8-1-2014

Development And Testing Of Data Driven Nowcasting Models Of Beach Water Quality

Jainy Mavani

Lianghao Chen

Darko Joksimovic

Songnian Li

DEVELOPMENT AND TESTING OF DATA DRIVEN NOWCASTING MODELS OF BEACH WATER QUALITY

JAINY MAVANI (1), LIANGHAO CHEN (1), DARKO JOKSIMOVIC (1), SONGNIAN LI (1)
(1): Department of Civil Engineering, Ryerson University, 350 Victoria Street, Toronto, Ontario, M5B 2K3, Canada

Recreational water users may be exposed to elevated pathogen levels that originate from various point and non-point sources. Current daily notifications practice depends on microbial analysis of indicator organisms (persistence model) such as Escherichia coli (E. coli) that require 18-24 hours to provide sufficient response. This research evaluated the use of Artificial Neural Networks (ANNs) and Evolutionary Polynomial Regression (EPR) for real time prediction of E. coli concentration in water at beaches in Toronto, Ontario, Canada. The nowcasting models were developed using readily available real-time environmental and hydro-meteorological data available for four bathing seasons (June-August). The results of the developed models were compared with historic data and found that the predictions of E. coli levels generated by ANN models slightly outperformed those generated using EPR. The best performing ANN models are able to predict up to 74% of the E. coli concentrations, offering an improvement over the currently employed persistence approach.

INTRODUCTION

Beaches are treasured natural resources that provide significant value, including recreational benefits in summer time. However, beach waters can contain various pathogenic micro-organisms, which are a potential threat to human health and can cause beach closures for periods of time. Due to the same reason, there has been an increasing interest for the last couple of decades to develop the models for fast assessment of beach water quality.

Out of these micro-organisms, elevated levels of Escherichia coli (E. coli) is used as an indicator of contamination and to indicate the presence of human or animal fecal wastes and other harmful bacteria in lakes and streams [1]. Numerous factors may explain fluctuations in E. coli level, including rainfall, wind speed and direction, wave height, turbidity, direction of flow and biological factors [2]. As a standard practice, measuring the geometric mean concentrations of the E. coli during the swimming season of each year has been the basis that establishes whether water quality meets/fails the safety levels for recreational purposes. However, due to the time required to obtain culture-based results (18-24 hours), E. coli concentrations are typically not available until the following day of the actual sample collection. The delay, coupled with the temporal and spatial variability associated with E. coli, sometimes results in unwarranted beach closures or the lack of a advisory when E. coli concentrations are, in fact, elevated and a public health risk exists [3].

Various approaches have been developed to address this time lag problem, including attempts to shorten analysis time for water quality monitoring, use of quicker predictive methods and communicating beach water quality information to the public on a timely (e.g., near-daily) basis so more informed decisions can be made by the public regarding recreational water use. [4]. Efforts have been made to develop predictive models for nowcasting and forecasting the level of E. coli around the world for beach condition. Nowcasting refers the current situation and what changes to expect over the next 2-6 hours. Nowcast systems operate continuously with little user intervention, which is appropriate for the time scales of the phenomena of interest [5].

The objective of this paper is to develop predictive models to nowcast beach status using the available readily available data measured by some sort of sensors or automated systems. In order to accomplish this objective, the following issues have been addressed:

- Exploring the correlations between indicator organism concentrations and other water quality and meteorological variables,
- Developing ANN and EPR models to forecast E. coli concentration for beach waters,
- Investigating the influence of different input parameter selection methods on models' development and performance,
- Investigating the influence of variable transformation on model performance,
- Comparing model performance to that achieved by the current practice.

STUDY AREA

Beaches are a key feature of Toronto's waterfront parks which contribute significantly to the quality of life in the city. Toronto's lakefront spans 157 kilometers of shoreline, 5.5 kilometers of which are supervised beaches (11 locations) designated for swimming during summer time. Eight beaches fly the Blue flag, which requires that individual beaches have water quality which enables them to be open for at least 80% of the swimming season and monitored by Environmental defense (www.blueflag.ca) in Canada. The remaining three beaches (Sunnyside, Rouge and Marie Curtis East Park), shown in Figure 1, are located near the mouth of major river systems and are with the poorest beach water quality and are regularly posted against swimming. As per Toronto City Council action plan, aimed at improving and enhancing the swimming beaches, the assessment of water quality at these three city beaches is of concern [6].

When E. coli is found in water samples collected at concentrations greater than 100 E. coli per 100 milliliters of water, the beaches are posted with advisory signs because swimming could lead to health effects such as skin rashes or gastro-intestinal illnesses [3]. Toronto Public Health (TPH) department determines the public health implications of the bacteria data, posts the result on their website (http://app.toronto.ca/tpha/beaches.html) and conveys this information to the jurisdiction that manages a particular beach; in most cases city's parks department.

The persistence model, based on traditional analysis method, that is currently employed uses last available value to manage beaches and has a significant lag period. As a result of limitations associated with laboratory quantification of microbial water quality and the need for beach managers to balance access to water recreation with protection of public health, real-time or near real-time predictive tools to aid in beach management decisions have been used, including rapid analytical techniques and deterministic models, regression models and artificial neural network based models.

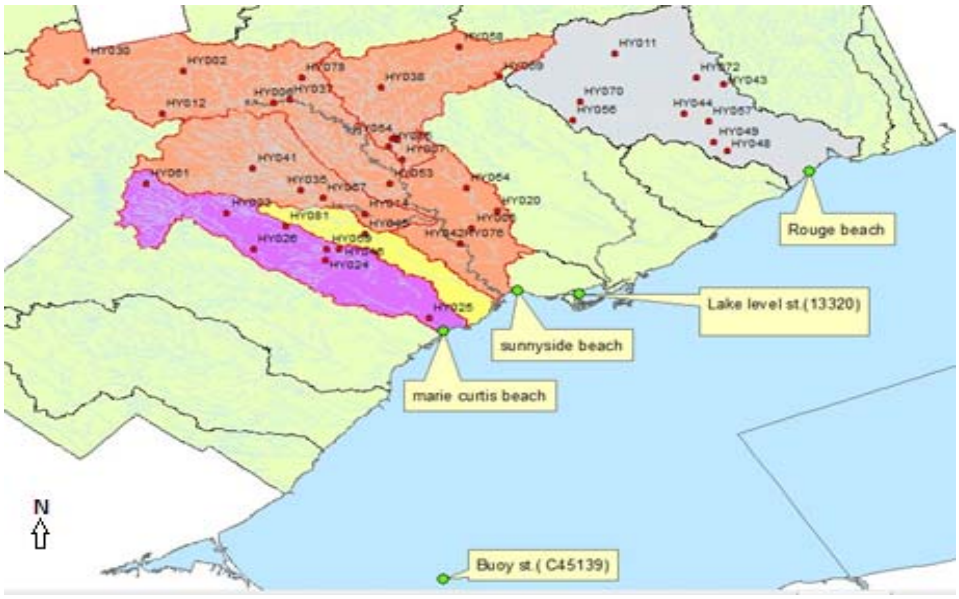


Figure 1 Locations of the beaches, buoy station and lake level station

NOWCASTING MODELS

An Artificial Neural Network (ANN) is a construct of software that partially mimics the workings of a biological neural network. ANNs are often used to model relationships between inputs and outputs or to find patterns and applied as nonlinear statistical data modelling tools. They can be used to model relationships between inputs and outputs or to find patterns. The technique is often useful when relationships between inputs and outputs are complex and not clearly understood. An ANN learns relationships between inputs and outputs using a learning algorithm [7]. ANNs exhibit many characteristics which make them attractive and appropriate for nowcasting/forecasting [8], including the ability to correctly generalize the unseen data even if the training data contained noise and the ability to learn (i.e. through examples), while refining their structure without the need of any predefined rules.

He and He [9] successfully used ANNs to predict indicator organism at marine recreational beaches receiving watershed base-flow and stormwater runoff in Southern California. Varma and Vijayan [10] carried out the research work to predict fecal coliform concentration in surface water of the Achancovil River in Kerala, India. Different inductive models were developed using ANN and found to perform better compared with statistical model that used the same parameters. Zhang et al. [11] compared ANN model for nowcasting and forecasting E. coli levels with other two models developed using US EPA Virtual Beach (VB) Program at Gulf Coast beaches in Louisiana, USA. The results indicated that the ANN model with 15 parameters performs better than the VB models with 6 or 5 parameters.

Evolutionary Polynomial Regression (EPR) is a novel data-driven technique, developed by Giustolisi and Savic in 2006 [17]. It has seen success in various fields since its introduction including geotechnical engineering [18, 19] and municipal engineering [20], though it has not been applied to environmental science as the problem presented in this paper. EPR first generates the model structure using genetic algorithm then selects model parameters using Least Square regression. The result obtained is polynomials in the form of:

$$Y = a_0 + \sum_{i=1}^{m} a_i \cdot \left(X_1\right)^{ES(j,1)} \cdot \ldots \cdot \left(X_k\right)^{ES(j,k)} \quad (1)$$

Where Y is the target objective, m is the maximum number of terms, X_i is the value for candidate variables, ES is the power of variables, a_j are estimated constants and a_o is an optional model bias. In this model, m, the range of ES and the presence of a_o can be defined by user [18]. Selection for best model is based on three criteria: 1. Reduce the sum of squared errors (SSE), 2. Reduce model complexity based on number of input combinations, 3. Reduce variance of estimated constant a_j [17]. An advantage EPR offers is that it is a highly automated process, very little input variable preparation is required, user may define the maximum number of terms in output expressions, EPR will select best variables from a list of candidate inputs. For the study, explanatory data set used is the same as ANN, models where constructed using both normalized data by taking natural logarithm of concentrations measured and non-normalized data.

DATA DESCRIPTION AND ANALYSIS

Data gathering and exploratory data analysis for the beach water quality of selected Rouge beach, Sunnyside beach and Marie Curtis Park East beach were performed. The most commonly used variables are those that both have some relationship to beach water quality and are typically readily available: rainfall, stream flow, solar radiation, lake level, wind speed and direction, turbidity, wave height and past E. coli levels. Data for hydro-meteorological parameters were considered only up to previous day's midnight taking into account maximum lag time out of all input parameters; this is to maintain the nowcasting ability if they are used as inputs to the water quality predictive models.

For this research work the July to August (2008 to 2012) precipitation data were collected for rain gauge stations located inside the relevant watershed areas. Stations that didn't have any major part of the data missing for bathing season and that had a higher value of correlation between precipitation and E. coli data were used. Real-time and historical solar radiation data were obtained from the Toronto and Region Conservation Authority's HY039 station located in the Humber river watershed, as shown in Figure 1. For all beaches, the previous day's solar radiation ending at midnight was considered which were available at 15 minutes interval. Historical and real-time stream flow data were downloaded through the Environment Canada website that includes river stage and a calculated river discharge. Hourly and historical wind and wave height data for all beaches were obtained from the Environment Canada Buoy station C45139 as shown in Figure 1, located the West of Lake Ontario. Real-time and historical lake levels (meters) were available for Lake Ontario from Toronto 13320 station from Fisheries and Oceans Canada, as shown in Figure 1. Beach water quality data were obtained from Toronto Public Health for 2008 to 2012 time periods. E. coli concentration data was transformed to natural logarithm (*lnEC*) before it was correlated with or predicted from different explanatory parameters. Several other parameters such as turbidity, wind direction and speed, waterfowl counts, wave height category (low, moderate, high) and water temperature were not considered mainly due to not being readily available on a daily basis or being difficult to be predicted or simply because of missing data for considerable period of time.

Scatter plots and correlation analysis were then used to detect potential relationships between variables. Scatter plots are obtained to visually investigate the relationship between *lnEC* and the environmental variables and to identify the critical factors that can affect beach water quality. Table 1 summarizes the parameters that showed statistically significant linear

correlations with E. coli at each of the beaches considered. These variables were used in the development of ANN models, while all potential variables were used in the EPR model development. Prior to inputting variables to model, preprocessing procedures were conducted to train the ANNs more efficiently. These procedures were: 1) solve the problem of missing data and 2) normalize data. Occasionally missing E.coli data were replaced by the average of neighboring values [12, 15]. Normalization of data into a particular range prior to applying transfer functions was performed [16].

Table 1 Explanatory variables at Toronto beaches considered for ANN model development

Explanatory Variables	Nomenclature	Sunnyside	Rouge	Marie Curtis
Previous day ln E. coli	*pr.lnEC*	☑	☑	☑
Flow of Humber River (m^3/s)	*st.fl.HR*	☑		
Flow of Mimico Creek (m^3/s)	*st.fl.MC*	☑		
Flow of Black Creek (m^3/s)	*st.fl.BC*	☑		
Flow of Rouge River (m^3/s)	*st.fl.RR*		☑	
Flow of Etobicoke Creek (m^3/s)	*st.fl.EC*			☑
Lake level(m)	*l.l*	☑		☑
Wave ht. (m)	*w.ht*		☑	☑
Wind Direction (deg)	*w.dir*		☑	
Wind speed (m/s)	*w.spd*	☑	☑	
solar radiation (MJ/m^2)	*slr*	☑	☑	☑
Cumulative 2 day rain (mm)	*r48*	☑	☑	☑

EVALUATION OF MODEL PERFORMANCE

Out of the numerous models generated, a summary of best performing ANN and EPR models for the three beaches, determined by the percent of instances of concentrations exceeding beach water quality standard, is summarized in Table 2. Also shown for comparison is the performance of the persistence model. All models use the transformed natural logarithm of past E. coli measurements as input, in addition to a number of other and differing parameters from the list of potential explanatory variables. Compared to the persistence model performance, it is evident that improvements in nowcasting ability for all three beaches can be made by employing either of the data driven models developed in this research. The two modelling approaches have similar predictive capability in terms of correct classification for the Sunnyside and Rouge beaches; however, the ANN model significantly outperforms both the persistence model and EPR approach for the Marie Curtis beach.

In addition to the quantitative assessment, the actual performance of the model to predict the exceedance of water quality threshold (e.g. 100 counts/100 ml) can also be visually assessed from scatter plots of the simulation data divided into quadrants differentiating true negatives, false positives, true positives, and false negatives. The scatter plot plots for the three beaches and models generated using the two modelling approaches (ANN on the left, EPR on the right) are shown alongside the persistence model results in Figure 2.

For all three beaches, overall correct classification percentage is slightly higher for the ANN models and they balance the rates of true positives and true negatives. However, in certain instances the ANN models leads to increase in false negatives. Although the overall predictive performance of the two approaches is similar, the transformed E. coli prediction produced by

the EPR models are generally closer to observed values, although their performance in terms of the number of false positives and false negatives is similar to the ANN models.

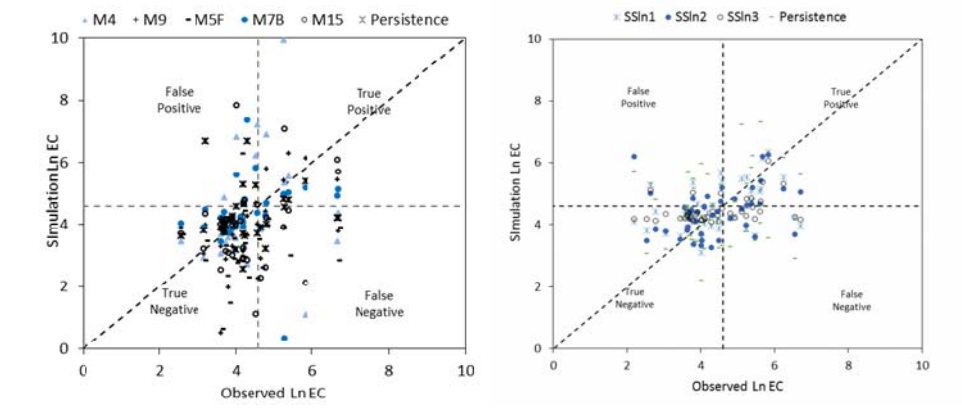
CONCLUSIONS

The nowcasting models, based on ANN and EPR methodologies, were developed using readily available real-time environmental and hydro-meteorological data available for four bathing seasons (June-August). Both models were found to perform better than the persistence model that bases the decision making on whether to post beaches on previous days measured E.Coli concentrations. The results of the developed models were compared with historic data and found that the predictions of E. coli levels generated by ANN models slightly outperformed those generated using EPR. The best performing ANN models are able to predict up to 74% of the E. coli concentrations, offering an improvement over the persistence model currently employed. Flows in nearest contributing creeks and rivers, solar radiation, previous day transformed E. coli concentration and 48-hour rainfall measured at respective watersheds were found to be important for all three beaches analyzed. In addition, wind direction and speed were found to be of importance in one of the three beaches.

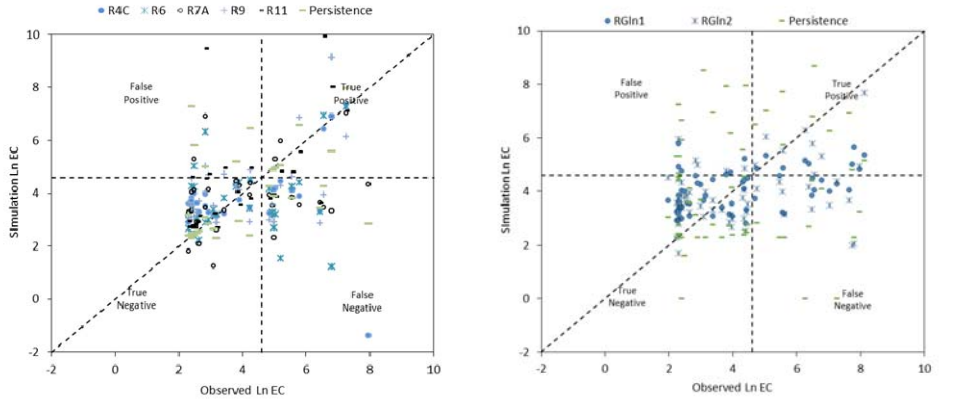
Table 2 A summary of the best performance achieved using ANN and EPR models

Beach	Model	Input combination (ANN) Model Structure (EPR)	Correct classification
Sunnyside	ANN	st.fl.HR, w.spd, slr, r48(HY041), r48(T.W.2), pr.lnEC, l.l	74%
	EPR	$0.22121\frac{w.t^{1.5}w.spd^{0.5}}{w.ht^{0.5}}+1.5876a.t^2l.l^2l.l^2pr.lnEC^{+}0.49931st.fl.BC^{0.5}+0.17436$	73%
	Persistence	-	66%
Rouge	ANN	st.fl, w.dir, w.ht **and** st.fl, r48(HY044), r48(HY070), pr.lnEC	71%
	EPR	$\frac{st.fl.RR^{w.ht0.5}w.spd^2pr.lnEC^{0.5}}{r48^{0.5}slr^{0.5}}+0.0040769st.fl.RR^{0.5}pr.lnEC^{0.5}+0.00012677st.fl.RR^{0.5}a.t^{w.dir0.5}+0.0012487st.fl.RR^{0.5}w.t^{0.5}w.ht^{0.5}+-5.7046e-008r48^{0.5}st.fl.RR^{0.5}w.dir^{0.5}pr.lnEC^{1.5}$	73%
	Persistence	-	67%
Marie Curtis	ANN	st.fl, w.ht, w.dir, slr, r48(HY025), r48(HY033), pr.lnEC, l.l	71%
	EPR	$0.53513\tanh(w.spd)$	59%
	Persistence	-	54%

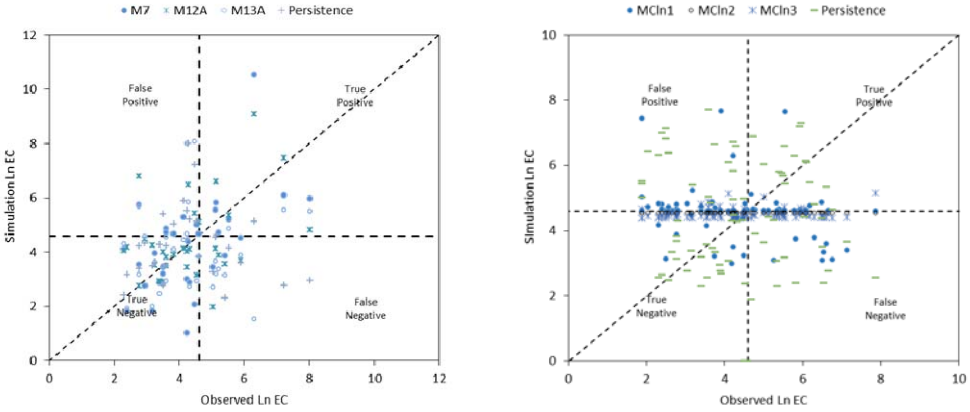
w.t - water temperature, *a.t* - air temperature



a)



b)



c)

Figure 2 Scatter Plots of Modelled and Observed E. coli Concentrations at a) Sunnyside Beach, b) Rouge Beach, c) Marie Curtis Beach

ACKNOWLEDGEMENTS

Authors of the paper would like to acknowledge Mahesh Patel and Jamie Duncan for their assistance in data collection, and Dr. Luigi Berardi for assistance in the EPR model development.

REFERENCES

1. MOE, *Water management policies & guidelines: Provincial water quality objectives* 1994, Ministry of Environment and Energy: Ontario, Canada. p. 1-67.
2. Nevers, M.B., et al., *Geographic relatedness and predictability of Escherichia coli along a peninsular beach complex of Lake Michigan.* J Environ Qual, 2009. **38**(6): p. 2357-64.
3. Amaral, N., *2009 Surface Water Quality Summary- Regional Watershed Monitoring Program.*, 2010. p. 26.
4. Ashbolt, N.J., et al., *Predicting pathogen risks to aid beach management: the real value of quantitative microbial risk assessment (QMRA).* Water Research, 2010. **44**(16): p. 4692-703.
5. Frick, W.E., Z. Ge., and R.G. Zepp, *Nowcasting and Forecasting Concentrations of Biological Contaminants at Beaches: A Feasibility and Case Study.* Environmental Science & Technology, 2008. **42**(13): p. 4818–4824.
6. *Great City, Great Beaches: Toronto Beaches Plan*, 2009, Toronto Water, City of Toronto: Toronto. p. 10.
7. USEPA, *Predictive Tools for Beach Notification : Review and Technical Protocol*, O.o.S.a.T. Office of Water, Editor 2010, U.S. Environmental Protection Agency. p. 1-61.
8. Zhang, G., B.E. Patuwo, and M.Y. Hu, *Forecasting with artificial neural networks:The state of the art.* Elsevier Science, 1998. **14**: p. 35–62.
9. He, L.M. and Z.L. He, *Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA.* Water Res, 2008. **42**(10-11): p. 2563-73.
10. Varma, S.S. and N. Vijayan, *Prediction of Fecal Coliform Concentration in Surface Water Using Artificial Neural Networks*, in *10th National Conference on Technological Trends*2009, College of Engineering Trivandrum.
11. Zhang, Z., Z. Deng, and K.A. Rusch, *Development of predictive models for determining enterococci levels at Gulf Coast beaches.* Water Res, 2012. **46**(2): p. 465-74.
12. Mas, D. and D. Ahlfeld, *The Development and Evaluation of Artificial Neural Networks for Modeling Indicator Organism Concentrations.* 2007.
13. Beale, M.H., M.T. Hagan, and H.B. Demuth, *Neural Network Toolbox-User's Guide*, 2013, The MathWorks, Inc.: Natick, MA, USA.
14. Legates, D.R., G. J., and M. Jr., *Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation.* Water Resources Research, 1999. **35**(1): p. 233-241.
15. Shamisi, M.H.A., A.H. Assi, and H.A.N. Hejase, *Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in Al Ain City – UAE*. Engineering Education and Research Using Matlab, ed. D.A. Ass. 2011, UAE: InTech. 480.
16. Tufail, M., L. Ormsbee, and R. Teegavarapu, *Artificial intelligence-based inductive models for prediction and classification of fecal coliform in surface waters.* Journal of Environmental Engineering, 2008. **134**: p. 789-799.
17. Giustolisi, O. and D.A. Savic. (2006). *A Symobolic Data-Driven Technique Based on Evolutionary Polynomial Regression*. Journal of Hydroinformatics, 8(3), 207-222. 2006.
18. Rezania, Mohammad, Akbar A. Javadi, and Orazio Giustolisi. *Evaluation of liquefaction potential based on CPT results using evolutionary polynomial regression*. Computers and Geotechnics 37.1 (2010): 82-92.
19. Doglioni, Angelo, et al. *Evolutionary polynomial regression to alert rainfall-triggered landslide reactivation*. Landslides 9.1 (2012): 53-62.
20. Savic, Dragan, et al. *Modelling sewer failure by evolutionary computing*. Proceedings of the ICE-Water Management 159.2 (2006): 111-118.