Proceedings of 2013 IAHR Congress

Comparison of machine learning classifier models for bathing water quality exceedances in UK

Andrew Duncan,
PhD Student, University of Exeter Centre for Water Systems, Harrison Building, North Park Road, Exeter, EX4 4QF, UK. Email: apd209@exeter.ac.uk

Deborah Tyrrell,
Scientific Officer Tidal Waters, Environment Agency of England and Wales, Manley House, Kestrel Way, Exeter, Devon, EX2 7LQ, UK. E-mail: debbie.tyrrell@environment-agency.gov.uk

Nicholas Smart,
Technical Specialist, Tidal Waters Team, Environment Agency of England and Wales, Manley House, Kestrel Way, Exeter, Devon, EX2 7LQ, UK. E-mail: Nick.Smart@environment-agency.gov.uk

Edward C Keedwell,
Senior Lecturer, University of Exeter Centre for Water Systems, Harrison Building, North Park Road, Exeter, EX4 4QF, UK. Email: E.C.Keedwell@exeter.ac.uk

Slobodan Djordjević,
Professor of Hydraulic Engineering, University of Exeter Centre for Water Systems, Harrison Building, North Park Road, Exeter, EX4 4QF, UK. Email: S.Djordjevic@exeter.ac.uk

Dragan Savić,
Professor of Hydroinformatics & Head of Engineering, University of Exeter Centre for Water Systems, Harrison Building, North Park Road, Exeter, EX4 4QF, UK. Email: D. Savic@exeter.ac.uk

ABSTRACT: The revised Bathing Water Directive (rBWD) (2006/7/EC) of the European Parliament requires monitoring of bathing water quality and, if early-warnings are provided to the public, it is permissible to discount a percentage of exceedance events from the monitoring process. This paper describes the development and implementation of both Decision Tree (DT) and Artificial Neural Network (ANN) based machine learning models for 8 beaches in south-west England, UK, as bases for early warning systems (EWS) and compares their performance for one beach. Weekly bacteria-count samples were gathered by the Environment Agency of England (EA) over a 12-year period from 2000-2011 during the 20-week bathing season and this data is used to calibrate and test the models. Daily sampling data were also collected at 5 of the beaches during the 2012 season to provide more robust validation of the models. As a benchmark, models are also compared with use of simple thresholds of antecedent rainfall to classify water quality exceedances. Evolutionary Algorithm-based optimisation of the ANN models is employed using single-objective approach using area under the Receiver Operating Characteristic (ROC) curve as fitness function. The optimum operating point is established using a weighting factor for the relative importance placed on false positives (passes) and false negatives (exceedances). The models use a number of input factors, including antecedent rainfall for the catchment adjacent to each bathing beach. A possible technique for automating selection of inputs is also discussed.

KEY WORDS: Artificial neural network, Bathing water directive, Decision tree, Early warning system, Water quality prediction.

1 INTRODUCTION

The revised Bathing Water Directive (rBWD) (2006/7/EC) was introduced by the European Commission in 2006. It will take over from the current Bathing Water Directive (76/160/EEC) in 2015 and sets more stringent water quality standards. The rBWD also places a strong emphasis on providing information to the public on the quality of bathing waters to allow them to make an informed choice where to bathe. Heavy rainfall can result in water running off the land, picking up contaminants and overloading the sewerage system. This can quickly have an adverse effect on water quality.

The Short Term Pollution provision of the rBWD allows for up to fifteen percent of samples taken during short term pollution events to be discounted from the four year compliance analysis, provided the public are advised in advance that water quality may be unsuitable for bathing, and measures are in place for water quality improvements. Where a bathing water fails to meet the "Sufficient" standard of the rBWD in 2015, signs will be put up from 2016 advising people not to bathe. This may impact the tourist industry/local economy.

The Environment Agency for England (EA) has developed a set of data-driven modelling tools which predict whether water quality is likely to be above or below a pre-determined bacteriological threshold each day, using multiple triggers from real-time rainfall data and tidal predictions. As part of the joint-agency 'Bacti' project, the University of Exeter Centre for Water Systems (CWS) has also developed machine-learning models, based on Artificial Neural Networks (ANNs). This paper summarises recent work on validation of these models and presents a comparison with water quality predictions using a simpler method of single rainfall triggers that can be applied to a larger number of bathing waters.

.

2 METHODS OF MODEL BUILDING AND TESTING

2.1 Decision Tree Models

Following previous studies conducted by the Environment Agency between 2007 and 2009 (Tyrrell, 2010), predictive tools for eight bathing waters were built and tested in 2012. The Environment Agency's models were built using a script developed from the Classification Trees module within IBM SPSS™ Statistics software (IBM, 2011; Mola, 1998). The decision tree procedure creates a tree-based classification by taking a set of data points and categorising them into groups of a dependent (target) variable based on values of independent (predictor) variables. The target variable was the 'pass' or 'fail' of a bacteriological sample against a threshold of 500 faecal coliforms/100ml and/or 200 faecal streptococci/100ml. The predictor variables were antecedent rainfall totals for 24hrs, 48hrs, 72hrs, 96hrs, and 120hrs, tidal range, and tidal state. The bacteriological threshold was developed by the World Health Organisation (WHO) and relates to accepted health standards for bathing water quality (Kay et al., 2004).

Of the various tree growing methods available in SPSS™, previous studies showed that the CART (Classification and Regression Trees) method gave the most accurate results. CART is a non-parametric technique that produces a binary decision tree constructed by splitting a node into two child nodes repeatedly, beginning with the root node (parent) that contains the whole data set. The data are split into segments that are as homogeneous as possible with respect to the target variable. Each branch of a tree ends with a terminal node which is uniquely defined by a set of rules that may then be applied to predict future events. The complexity of the tree depends on the underlying distribution of data, and it follows that the stronger the relationship between target and predictor variables the simpler the tree. For this study we were focussed on correctly predicting poor water quality, and protecting public health, therefore the models were weighted to minimise the number of incorrectly predicted exceedances of the bacteriological threshold, but could be modified to minimise restrictions on bathing. An example decision tree is presented in Figure 1.

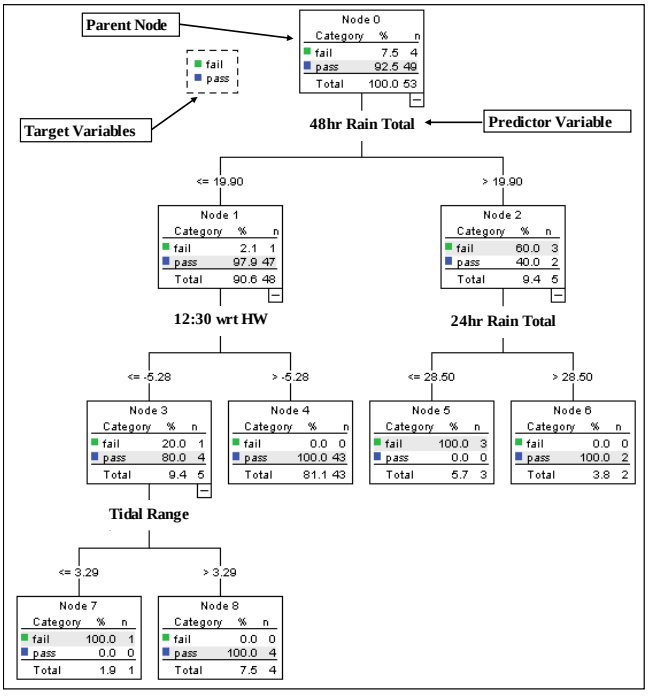

Figure 1 Example Decision Tree

Models were built and tested for the following bathing waters located in South West England: Mothecombe, Seaton (Cornwall), East Looe, Readymoney, Par, Porthluney, Ilfracombe Wildersmouth, and Burnham Jetty. The bacteriological and environmental data for each bathing water for 2000 to 2011 were compiled and analysed by a database tool. Samples exceeding the bacteriological threshold in dry weather (quantified by a 96hr rainfall total <5mm) were removed from the dataset, since these would have occurred by events we cannot predict using rainfall data, e.g. wrongly connected waste water systems, bird and dog fouling. For each bathing water, an initial decision tree was built using the data from 2000 to 2006, and the resulting rules from the terminal nodes were applied blindly to the 2007 data for validation. The models were then rebuilt including the 2007 data and resulting rules applied blindly to the 2008 data. This iterative process of calibration and validation was continued until all the data to 2011 had been included in the model, giving a total of five validation trees per bathing water plus a tree for investigational use in the 2012 bathing season. The results for Seaton (Cornwall) bathing water are presented in Table 1 (see section 3.1).

2.2 ANN Models

Using the same datasets, ANN classifier models were built and tested using MATLAB ® V2012a (Mathworks, 2012). The models were based on the RAPIDS package described in earlier publications: (Duncan et al., 2011, 2013a, 2013b). Two-layer fully-connected feedforward ANNs were used. Inputs were not time-lagged explicitly, due to the very long timestep for the samples. Instead, the implicit time-lagging present in the antecedent rainfall totals was exploited. Figure 2 illustrates the architecture of the learning scheme used for training the ANNs. For each beach modelled, a Leave-One-Out-Cross-Validation (LOOCV) methodology (Cawley and Talbot, 2003) was used, in which the samples for each bathing season 2000-2011 were used as the data blocks. Thus an ensemble of 12 ANN models was trained, and then each tested on a different remaining ("left-out") block (bathing season of samples). The results for all 12 models were then aggregated and are summarised in section 3.2.

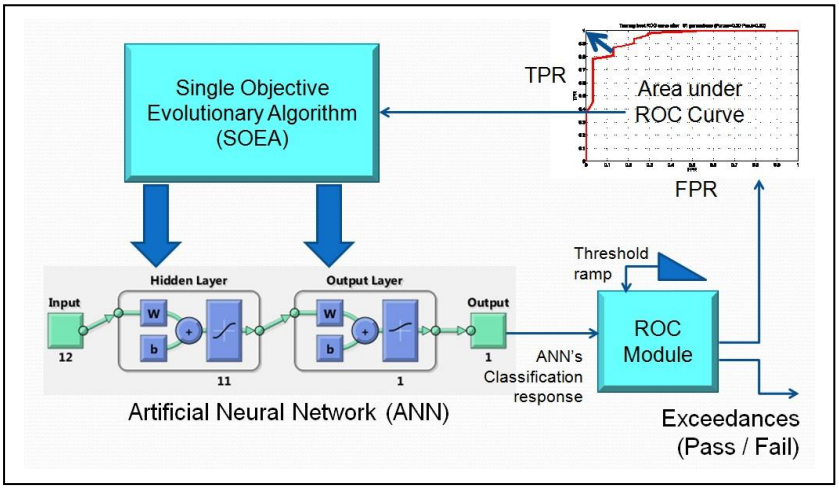

Figure 2 ANN Learning Scheme Architecture

For each member of the ensemble, the same set of 12 input signals was applied to the ANN; these included antecedent rainfalls (x5) and tide height and state (as for DT models), plus salinity, tidal height at high water and at sampling time, tidal range at standard port, tide level class (spring/mean/neap) and timestamp. The number of neurons in the hidden layer was varied during experimentation. The classification responses from the single output neuron were compared to a (ramped) set of threshold values covering the span of actual model output values, and the area under the trade-off curve of true positive (pass) rate (TPR) versus false positive rate (FPR) was computed. Such a curve is known as a Receiver Operating Characteristic (ROC), the purpose of which is to establish the optimum trade-off between FPR and False Negative Rate (FNR = 1 – TPR). Optimisation of the ANN weights and biases was achieved using a single-objective realisation of NSGA-II (Deb et al., 2002), a widely used evolutionary algorithm. In this single-objective case, crowding distance was neglected, since each rank of pareto dominance had exactly one member. The chromosome of values ($[-1,+1] \in R^N$) in the decision space for each member of the population represented the values of ANN weights and biases (Yao, 1999); whereas the fitness function used minimization of cost (1- the area under the ROC curve) in the 1-dimensional objective space. At each generation (epoch) of the training, fitter members of the population had a higher probability of being selected as parents for reproduction of the new child solutions for the next generation. Population size of 100 was found to be adequate and probabilities of crossover and mutation were also varied during experimentation to determine reasonable value ranges.

Stidson et. al. (2012) proposed use of a modified F measure to evaluate model performance using a weighting (a=4 in (1)) to minimise the number of incorrectly predicted passes (levels below the bacteriological threshold): False Positives (*FP* in (1)).

$$F = \frac{(1+a)TN}{(1+a)TN + aFP + FN} \tag{1}$$

where: F = modified F measure; *TN* = number of true negative samples; *FP* = number of false positive samples and *FN* = number of false negative samples (negative = fail; positive = pass). This study adapted this method to use ROC curves, effectively stretching the x-axis (FPR) for values of a>1 and shrinking it for a<1. Euclidean distance (E) of each point, to the ideal (FPR=0; TPR=1), on the scaled ROC curve was calculated and the optimum operating point (E_{opt}) determined using (2).

$$E_{\text{opt}} = \min\left(\sqrt{(a \cdot FPR)^2 + (1-TPR)^2}\right) \tag{2}$$

Analysis of influence of inputs on the models' output responses was performed using:

$$W_{io} = W_1 \bullet W_2 \quad (3)$$

where: W_{io} = input-to-output influence vector; W_1 = ANN hidden layer weight matrix; W_2 = ANN output layer weight vector and • represents matrix multiplication. Thus W_{io} has dimensions of N_{in} x N_{out} where N_{in} is number of input signals and N_{out} = 1, the number of output neurons (signals). By comparing the ranges of each coefficient in W_{io}, over the ensemble of ANN models produced, it may be possible to rank the relevance of each input based on inter-quartile range.

3 RESULTS & DISCUSSION

3.1 Decision Tree Models

Table 1 shows that the Decision Tree models for Seaton (Cornwall) are capable of predicting both "fail" and "pass" reasonably well, with a total of 10 out of 13 "fails" predicted correctly, and 79 out of 86 "passes" predicted correctly. The models were also tested blindly on data for a full bathing season (1st May to 30 September) for each year from 2007 to 2011 to count how many exceedances would typically be predicted (advisory signs put up). The number of public advisories per bathing season ranged from 10 in 2009 and 2011 to 28 in 2008 (a very wet year with regard to long-term average rainfall).

The final models, which included all the data to 2011, were tested in 2012. Five of the eight bathing waters (Seaton (Cornwall), East Looe, Readymoney, Par, and Porthluney) were sampled daily (excluding weekends and Bank Holidays) throughout the bathing season to provide a larger dataset, and hence a more robust model validation than previously available. A second assessment was completed using single 24hr rainfall triggers of 0, 5, 10 and 15mm, and the two sets of results compared. The results are presented in Figure 3 and Table 2.

The decision tree models gave a similar level of accuracy to that produced using a single 24hr rainfall trigger of 10mm (red diamonds and green squares in Figure 3). This is important because the decision tree method relies on having a good number (typically >10%) of data exceeding the bacteriological threshold in order to train the model (i.e. the bathing waters with poorer water quality), whereas a single 24hr rainfall trigger may be applied to any number of bathing waters of varying quality.

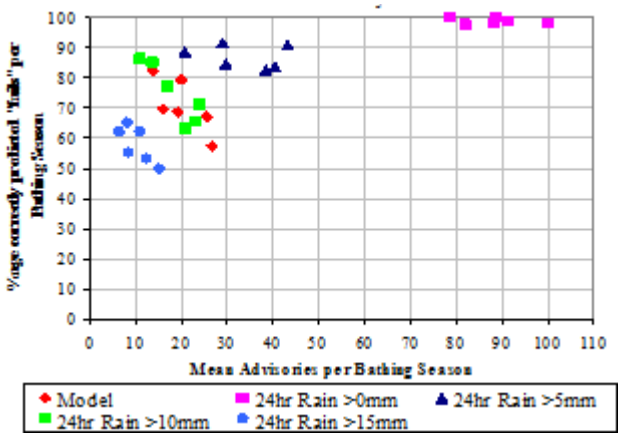

Figure 3 Comparisons of Decision Tree Models with Simple Triggers – Advisories vs Model Accuracy 2007-2012

Figure 3 also shows that as the rainfall trigger decreases, the number of "fails" correctly predicted increases, and so does the number of public advisories. This is important because to implement an operational bathing water warning system, the rainfall triggers set will be determined not only by the absolute accuracy of the predictions, but also with a recognition of the number of public advisories deemed acceptable by beach managers.

Table 1 Decision Tree Validation Results for Seaton (Cornwall)

Data Used to Build Model	Samples Used to Build Model	Test Year	Samples In Test Year	Sample Exceedances in Test Year	Predicted Pass, Sample Pass (PP)	Predicted Fail, Sample Pass (FP)	Predicted Fail, Sample Fail (FF)	Predicted Pass, Sample Fail (PF)	Total Signs in Bathing Season (153 days)	Model Predictor Variables
2000-06	140	2007	20	1	16	3	0	1	17	Rainfall only (24, 48, 72, 96, and 120hr)
2000-07	160	2008	20	7	10	3	5	2	28	Rainfall only (24, 48, 72, 96, and 120hr)
2000-08	180	2009	19	1	18	0	1	0	10	Rainfall only (24, 48, 72, and 96hr)
2000-09	200	2010	20	3	17	0	3	0	11	Rainfall only (24, 48, 72, and 96hr)
2000-10	220	2011	20	1	18	1	1	0	10	Rainfall only (24, 48, 72, and 96hr)

Table 2 Prediction Results for 2012

Bathing Water	Sample Fails	Dry Weather Fail	Samples without Discount	rBWD 2012 without Discount	24hr Rain Trigger	48hr Rain Trigger	%age Correct Predictions	Samples with Discount	Samples Discounted	rBWD 2012 with Discount	Advisories 2010	Advisories 2011	Advisories 2012	Total Advisories 2010 to 2012	Method
Mothecombe	15	0	80	Poor	10		**90**	70	10	Good	13	18	23	54	**Simple trigger**
							84	73	7	Sufficient	7	12	33	**52**	Decision tree
Seaton (Cornwall)	9	0	80	Poor	8.5		**96**	72	8	Good	15	11	22	48	Simple trigger
							94	74	6	Sufficient	11	10	12	**33**	**Decision tree**
East Looe	13	4	80	Poor		19	84	75	5	Poor	17	8	25	**50**	Simple trigger
							86	75	5	Poor	12	12	30	54	**Decision tree**
Readymoney	8	0	80	Sufficient	10		89	75	5	Good	11	7	22	40	Simple trigger
							90	76	4	Good	8	13	18	**39**	**Decision tree**
Par	6	0	80	Sufficient	15		**94**	77	3	Good	10	3	11	**24**	**Simple trigger**
							94	78	2	Sufficient	9	6	19	34	Decision tree
Porthluney	15	2	80	Poor	7.7		**90**	70	10	Sufficient	13	8	23	**44**	**Simple trigger**
							86	72	8	Sufficient	13	15	26	54	Decision tree
Ilfracombe Wildersmouth	22	3	80	Poor		9	66	70	10	Poor	34	45	61	140	Simple trigger
							70	69	11	Poor	34	38	40	**112**	**Decision tree**
Burnham Jetty	9	0	80	Poor	10		89	76	4	Poor	3	6	14	**23**	Simple trigger
							93	76	4	Sufficient	2	14	25	41	**Decision tree**

The results for the 2012 investigations are presented in Table 2 with additional fine tuning of the single 24hr or 48hr rainfall trigger levels (mm). Also presented is an assessment of how the rBWD classification for 2012 would be improved by being able to discount samples that were correctly predicted to exceed the bacteriological threshold.

It is clear from Table 2 that the rBWD classification can be improved with discounting, and that both methods are applicable. The conclusion whether to use single rainfall triggers or decision trees is a balance between the accuracy of the predictions, the number of advisories, and whether rBWD class change is achieved. The simple rainfall trigger method was considered appropriate for Mothecombe, Par, and Porthluney. The decision tree method was considered appropriate for Seaton (Cornwall), East Looe, Readymoney, Ilfracombe (Wildersmouth), and Burnham Jetty. For East Looe and Ilfracombe (Wildersmouth), further work will be required to improve the water quality and reduce the number of times water quality is impacted in dry weather, so that rBWD class change through discounting can be achieved.

3.2 ANN Models

Figure 4 illustrates a set of ROC curves for the Seaton catchment produced from an ensemble of 12 ANN models, one for each of the sampling years 2000-2011. All ANNs have 12 neurons on the hidden layer and 12 input signals. Three trade-off curves are shown: training, validation and test. The test curve represents the aggregation of test results from all of the models, each model having used the sample results from one of the 12 (Y_t) years (2000-2011) against which to test its performance following training. For each of these 12 models, the succeeding year of samples (Y_t + 1) mod 12 was excluded from the training dataset and used to validate progress during training, to perform early-stopping and avoid over-fitting. The remaining 10-years' samples for each model were used as the training dataset. Validation and training curves are similarly aggregations of results from all 12 models. Results are preliminary and demonstrate proof of concept. Using the FP:FN importance weighting factor, a=4, the optimum operating point for the ensemble, based on minimum Euclidean distance to the ideal, is also shown as -×. This is compared to the test operating point for the ensemble of 5 DT-based models for Seaton, shown as + and to the simple threshold of 24-hour antecedent rainfall greater than 10mm, shown as ◆.

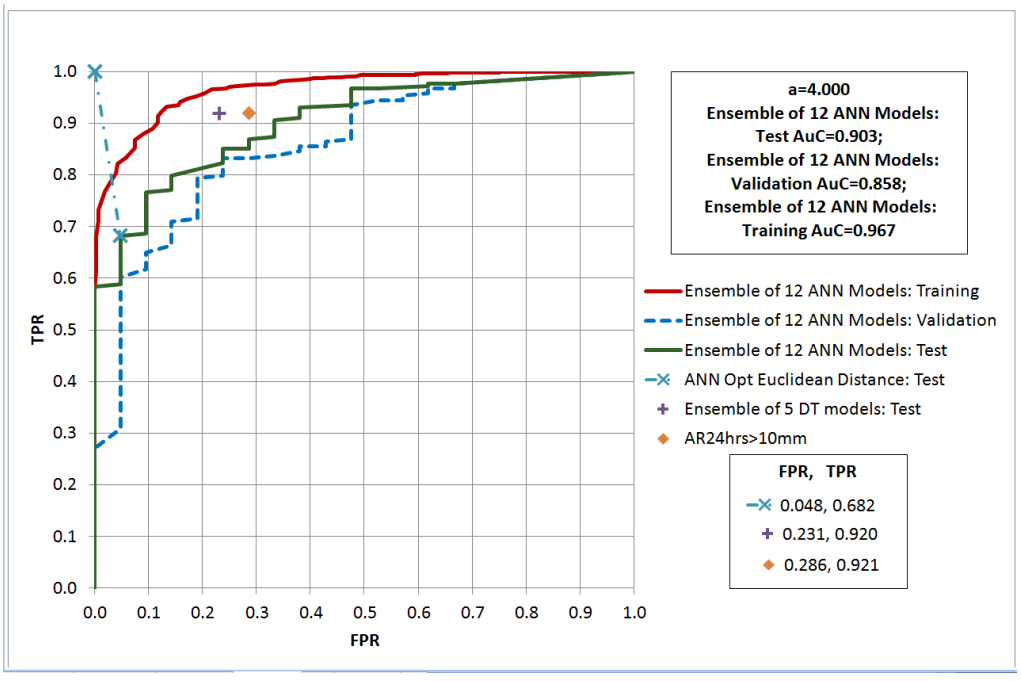

Figure 4 Seaton: ANN and DT model ensemble performance compared (2000-2011)

Although the performance of the ANN model ensemble during training exceeds the test performance of the ensemble of 5 Decision Tree models, the ANN test performance is significantly worse. However, the ROC curve approach illustrates the importance of choosing the optimum operating point based on the weighting factor (a). Further work is needed to ascertain if more robust and/or improved results could be achieved by instituting a voting scheme for the ANN ensemble members, rather than merely aggregating the results from all ANN models, as at present.

Analysis of influence of inputs on the models' output responses is shown in figure 5 for the ensemble of 12 ANN models, using equation (3) for each model. Each box and whisker shows the spread of values of $W_{io}=W_1xW_2$ for the given input signal over the ensemble of 12 models. In addition to 24hrs antecedent rainfall (AR24), Salinity and Tidal Range at Standard Port were also shown to be relevant signals.

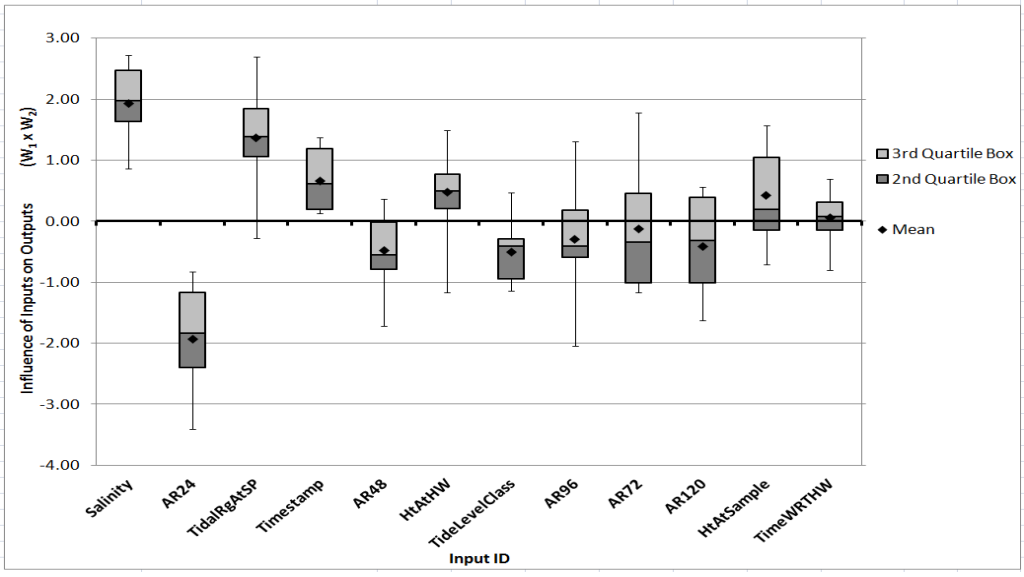

Figure 5 Seaton model: 12 ANN ensemble analysis of influence of inputs on outputs

By using (for example) the inter-quartile range, it may be possible to rank the input signals in terms of their relevance to the skill of the model ensemble and hence provide a means of automating input selection during a hybrid training algorithm. Further work is needed on this.

4 CONCLUSIONS

Both decision tree and artificial neural network models performed well in classifying potential bathing water quality exceedances and could therefore be used in a real-time EWS. However, performance of both has been shown only to be comparable with a classifier using a simple threshold of antecedent rainfall for 24 hours exceeding 10mm. Further work is needed to establish if model skill can be improved through improvements to optimisation through automating selection of number of ANN hidden units and input signals used.

Ultimately, effectiveness of models will be evaluated on whether rBWD class change through discounting can be achieved. For deployment nationally in a live system, such models will also need to use only input signals that are readily available from existing data sources, although these sources are increasing all the time. This study has not included for example combined sewer overflow spill data as this was not available for the period 2000-2011. However, this may provide a means of enhancing performance of models for some catchments. It is expected that the performance will be significantly enhanced because CSO activity will provide additional valuable information, which is the next step in this research.

ACKNOWLEDGEMENT

Our thanks are also due to our partners in Environment Agency and South West Water Plc, who have provided funding for the project as well as invaluable information, advice and discussion.

References

Cawley, G.C., Talbot, N.L.C., 2003. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. Pattern Recognition 36, 2585–2592.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. Evolutionary Computation, IEEE Transactions on 6, 182–197.

Duncan, A., Chen, A.S., Keedwell, E., Djordjevic, S., Savic, D.A., 2011. Urban flood prediction in real-time from weather radar and rainfall data using artificial neural networks, in: IAHS Red Book Series No. 351, 58. Presented at the Weather Radar and Hydrology International Symposium, International Association of Hydrological Sciences, Exeter, UK.

Duncan, A.P., Chen, A.S., Keedwell, E.C., Djordjevic, S., Savic, D.A., 2013a. RAPIDS: Early Warning System for Urban Flooding and Water Quality Hazards (Extended Abstract), in: AISB 2013. Presented at the Artificial Intelligence and Simulation of Behaviour Conference; Machine Learning in Water Systems Symposium, AISB, University of Exeter.

Duncan, A.P., Keedwell, E.C., Djordjevic, S., Savic, D.A., 2013b. Early Warning System for Bathing Water Quality (Poster), in: Bathing Waters 2013. Presented at the Bathing Waters 2013, Defra, England, Southport, UK.

IBM, C., 2011. IBM SPSS Decision Trees 20 User Manual.

Kay, D., Bartram, J., Prüss, A., Ashbolt, N., Wyer, M.D., Fleisher, J.M., Fewtrell, L., Rogers, A., Rees, G., 2004. Derivation of numerical values for the World Health Organization guidelines for recreational waters. Water Research 38, 1296–1304.

Mathworks, T., 2012. MATLAB® & Simulink® Release Notes for R2012a.

Mola, F., 1998. Classification and Regression Trees Software and New Developments, in: Rizzi, A., Vichi, M., Bock, H.-H. (Eds.), Advances in Data Science and Classification, Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg, pp. 311–318.

Stidson, R.T., Gray, C.A., McPhail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water and Environment Journal 15.

Tyrrell, D., 2010. Bathing Water Quality Forecasting System 2009 Trial, Modelling Report (Internal, unpublished). Environment Agency.

Yao, X., 1999. Evolving artificial neural networks. Proceedings of the IEEE 87, 1423–1447.