
Average model performance for *k*-fold cross validation across the years. For proportion of correct mode predicted the closer the value is to 1 the better the performance and the value closer to zero for the cross validation error rate (CVER) indicates superior performance.

	Model	Proportion of Correct Mode Predicted for Succeeding Week			CVER
		Green	Amber	Red	
Continuous Response	Dynamic Regression	1.00	0.00	0.62	0.18
	Naive	0.75	0.23	0.03	0.38
	Regression	0.99	0.00	0.625	0.21
	Regression Tree	0.95	0.00	0.68	0.22
	RF Regression	1.00	0.00	0.80	0.17
Categorical Response	Bayesian Network	0.92	0.00	0.95	0.21
	Classification Tree	0.95	0.00	0.68	0.22
	Linear Discriminant Analysis	0.98	0.00	0.74	0.19
	Markov Chain	1.00	0.00	0.00	0.24
	Multinomial Logistic Regression	0.97	0.00	0.74	0.19
	Quadratic Discriminant Analysis	0.85	0.13	0.86	0.26
	RF Classification	0.94	0.00	0.71	0.22

mode Red observations correctly assigned for the leave-one-out and 95% for the *k*-fold cross validation. Therefore, we conclude that Bayesian network is the most suitable model for water quality prediction.

The water quality mode is assigned based on the reported *E. coli* MPN counts from the water body. Although other procedures to quantify *E. coli* concentration exist the Microbiological Water Quality Guidelines 2003 state that either the Membrane Filter Method, the MPN method or another accepted method must be used for *E. coli* in determining water quality (Ministry for the Environment, 2002). In our study only the MPN result was available and therefore used for analysis. However, it is important to note that the MPN is a positively-biased estimate of fecal coliform concentration (Garthright, 1997). It is also known to have wider variability in its estimates than the colony-forming-unit (CFU), another common measure of water quality, due to probabilistic basis of calculation of the MPN (Gronewold and Wolpert, 2008). Therefore, it should be noted that the phenomena modelled here may not be entirely reflective of the true and underlying process.

The ability of Bayesian networks to easily and flexibly handle missing data adds to their desirability as a modelling tool. In parametric models, missing observations are either omitted or imputed. The former may not be cost-effective: when observations are few, each one is valuable. The latter is often cumbersome, especially if the need for imputation is frequent. For example, a useful predictor for water quality is water temperature (Pratt and Chang, 2012; Carrillo et al., 1985; Faust et al., 1975). In our data set, however, it was only recorded occasionally. We have therefore excluded it from the analysis. However, in the future work, concentrating on Bayesian networks, we intend to add this variable (amongst others) to our model and investigate its effects on prediction accuracy noting its particular effect on Amber modes.

The estimation and fit of parametric models can also be affected by the presence of extreme values or outliers (Hastie et al., 2009). Bayesian networks circumvent this as the variables are discretized, thus ignoring the magnitude and influence of individual unusual observations. Furthermore Bayesian networks are also suitable for prediction in real time, as they are easily updated and there are no assumptions to check for.

However, it also known that despite its high predictive performance and aforementioned advantages, the Bayesian networks performance can be severely altered by the choice of discretization as well as the number of intervals used (Nojavan et al., 2017). In this study exploratory data analysis was used to aid where the variables should be discretized. By doing so, this allowed a better understanding in the underlying process which drives the transitional changes between modes and provides a justification of the choice of discretization. For the Bayesian network model, splitting the variables into two groups, i.e., dichotomizing, was sufficient to obtain high prediction rates. However, for other study areas, dichotomizing may not produce good enough results, and the discretization may need to be reconsidered on a site-by-site basis.

Furthermore the poor results for mode Amber demonstrate that further modelling work is required. All models performed especially poorly for the Amber mode. This may be due to the fact that it is rare (13.9% of all occurrences in our data) and transient

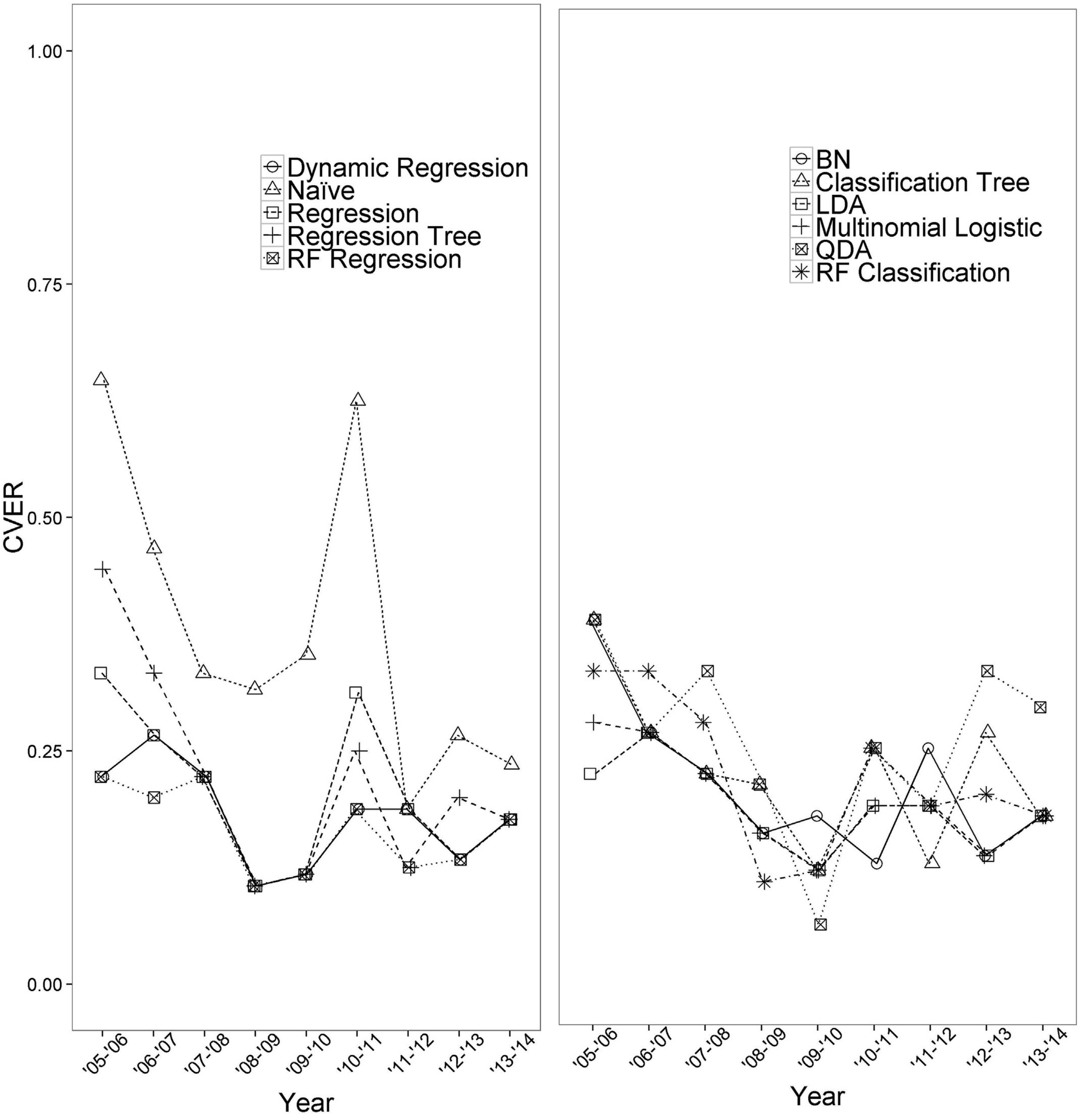

Fig. 4. Model performance by year using *k*-fold cross validation. The closer the cross validation error rate (CVER) is to zero the better the performance. Here the results for the modelling approaches are split up, the continuous response is on the left and the categorical response is on the right.

(probability of switching to another state 0.76 compared with 0.23 for Green and 0.94 for Red). Although, the transient probability is higher for Red modes, it is well established that increased rainfall corresponds to higher *E. coli* levels. However the rainfall threshold is not clearly observed or established for Amber modes. In general, the model accuracy specifically with respect to the Amber model, may be improved either by obtaining more data or by better understanding the processes behind changes in water quality and incorporation of spatial information such as land use and its proximity to the FMU as well as adding other covariates such as water temperature, electrical conductivity and other water quality indicators such as Halides.

Previous studies have explored prediction of water quality at recreational level, and evaluated changes in rivers and lakes through space and time. Deciding on an optimum model depends on the objectives of a study, with each model having its own advantages and limitations. Thoe et al., 2014 explored five statistical models; multiple linear regression, logistic regression, partial least squares regression, artificial neural networks and classification tree to predict increased levels of fecal indicator bacteria (FIB) in Santa

Monica Beach. The aim was to propose a model that would better the naive model approach that was utilised at the time. Their results showed an improvement over the naive model, with the classification tree achieving the best performance, predicting 42% of unsafe FIB levels compared to 28% by the naive. These results are consistent with our findings, as the naive and Markov chain had poor model performance.

Another example of modelling unsafe levels of fecal indicator bacteria is by Stidson et al. (2012), using regression trees they predicted 81% of unsafe levels correctly, and is the current method used for real-time water quality prediction across bathing sites in Scotland. Moreover regression trees were also used in a study conducted by De'ath et al., 2010 to evaluate the health status of the Great Barrier Reef, concluding that decreased water clarity and increased chlorophyll degrades the reef's health. In addition to high predictive power, regression trees can aid in understanding the relationship between response and predictor as the model is given by a decision tree. When modelling water quality parameters such as discharge, water temperature, dissolved oxygen etc. of Slovenian rivers, Džeroski et al., 2000 preferred the regression tree over multiple regression and the nearest neighbour method as it illustrated how the predictors affected the response. The results from our study also show that regression trees are powerful for prediction with a leave-one-out CVER of 20% and *k*-fold CVER of 22%. It also achieved one of the higher performances for mode Red, with the leave-one-out and *k*-fold cross validation yielding 62% and 68% correct respectively.

Trees are known to suffer from instability where small changes in the data can result in different partitions thus making interpretation precarious (Hastie et al., 2009). This is particularly problematic when the number of predictors are high and to address this, random forest's can be used. The number of predictors in this case study was low and therefore the results of the random forest did not differ significantly from the single classification or regression tree. For a single classification and regression tree it is easy to get an insight into decision rules if the tree is small. However for RF's this is no longer the case, as the outcome is the average result of many trees. Therefore the loss of insight of the decision rules may not be desirable when working with small data sets.

In our analysis, discriminant analysis had one of the highest performance with LDA and QDA predicting mode Red correctly with 69% and 75% respectively for leave-one-out cross validation and 74% and 86% respectively for *k*-fold cross validation. Despite the high proportion of correct mode Red predictions, for both discriminant analysis and regression trees, this level may not be high enough for policy makers and users, as the cost of false negatives can have an adverse effect on human health. Previous studies have also demonstrated Discriminant analysis' high predictive power. For instance, Shrestha and Kazama 2007 used it to model seasonal variations of water quality parameters found in surface water in the Fuji River Basin, with discriminant analysis correctly identifying 85% of the parameters variability. Moreover, Wunderlin et al., 2001 also evaluated spatio-temporal changes of water quality parameters in the Suquia River Basin, Argentina, resulting in 87% correctly predicted for temporal analysis and 75% in spatial analysis. Similarly, discriminant analysis was also fitted by Singh et al., (2004) to model spatio-temporal variations of water quality parameters in the Gomti River, resulting in 88% correctly predicted for temporal analysis and 91% in spatial analysis.

Despite their drawbacks in the choice discretization method (Gronewold and Wolpert, 2008) results of this study suggest that Bayesian networks are an ideal tool for water quality prediction as they are capable of high predictive power, see Tables 4 and 5 Like the regression tree, Bayesian networks are graphically given in a DAG, resulting in a better understanding of the relationship between the response and its predictors see Fig. 2. Other applications of Bayesian networks are by Ha and Stenstrom (2003), with their aim to differentiate between storm water origins based on land use in the Santa Monica Bay. The results of their Bayesian network correctly identified 92.3% of storm water origins. Furthermore Bayesian network's can help identify high risk groups to disease in relation to polluted water. For example Donald et al., 2009 modelled the risk of gastroenteritis associated with recycled water, with the model results indicating that the young and elderly were most susceptible to gastroenteritis. Although this is common knowledge, for other applications it shows it is capable in identifying previously unknown high risk groups. Therefore, the results from this study and those previous, suggest that a Bayesian network model should be preferred for water quality prediction as it is capable of high predictive power.

FMUs are also assigned a long term water quality grade, which is based on long term *E. coli* data trends (Ministry for the Environment, 2014). It would be of interest to investigate how BNs can be used for lakes and river grading to compare if the set limits would be similar. In addition it may be of use to evaluate how risk to GI illness differs between FMUs based on the surrounding catchment area as well as other risk factors. This can help determine if different factors unique to a site should be considered when allocating a water quality grade or if the current method is sufficient. Future modelling work will also see the Bayesian network extended to help identify possible sources of pollution and its potential in river grading.

5. Conclusion

The results from our analysis indicate that the most suitable model for real time water quality is the Bayesian network, as it could correctly predict the majority of mode red days and had a low CVER. Furthermore its ability to handle missing values, outliers and its updatability capability make it ideal for real time prediction. Future modelling work is to fit the Bayesian network model to other areas and assess its overall performance. In addition a spatial component will be included, allowing connected upstream sites and surrounding land to have an influence on the FMU, with the aim of increased accuracy of mode Amber and Red predictions. Finally, we hope that the conditional dependencies displayed in the network will aid in policy decisions regarding water quality at the recreational level.

Acknowledgements

Rodelyn Avila was jointly supported by the ESR Postgraduate scholarship and UC Connect Doctoral Scholarship (The University of Canterbury, New Zealand). Gratitude must also be extended to Environment Southland for use of their data and willingness to share their expertise and local knowledge.

References

Agresti, Alan, 1996. An Introduction to Categorical Data Analysis, vol. 312. Wiley, ISBN 0471113387.

Bridle (Heriot Watt University), Helen, 2014. Waterborne Pathogens; Detection Methods and Applications, vol. 401. Elsevier B.V, 9780444595430 (hbk.).

Carrillo, M., Estrada, E., Hazen, T.C., 1985. Survival and enumeration of the fecal indicators bifidobacterium-adolescentis and Escherichia-Coli in a tropical rain forest watershed. Appl. Environ. Microbiol. ISSN: 0099-2240 50 (2), 468–476.

Cheng, Tseng Tung, 1949. The normal approximation to the Poisson distribution and a proof of a conjecture of Ramanujan. Am. Math. Soc. 55, 396–401. https://doi.org/10.1090/S0002-9904-1949-09223-6.

De'ath, Glenn, et al., 2010. Water quality as a regional driver of coral biodiversity and macroalgae on the Great Barrier Reef Water as a driver of coral quality biodiversity regional on the Great Barrier Reef and. Ecol. Soc. Am. ISSN: 10510761 20 (3), 840–850. https://doi.org/10.1890/08-2023.1.

Donald, Margaret, Cook, Angus, Mengersen, Kerrie, 2009. Bayesian network for risk of diarrhea associated with the use of recycled water. Risk Anal. ISSN: 02724332 29 (12), 1672–1685. https://doi.org/10.1111/j.1539-6924.2009.01301.x.

Draper, N.R., Smith, H., 1998. Applied regression analysis. Technometrics. ISSN: 00359254 47 (3), 706. https://doi.org/10.1198/tech.2005.s303.

Džeroski, Sašo, Demšar, Damjan, Grbović, Jasna, 2000. Predicting chemical parameters of river water quality from bioindicator data. Appl. Intell. ISSN: 0924669X 13 (1), 7–17. https://doi.org/10.1023/A:1008323212047.

Edberg, S.C., et al., 2000. *Escherichia coli*: the best biological drinking water indicator for public health protection. In: Symposium Series (Society for Applied Microbiology). ISSN: 1467-4734, vol. 88(29), pp. 106S–116S. https://doi.org/10.1111/j.1365-2672.2000.tb05338.x.

Efron, Bradley, 1983. Estimating the error rate of a prediction rule : improvement on cross-validation. Am. Stat. Assoc. 78 (382), 316–331.

Environment Southland, 2010. Regional water plan for Southland. Visited on 09/05/2016. http://www.es.govt.nz/DocumentLibrary/Plans,policiesandstrategies/Regionalplans/RegionalWaterPlan/regional/_water/_plan.pdf.

Environment Southland, Te Ao Marama Inc, 2010. Our Health: Is Our Water Safe to Play in, Drink and Gather Kai from? Part 1 of Southland Water 2010: Report on the State of Southland's Freshwater Environment. Visited on 09/05/2016. http://www.es.govt.nz/DocumentLibrary/Researchandreports/SOEreports/water-2010-our-health.pdf.

European Parliament, 2006. Directive 2006/7/EC of the european parliament and of the council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC. Off. J. Eur. Commun. 64, 37–51.

Fabaozzi, Frank J., Focardi, Sergio M., Kolm, Petter N., 2006. Financial Modeling of the Equity Market from CAPM to Cointegrations, vol. 648. Wiley, New Jersey. http://vk.com/doc215711421/_317805266?hash=35172a618cc7347479/&dl=35d467177215b1d495.

Faust, Maria a., Aotaky, a. E., Hargadon, M.T., 1975. Effect of physical parameters on the in situ survival of *Escherichia coli* MC-6 in an estuarine environment. Appl. Microbiol. ISSN: 0003-6919 30 (5), 800–806. https://doi.org/10.1007/BF02090102.

Fenton, Norman E., Martin (Martin D.) Neil, 2012. Risk Assessment and Decision Analysis with Bayesian Networks, vol. 503. Taylor & Francis isbn: 9781439809105.

Fewtrell, Lorna, Kay, David, 2015. Recreational water and infection: a review of recent findings. Curr. Environ. Health Rep. ISSN: 2196-5412 2 (1), 85–94. https://doi.org/10.1007/s40572-014-0036-6.

Freedman, David, 1971. Markov Chains, vol. 382. Holden-Day, ISBN 0816230048.

Garthright, W.E., 1997. A Bayesian analysis of serial dilutions offers a worse positive bias than the MPN and proposes an inappropriate interval estimate. Food Microbiol. 14, 515–517.

Given, Suzan, Pendleton, Linwood H., Boehm, Alexandria B., 2006. Regional public health cost estimates of contaminated coastal waters: a case study of gastroenteritis at southern California beaches. Environ. Sci. Technol. ISSN: 0013936X 40 (16), 4851–4858. https://doi.org/10.1021/es060679s.

Gleick, P.H., 2002. Dirty Water: Estimated Deaths from Water-related Diseases 2000-2020. Tech. rep. Pacific Institute Research Report.

Gronewold, Andrew D., Wolpert, Robert L., 2008. Modeling the relationship between most probable number (MPN) and colony-forming unit (CFU) estimates of fecal coliform concentration. Water Res. 42, 3327–3334. https://doi.org/10.1016/j.watres.2008.04.011.

Ha, Haejin, Stenstrom, Michael K., 2003. Identification of land use with water quality data in stormwater using a neural network. Water Res. ISSN: 00431354 37 (17), 4222–4230. https://doi.org/10.1016/S0043-1354(03)00344-0.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, 2009. Second. The Elements of Statistical Learning, vol. 1. Springer, ISBN 9780387848570, pp. 337–387. https://doi.org/10.1007/b94608.

Hunter, Paul R., Zmirou-Navier, Denis, Hartemann, Philippe, 2009. Estimating the impact on health of poor reliability of drinking water interventions in developing countries. Sci. Total Environ. ISSN: 00489697 407 (8), 2621–2624. https://doi.org/10.1016/j.scitotenv.2009.01.018.

Kang, Hyon, Joo, et al., 2010. Linking land-use type and stream water quality using spatial data of fecal indicator bacteria and heavy metals in the Yeongsan river basin. Water Res. ISSN: 00431354 44 (14), 4143–4157. https://doi.org/10.1016/j.watres.2010.05.009.

Liaw, Andy, Wiener, Matthew, 2002. Classification and regression by randomForest. R. News 2 (3), 18–22. http://CRAN.R-project.org/doc/Rnews/.

Mallin, Michael A., et al., 2016. Effect of human development on bacteriological water quality in coastal watersheds. Ecol. Appl. 10 (4), 1047–1056.

Maniquiz, Marla C., Lee, Soyoung, Kim, Lee-hyung, 2010. Multiple linear regression models of urban runoff pollutant load and event mean concentration considering rainfall variables. J. Environ. Sci. ISSN: 1001-0742 22 (6), 946–952. https://doi.org/10.1016/S1001-0742(09)60203-5.

Marcara, G.R., 2013. The Climate and Weather of Southland. Visited on 09/26/2017. NIWA Science and Technology Series. https://www.niwa.co.nz/sites/niwa.co.nz/files/Southland_Climate_WEB.pdf.

McDowell, Rw, Wilcock, Rj, 2008. Water quality and the effects of different pastoral animals. N. Z. Vet. J. ISSN: 0048-0169 56 (6), 289–296. https://doi.org/10.1080/00480169.2008.36849.

Ministry for the Environment, 2015. A Guide to the Ment for Freshwater Management 2014. Tech. rep. New Zealand government, Wellington.

Ministry for the Environment, 2002. Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas. Tech. rep. New Zealand government.

Ministry for the Environment, 2014. National Policy Statement for Freshwater Management 2014. Tech. rep. New Zealand government.

Muirhead, R.W., et al., 2004. Faecal bacteria yields in artificial flood events: quantifying in-stream stores. Water Res. ISSN: 00431354 38 (5), 1215–1224. https://doi.org/10.1016/j.watres. 2003.12.010.

Muirhead, Richard William, Collins, Robert Peter, Bremer, Philip James, 2006. Interaction of *Escherichia coli* and soil particles in runoff interaction of *Escherichia coli* and soil particles in runoff. Appl. Environ. Microbiol. ISSN: 0099-2240 72 (5), 3406–3411. https://doi.org/10.1128/AEM.72.5.3406.

NIWA, 2005. National climate summary - summer 2004/05. Visited on 08/24/2016 https:/www.niwa.co.nz/sites/niwa.co.nz/files/import/attachments/sclimsum/_05/_1/_summer.pdf.

Nojavan, A., Farnaz, Qian, Song S., Stow, Craig A., 2017. Comparative analysis of discretization methods in Bayesian networks. Environ. Model. Softw. ISSN: 13648152 87, 64–71. https://doi.org/10.1016/j.envsoft.2016.10.007.

Odonkor, Stephen T., Ampofo, Joseph K., 2013. *Escherichia coli* as an indicator of bacteriological quality of water: an overview. Microbiol. Res. ISSN: 2036-7481 4 (1), 5–11. https://doi.org/10.4081/mr.2013.e2.

Pearson, L., Couldrey, M., 2016. Methodology for GIS-based land use maps for Southland. Environment Southland Publication No 2016-10. http://www.es.govt.nz/Document%20Library/Research%20and%20reports/Various%20reports/Science%20reports/Land%20use%20inputs/Report%20-%20Methodology%20for%20GIS-based%20Land%20Use%20Maps%20for%20Southland.pdf.

Peizer, David B., Pratt, John W., 1968. A normal approximation for binomial, F, beta, and other common, related tail probabilities, I. J. Am. Stat. Assoc. ISSN: 01621459 63 (324), 1416. https://doi.org/10.2307/2285895.

Pratt, Bethany, Chang, Heejun, 2012. Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. J. Hazard Mater. ISSN: 1873-3336 209210, 48–58. https://doi.org/10.1016/j.jhazmat.2011.12.068.

Prüss, Annette, 1998. Review of epidemiological studies on health effects from exposure to recreational water. Int. J. Epidemiol. ISSN: 0300-5771 27 (1), 1–9 doi:10. 1093/ije/27.1.1.

R Core Team, 2015. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Sampson, Reyneé W., et al., 2006. Effects of temperature and sand on *E. coli* survival in a northern lake water microcosm. J. Water Health. ISSN: 14778920 4 (3), 389–393. https://doi.org/10.2166/wh.2006.024.

Shrestha, S., Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ. Model. Softw. ISSN: 13648152 22 (4), 464–475. https://doi.org/10.1016/j.envsoft.2006.02.001.

Singh, Kunwar P., et al., 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) - a case study. Water Res. ISSN: 00431354 38 (18), 3980–3992. https://doi.org/10.1016/j.watres.2004.06.011.

Soller, Jeffrey a., et al., 2010. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. Water Res. ISSN: 00431354 44 (16), 4674–4691. https://doi.org/10.1016/j.watres.2010.06.049.

Spedicato, Giorgio Alfredo, 2015. Markovchain: Discrete Time Markov Chains Made Easy. R Package Version 0.3.1.

Stidson, R.T., Gray, C.A., McPhail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water Environ. J. ISSN: 17476585 26 (1), 7–18. https://doi.org/10.1111/j.1747-6593.2011.00258.x.

Therneau, Terry, Atkinson, Beth, Ripley, Brian, 2015. Rpart: Recursive Partitioning and Regression Trees. R Package Version 4.1-10. http://CRAN.R-project.org/package=rpart.

Thoe, W., et al., 2014. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. ISSN: 1879-2448 67C, 105–117. https://doi.org/10.1016/j.watres.2014.09.001.

Venables, W.N., Ripley, B.D., 2002a. Modern Applied Statistics with S. Fourth. Springer, New York, ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.

Venables, W.N., Ripley, B.D., 2002b. Modern Applied Statistics with S. Fourth. Springer, New York, ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.

Wade, Timothy J., et al., 2003. Do U.S. Environmental protection agency water quality guidelines for recreational waters prevent gastrointestinal Illness? A systematic review and meta-analysis. Environ. Health Perspect. 111 (8), 1102–1109 doi:10. 1289/ehp.6241.

Winfield, M.D., Groisman, E.A., 2003. Role of nonhost Environments in the lifestyles of Salmonella and *Escherichia coli*. Appl. Environ. Microbiol. ISSN: 0099-2240 69 (7), 3687–3694. https://doi.org/10.1128/AEM.69.7.3687-3694.2003.

World Health Organization, 2003. Coastal and Fresh Waters". Geneva 1:219. Guidelines for Safe Recreational Water, vol. 1. http://www.who.int/water/_sanitation/_health/bathing/srwe2full.pdf.

Wunderlin, Daniel Alberto, et al., 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River basin (Cordoba-Argentina). Water Res. 35 (12), 2881–2894.

Yoder, Jonathan S., et al., 2008. Surveillance for waterborne disease and outbreaks associated with recreational water use and other aquatic facility-associated health events United States, 2005-2006. Visited on 09/06/2016. http://www.cdc.gov/mmWR/preview/mmwrhtml/ss5709a1.htm.