



A neural-network-based classification scheme for sorting sources and ages of fecal contamination in water

Gail M. Brion*, T.R. Neelakantan, Srinivasa Lingireddy

Department of Civil Engineering, University of Kentucky, 161 Raymond Bldg., Lexington, KY 40506-0281, USA

Abstract

Artificial neural networks (ANNs) were successfully applied to data observations from a small watershed consisting of commonly measured indicator bacteria, weather conditions, and turbidity to distinguish between human sewage and animal-impacted runoff, fresh runoff from aged, and agricultural land-use-associated fresh runoff from that of suburban land-use-associated-fresh runoff. The ANNs were applied in a cascading, or hierarchical scheme. ANN performance was measured in two ways: (1) training and (2) testing. An ANN was able to sort sewage from runoff with <1% error. Turbidity was found to be relatively unimportant for sorting sewage from runoff, while gross measurements of gram-negative and gram-positive bacteria were required. Predictions clustered tightly around the known values. ANN classification of aged suburban runoff from fresh, and agricultural runoff from suburban was accomplished with >90% accuracy. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Sewage; Runoff; Neural networks; Fecal indicators; Land-use

1. Introduction

Artificial neural network (ANN) modeling framework has been used increasingly in various fields of science and technology. In recent years, ANNs have found a number of applications in the area of water quality modeling. Water quality is influenced by many factors such as flow rate, contaminant load, medium of transport, water levels, initial conditions and other site-specific parameters. The estimation of such variables is often a complex and non-linear problem, making it suitable for ANN application. Rogers and Dowla [1] employed an ANN, which was trained using a solute transport model, to investigate hypothetical scenarios of one or several contaminant plumes moving through a ground-water region with a number of pumping wells. Basheer and Najjar [2] used a three-layer ANN to predict the breakthrough time in a fixed-bed adsorption system. Maier and Dandy [3] illustrated the utility of ANNs for estimating salinity at the Murray Bridge on the River

Murray in South Australia. Hutton et al. [4] used ANNs to enhance the capability of an existing model for predicting THM formation by including variable reaction conditions. Sandhu and Finch [5] used ANNs to relate flow conditions and gate positions in the Sacramento San Joaquin Delta to salinity levels in the interior and along the boundary of the delta to estimate flow in the Sacramento River thereby meeting salinity standards. They found that ANN models are better than simulation models and the commonly used statistical models for this application. Starrett et al. [6] employed an ANN to predict pesticide leaching through turf grass-covered soil. Lek et al. [7] developed ANN model for relationships between macro-drainage area characteristics and nutrient levels in streams. It had eight independent input variables of watershed parameters (five land-use features, mean annual precipitation, animal unit density and mean stream flow) and two dependent output variables (total and inorganic nitrogen concentrations in the stream). Recently, neural networks were used for identifying land-use-related non-point sources of microbial contamination [8] and identification of *Cryptosporidium* peak concentrations at the inlet to a water treatment plant on the Delaware River [17].

*Corresponding author. Tel.: +1-606-257-4467; fax: +1-606-257-4404.

E-mail address: gbrion@engr.uky.edu (G.M. Brion).

To fully understand the risks involved with fecal pollution in water, and develop risk-based early warning systems for public treatment works relying upon polluted surface water sources, the source of fecal material and the time the fecal material has been in the environment must be known. By studying shifting bacterial populations, and being aware of their constraints, an estimate of the relative age of fecal pollution in runoff [9,10], and the most probable fecal source [8] has been demonstrated to be predictable on a watershed level. It was the objective of this study to extend our original ANN-based non-point source classification approach, in a hierarchical manner, and determine if neural network classification could identify the unique signature of human sewage from animal-impacted runoff, as well as recent runoff from aged in a local watershed.

2. Materials and methods

2.1. Site description and sampling

In this multi-year study, water samples were collected at least weekly from nine runoff collection sites within a small watershed, and at the inlet to the Town Branch wastewater treatment plant of Lexington, KY, USA. This watershed has been under study since 1996 and the sources of runoff and collection system are well defined. Samples were taken during all the seasons except deep winter where freezing or dry precipitation in the form of snow was problematic. Runoff sample sites in the watershed were predominantly impacted by animal fecal material, not sewage, as has been indicated before by specialized analysis for human and environmental lipopolysaccharides, hydroxy fatty acids, coprostanol, and bacteriophage testing [8]. Eight runoff-impacted sites were classified according to associated land use, estimated fecal age, and flow characteristics as follows:

- flowing agricultural (Evans Mill, Gess, Jacobson),
- flowing suburban (Armstrong, Shelby, Stonewall),
- impounded suburban (Andover, Squires).

Inlet samples of raw sewage (RS) from the treatment plant were classified as RS. This brought the number of sample sites included in the study to nine. The flowing agricultural (FA) sites are agricultural creeks with direct contact between cattle and water. The flowing suburban (FS) sites are small creeks and drainage canals that carried runoff through well-sewered, residential neighborhoods. FS sites received less total quantities of direct fresh fecal input than FA during dry weather, but during rainfall, FS sites would be inundated with fecally contaminated suburban scour. The impounded subur-

ban (IS) sites collect drainage and runoff into small retention ponds. The average age of fecal materials in these ponds was greater than all other sample sites due to retention. The order of the categories in terms of fecal age, from freshest human to oldest animal would be RS, FA, FS, and then IS.

For each site in the study, samples were collected in sterile, 3-L polypropylene jugs, and immediately placed on ice to be transported back to the lab. Mid-stream, mid-depth sampling protocol was used for collection so that streambeds remained undisturbed. Sampling was halted at a site if the stream became dry, or ceased to have observable flow, or if an impoundment dried up. Samples were received and processed within 6 h from the time of collection. For each sample, various water quality and watershed condition parameters were measured and recorded. These included (1) weather (rain or no rain), (2) background colonies, (3) atypical coliforms, (4) total coliforms, (5) fecal coliforms, (6) fecal streptococci, and (7) turbidity. Measured precipitation was obtained from records maintained at a nearby water treatment plant weather station. Sampling observations were coded as rain if there had been 0.5" of wet precipitation in 24 h, or 1.0" in 48 h, or 2.5" in 96 h prior to sampling. A record of the visual appearance of the sample site was maintained to provide a record of pertinent observations, such as high or low flow conditions, the degree of direct animal contact with water at the site, and sanitary survey observations.

Each data observation comprised a definition of meteorological conditions (0 or 1 for non-rain or rain, respectively), log-transformed bacterial concentrations for total, atypical and background colonies (BG) resulting from the coliform analysis, log-transformed fecal coliforms, log-transformed fecal streptococci, and turbidity as NTU. Although pH and coliphage were measured on all samples, these results were not used as model input data, nor were the results from biolipid and male-specific RNA coliphage typing analyses, although these were critical in excluding the presence of overt human sewage in the runoff. The result of this sampling effort yielded a total of 343 data observations. Some of these observations contained values of too numerous to count (TNTC) or below detection level (BDL) for bacteria, which were assigned values of 10^9 and 10^2 CFU/L, respectively, to provide a full database, complete with all the 7 measured parameters for each observation for model training and verification. Before attempting a statistical analysis or neural network modeling, the entire database was randomly shuffled to prevent any seasonal or order influences upon training and prediction. Table 1 shows typical data observations used for training/testing of all neural network models. Statistical analysis of the database was done using SigmaStat (SPSS, version 2.03) on the incomplete, or unfilled, database, before values of

Table 1
A typical sample of data observations used as input to ANN

Weather	BG	AC	TC	FC	FStrep	Turbidity	Site name/type	Date
0	6.332	7.237	7.138	9.000	5.985	190	Gess/FA	07/21/99
0	5.019	5.911	4.859	5.002	5.061	5.5	Evans Mill/FA	05/24/99
1	5.813	5.431	5.301	4.431	4.975	22	Andover/IS	12/14/98
0	4.477	5.842	4.484	9.000	4.431	1.5	Stonewall/FS	07/12/99
1	4.243	5.574	4.120	4.168	3.922	8.2	Squire/IS	06/28/99
0	4.727	4.320	4.327	2.824	3.283	2.5	Shelby/FS	11/23/98
1	5.019	5.130	4.038	2.736	5.081	11	Gess/FA	01/14/99
0	5.336	5.255	4.339	3.351	3.599	41	Stonewall/FS	11/19/98
0	2.000	5.439	3.980	3.788	3.791	2.4	Shelby/FS	06/02/99
0	4.237	4.765	3.772	3.226	3.073	1.4	Armstrong/FS	03/02/99
0	0.000	4.761	4.202	4.126	3.714	2.1	Armstrong/FS	05/11/99

Note: All bacterial concentrations are expressed as log CFU/L.

TNTC and BDL were replaced by 10^9 and 10^2 CFU/L, respectively. Before ANN modeling, all input parameter values to the neural network were rescaled to follow a maximum of 1 and a minimum of 0 by dividing with the maximum value.

2.2. Bacterial analysis

2.2.1. Total coliforms

Membrane filtration with mEndo broth- (Difco) soaked adsorbent pads, incubation at $35 \pm 0.5^\circ\text{C}$ for 22–24 h, after which total, atypical, and BG were counted. Three dilutions and two replicates per dilution were examined per sample. All organisms that produced a red colony with a metallic sheen were counted as total coliforms. Atypical colonies appeared as pink to dark red without sheen, and the BG appeared white or colorless [11].

2.2.2. Fecal coliforms

Membrane filtration using M-FC broth- (Difco) soaked adsorbent pads, incubation at $44.5 \pm 0.2^\circ\text{C}$ submerged in a water bath for 22–26 h incubation after which blue colonies were counted as fecal coliform bacteria [11]. Three dilutions and two replicates per dilution were examined per sample.

2.2.3. Fecal streptococci

Membrane filtration using a KF streptococcus agar (Difco), incubation at $35 \pm 0.5^\circ\text{C}$ for 48 h after which dark red to pink colonies were counted as fecal streptococci [9]. Three dilutions and two replicates per dilution were examined per sample.

2.2.4. Turbidity

Turbidity for each site was measured in the laboratory using a Hach 2100 A turbidimeter.

2.3. Neural network model

The model used for all classification attempts was a standard, three-layer, backpropagation, neural network with N input nodes, L hidden nodes, and K output nodes, which can be mathematically expressed as:

$$O_{Pk} = f_1 \left(\sum_{j=1}^L w_{jk}^o f_2 \left(\sum_{i=1}^N w_{ij}^h x_{Pi} + b_1^j \right) + b_2^k \right) \quad \forall k \in 1, 2, \dots, K, \quad (1)$$

where O_{Pk} is the output from the k th node of the output layer of the network for the P th vector (data point), x_{Pi} the inputs to the network for the P th vector (data point), w_{jk}^o the connection weight between the j th node of the hidden layer and the k th node of the output layer (refer Fig. 1), w_{ij}^h the connection weight between the i th node of the input layer and the j th node of the hidden layer, b_1^j and b_2^k are bias terms, and $f_1(\cdot)$ and $f_2(\cdot)$ are activation functions. The activation function used was a logistic sigmoid function which produces output in the range 0–1. Logistic sigmoid function has a form as given below:

$$f(x) = \frac{1}{1 + e^{-x}}. \quad (2)$$

Suitable activation function for the hidden units, $f_2(\cdot)$, is needed to introduce non-linearity into the network, which gives the power to capture non-linear relationship between input and output. For the output units, the selection of activation function, $f_1(\cdot)$, is based on the distribution of the target values. Bounded activation functions such as the logistic sigmoid function are particularly useful when the target values have a bounded range. The number of nodes in the hidden layer was set to be twice that of the input layer based on the experience gained from earlier studies by the authors [8,12].

The method of training the model was an iterative non-linear optimization approach using gradient descent

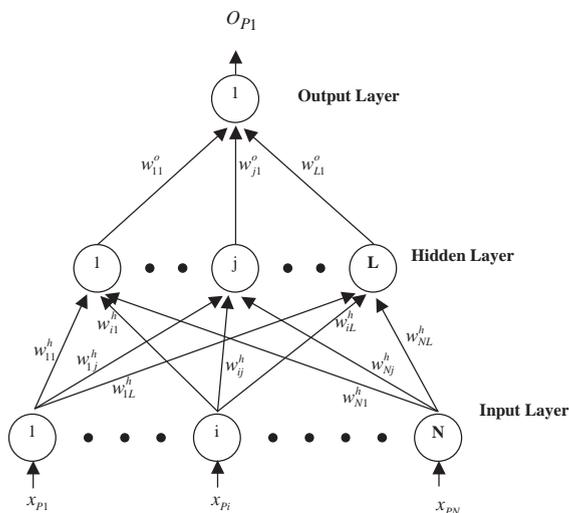


Fig. 1. Three-layer feedforward neural network architecture.

search technique on a randomized sort of the database of observations. In training, the weights on each connection were adjusted to yield the minimum error between the computed output pattern and the desired output pattern based on the method of least squares. The derivation of the steps of back propagation network is available in the books of Freeman and Skapura [13], Master [14] and many others. The basic procedure used to train the network is embodied in the following steps: (a) apply input observations from training set to the network and calculate the corresponding output values, (b) compare the computed output with the known output values and determine a measure of the error, (c) determine corrections (increase or decrease) to the connection weights, (d) apply the corrections to the weights and (f) repeat items (a)–(e) with all the training vectors until the error for all vectors in the training set is reduced to an acceptable value. The procedure outlined in steps (a)–(f) was developed into a computer program “NuroSort”. The entire program including data interface, training and testing phases of the neural network, was developed in-house at the University of Kentucky.

The same type of basic three-layer model structure was trained and tested repeatedly on a single randomly shuffled database to determine if it could distinguish between human-sewage- and animal-impacted runoff, and in-between the types and ages of animal-impacted runoff. Two strategies adopted for testing the applicability of ANN for each selected classification were (1) to test the capability of the ANN to correctly predict the output for the given input sets that were originally used to train the network (accuracy performance) and, (2) to test the capability of the ANN to correctly predict the output for the given input sets that were not in the training set (generalization performance). An individual

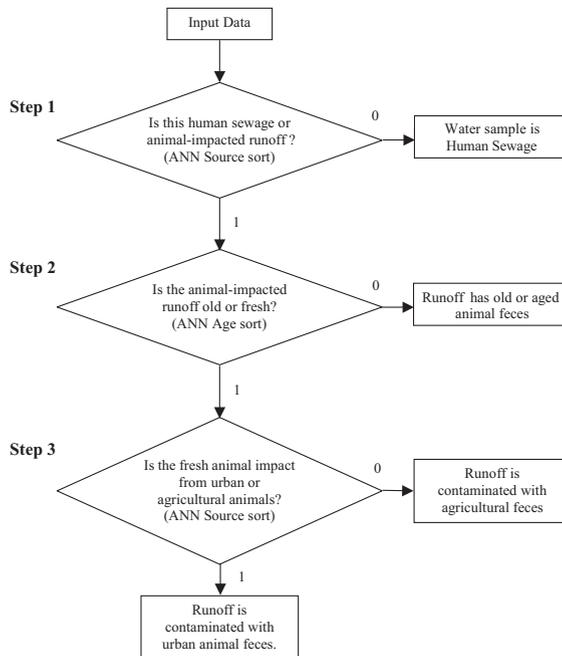


Fig. 2. Algorithm for hierarchical ANN classification scheme.

model’s performance was judged upon the ratio of the total number of inaccurate predictions to the data set under consideration. Inaccurate predictions were investigated for reasons that might have confounded the model by going back to the field testing log and reviewing the conditions observed at the site on that date.

The modeling effort followed the flowchart shown in Fig. 2. First, sewage was identified from all other runoff, then the runoff types were sorted from the data set left after sewage was eliminated, with IS runoff samples sorted from all others in the last step. Each step used a neural network model, trained on data specific to the classification being made, of N:2N:1 architecture and the same set of input parameters on databases that decreased with each step. In all cases, the number of iterations (epoches) required for training was $\leq 20,000$.

3. Results and discussion

3.1. Parameter selection

It must be noted that the fecal indicator parameters used for neural network modeling are generally inclusive rather than exclusive, and were selected for this characteristic, and with regard to analyses used in historical surface water quality databases of the Kentucky River. The term inclusive as used in this paper means that the media were designed to detect a group of related organisms rather than a single

organism. Two of the media used, mEndo broth and KF streptococcus agar, have been criticized for a general lack of specificity, need for further confirmation testing, and high rates of false positives [11,15]. These media may also not recover injured or stressed bacteria when used in conjunction with membrane filtration. The implications of these published media deficits have led to the intensified use and recommendation of more specific and exclusive indicator analyses based on enzyme-specificity or molecular techniques for generalized testing of environmental samples. Yet, prior work shows that data from these older, broadly inclusive, interrelated, and imperfect bacterial methods are critical for neural network source classification as ANN models would not train without their inclusion [8].

The number of input parameters must be balanced with neural network architecture and observation sample size to avoid overtraining. Although more exclusive fecal indicators such as total and male-specific coliphage, and the spores of *Clostridia perfringens* were measured and included in the resultant database, preliminary modeling efforts showed limited impact upon model training and prediction (data not shown). Therefore, although these more specific fecal indicators have been shown to be useful in predicting the presence of fecal contamination associated with viral and protozoan pathogens [15,17], they were not included as input parameters for the final model. The input parameters selected were identical to the one used in the previous source classification attempts [8]. The authors do not recommend precluding more specific analytical parameters, but they affirm the utility of widely used, and inexpensive methodology (total and fecal coliforms) for the approach presented herein. Preliminary modeling attempts would not train without the inclusion of the bacterial input parameters while exclusion of parameters that influenced fecal mobility such as turbidity and rainfall reduced predictive power [8].

For Step 1 model trained to sort sewage from all the combined form of runoff, the top three parameters that needed to be included for best prediction were total coliforms (TC), fecal coliforms (FC), and fecal streptococci (FStrep). Atypical coliforms (AC) and BG read from the mEndo coliform tests followed these in importance. These abnormally distributed, bacterial populations are intertwined and interdependent which has made linear-based modeling and regression of limited utility, while neural network modeling has been shown to be successful [8]. The TC membrane filtration test utilizing mEndo broth provides a rough estimate of the total number of recoverable, gram-negative bacteria in water, and differentiates this total number into those that do not ferment lactose (BG), those that ferment lactose slowly (AC), and those that ferment lactose very quickly (TC). Previous work [10] has shown the utility in ranking fecal sources based on the ratio of AC to TC,

with lower ratios being indicative of fresh fecal inputs. Higher ratios are thought to be due to non-coliforms such as *Aeromonas hydrophila*, slower growing bacteria common in nutrient-enriched waters, presenting an atypical colony appearance on mEndo media. Oxidase testing results of the AC isolated from these sites have supported this theory (data not shown). Fecal coliform testing identifies that subset of TC that can still grow in the presence of bile and at elevated temperatures, also known as thermotolerant coliforms. Combining this new ratio with the information obtained by the relative concentrations of FC and FStrep allows for estimates of the age and source of fecal contamination to be predicted. One parameter that the authors suggest including in future studies is water temperature as it directly impacts the rate of bacterial acclimation, growth, and die-off and is related to fecal mobility parameters such as rainfall.

3.2. Sorting raw sewage from animal-impacted runoff

The first step in Fig. 2 is to sort data observations into general source categories (raw domestic sewage versus animal-impacted runoff). While sewage generally had greater average concentrations of indicator bacteria than animal-impacted runoff, the variability around these averages is great (Table 2). On an average, sewage looks different from animal-impacted runoff, but the difference is slight and may not be reliable for prediction of the source of a single observation. Looking at individual observations in Table 1, and the standard deviation about the mean values reported in Table 2, it is conceivable that a single observation could be classified as sewage based upon bacterial concentrations alone. An example of this is an FA site observation on August 21, 1999 that had reported concentrations of BG—6.3, AC—7.2, TC—7.1, FC—TNTC, and FStrep 6.0 log C-FU/L. This observation could be easily classified as sewage based on average indicator bacteria values.

This led us to question the differences that exist in the data that are sufficient for individual observation source classification. One of the criticisms with ANN modeling is that it is a “black-box” approach and the underlying pattern may not be obvious. However, basic statistical analysis of the data yields an insight into inherent data patterns that an ANN can use. If true statistical differences underlie the established classifications, then an ANN has a strong signal upon which to train. Kruskal–Wallis one-way ANOVA on ranks was done for each classification, for each bacterial indicator or water quality parameter measured. Then pairwise comparisons between classifications were made using Dunn’s method to determine if underlying differences existed. The results are summarized in Table 3 where a significant difference ($P < 0.05$) in the concentration of an input parameter between two classification types is

Table 2
Summary of average parameter concentrations

	Sewage (log CFU/ L or NTU) (Std. dev.)	Flowing agricultural (log CFU/L or NTU) (Std. dev.)	Flowing suburban (log CFU/L or NTU) (Std. dev.)	Impounded suburban (log CFU/L or NTU) (Std. dev.)
Background colonies	5.9 (0.8)	4.7 (0.7)	4.5 (0.7)	4.6 (0.7)
Atypical coliforms	6.3 (1.2)	5.2 (0.7)	5.3 (0.6)	5.3 (0.7)
Total coliforms	6.0 (1.5)	4.6 (0.6)	4.3 (0.6)	3.9 (0.8)
Fecal coliforms	6.0 (1.5)	4.3 (0.9)	3.8 (0.8)	3.4 (0.9)
Fecal streptococci	5.6 (1.2)	4.1 (0.8)	3.8 (0.7)	3.2 (0.9)
Turbidity	62.6 (45.3)	23.8 (40.6)	5.7 (12.3)	16.4 (8.2)

Table 3
Summary of ANOVA comparisons between input parameters

	Sewage		
Flowing agricultural	y,y,n,n,y,y	Flowing agricultural	
Flowing suburban	y,y,y,y,y,y	n,n,y,n,n,y	Flowing suburban
Impounded suburban	y,y,y,y,y,y	n,n,y,y,y,n	n,n,y,y,y,y

Order of comparison is BG, AC, TC, FC, FStrep, and turbidity.

y = significant difference in concentrations, n = no significant difference ($P < 0.05$).

denoted by “y”. Sewage (RS) is clearly different from suburban runoff, either flowing (FS) or impounded (IS), as significant differences between all the bacterial concentrations and turbidity are found (y,y,y,y,y,y). RS is less different from agricultural runoff (FA) than FS or IS, but still has differences in BG, AC, FStrep, and turbidity concentrations. These underlying statistical differences between bacterial communities and water quality in sewage- and animal-impacted runoff should allow an ANN to readily classify data observations as either sewage- or animal-impacted, with greater difficulty for agriculturally impacted sites than suburban.

The objective of the ANN modeling effort was to see if RS had a distinctive microbial fingerprint from that of animal-impacted runoff. To train the ANN, individual data observations were classified as output result ‘0’ for sewage (RS) samples and ‘1’ for all other animal-impacted runoff samples (FA, FS, and IS) (Fig. 2). Output results between 0 and 0.5 were classified as RS and results from 0.5 to 1.0 were classified as animal impacted runoff. Using random numbers, the data observations were shuffled to disturb any time sequence, then 229 points were used for training, and 114 points were used for testing of the trained model. Results from ANN classification are presented in Table 4.

When all seven parameters (weather, BG, AC, TC, FC, FStrep, turbidity) are used as inputs to an ANN to sort sewage from all other forms of runoff, two mispredictions occurred, both in the classification of

animal-impacted runoff (FA, FS, IS) (Table 4). One misprediction was noted in training (FS—Armstrong) and one in testing (FA—Evan Mills). Both the mispredicted observations were associated with data gathered during rain events. These mispredictions might be attributed to the relatively few data points during rain events (76 total) in the training set (25, 33, and 18 for FA, FS, and IS, respectively) providing less opportunity for the neural network to recognize the patterns associated with rain events. While the FS misclassification in the training set was exactly opposite of the expected (0 rather than 1), the second misclassification in the test set had an error of only about 32.5%. Apart from these two misclassifications all other observation classifications were correctly predicted and are clustered closely around either 0 or 1 (total error 0.0791). Out of a total of 343 points used for training and testing, 301 points were predicted within $\pm 1\%$ of error. The distribution of training and testing errors is presented in Figs. 3 and 4. The ANN performed this classification very reliably in terms of both total misclassification (0.6%) and the sum total error (0.0791).

To further test the relevancy of each input parameter to the ANN model, and compare this to the ANOVA predictions, another seven ANNs were trained and tested such that each time, only six inputs were used to sort RS from all animal-impacted runoff. By eliminating one of the seven inputs in each model, the relative importance of each input was evaluated by comparing

Table 4
Results of ANN classification of sewage from animal-impacted runoff

Parameter	Total data		Train data		Test data		Mispredictions																	
	data	60	44	16	1	1	2	Mispredictions by ANN, all inputs		W/O weather		W/O BG		W/O AC		W/O TC		W/O FC		W/O FS		W/O turbidity		
								Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
Flowing agricultural																								
Flowing suburban		136	86	50	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Impounded suburban		83	59	24																				
Sewage		64	40	24																				
Total		343	229	114	1	1	2	3	2	3	2	2	4	4	3	5	4	4	0	4	4	0	2	2

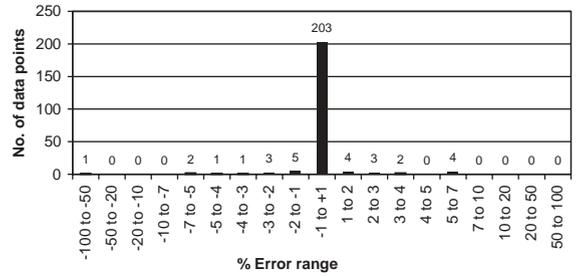


Fig. 3. Error frequency of training set predictions.

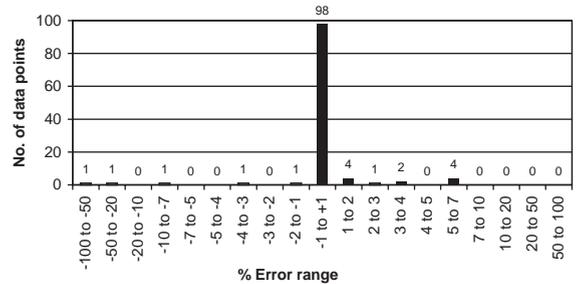


Fig. 4. Error frequency of testing set predictions.

the total misclassifications that result from the exclusion of that input parameter. The results of these incomplete ANN training and testing attempts are presented in Table 4. The removal of any single bacterial input parameter doubles the number of mispredictions, but the total misprediction error remains small (1–2%). Although turbidity is one of the parameters that ANOVA would predict to be important, removing it from the input set has little impact on the predictive power of the model and only changes the observations that would be mispredicted. Error distributions in training and test set continued to be small even when turbidity was excluded. Out of 229 training observations, only 2 data points showed more than $\pm 10\%$ error, while 80% of the predicted output values are within $\pm 1\%$ error of the expected one. In the test set, out of 114 observations, 2 points are beyond $\pm 50\%$ error and about 77% of the data points are within $\pm 1\%$ error. In essence, removing turbidity enables the model to perform better in training, whereas it produces two new errors in testing.

3.3. Sorting fresh from aged animal-impacted runoff

The second step presented in Fig. 2 is that of identifying aged (IS) from fresher animal-impacted runoff (FS and FA) after sewage observations are removed from the database. Looking at the differences between parameter concentrations for animal-impacted runoff classifications in Table 3, it would appear that

sorting IS from FS and FA is possible, as the majority of concentrations are significantly different. However, it is more complicated than merely comparing individual parameter concentrations. Sorting fecal age is dependent upon the absolute numbers, and the relative concentrations, of interdependent groups of gram-negative bacteria (BG, AC, TC, FC), and the number of gram-positive bacteria (FStrep). Although the FC:FStrep is no longer recommended for source classification due to differential die-off of bacteria isolated by KF streptococcus agar, and a documented rate of false positives [11], the relative numbers of FC to FStrep and other indicator bacteria are critical for the ANN model to differentiate fecal source and age. It is thought that the FC:FStrep is modified by the AC:TC in a manner that provides information about fecal age. Statistical analysis (data not shown) showed that the different classifications defined for this study had statistically different AC:TC ratios. The AC:TC ratios for IS sites were greater than for FS and FA, and the variance around the IS mean is twice as large. The ratios decreased for all animal-impacted sites when fresh feces were flushed into surface water by rain events, further supporting the findings of previous research that suggested the AC:TC to be fecal-age-related and more variable in impounded runoff than a free flowing one [10].

Table 5 presents the results of ANN modeling and prediction for sorting animal-impacted runoff by age (IS versus FS+FA). Of the total 279 observations, 12 training and 13 testing misprediction of classifications occurred. This is an overall misclassification rate of about 9% for individual observations. This is not an unacceptable rate of misprediction when considering

that this ANN sort is primarily based on fecal age indicated by indicator die-off and growth; factors controlled to a great degree by temperature, which was not measured for this study. Looking at the individual dates of the mispredictions, one is struck by the number of times (7 of 12 and 8 of 13 for training and testing respectively) that the data were collected during the six coldest months of the year (November–May). Field notes indicate snowfall within a week of several of the mispredictions in February. This suggests that the performance of the model could be improved by collecting and/or coding for additional input parameters such as season, solar insolation, and water temperature which directly impact bacterial growth and survival.

Table 6 is a summary of the results from an ANN model trained to distinguish aged suburban runoff from that of fresh (IS versus FS). This model shows the strength of the IS classification which is based upon <100 observations, and the confounding influence of rainy, cold days on the microbial fingerprints. Three of 4 mispredictions from 83 individual IS observations, and 7 of 11 mispredictions from 136 FS observations occurred on wet cold days. As is mentioned earlier, inclusion of water temperature and season may eliminate some of these errors, but these few errors should not draw undue attention from the ability of ANN to sort the average fecal age of runoff that is generated from the same land-use-related activities, from individual observations of a water sample, with <7% error overall, and <5% for classifying IS. We know of no other modeling approach that can match this performance utilizing inexpensive, inclusive, and interdependent input parameters.

Table 5
Results of ANN classification of aged versus fresh animal-impacted runoff

Parameter	Total data	Train data	Test data	Mismatches in model, all seven inputs	
				Train	Test
Flowing agricultural	60	44	16	2	5
Flowing suburban	136	86	50	3	2
Impounded suburban	83	59	24	7	6
Total	279	186	93	12	13

Table 6
Results of ANN classification of aged versus fresh suburban runoff

Parameter	Total data	Train data	Test data	Mispredictions, all seven inputs	
				Train	Test
Flowing suburban	136	86	50	4	7
Impounded suburban	83	59	24	0	4
Total	219	146	73	4	11

Table 7
Results of ANN classification of fresh suburban versus agricultural animal-impacted runoff

Parameter	Total data	Train data	Test data	Mispredictions, all seven inputs	
				Train	Test
Flowing suburban	136	86	50	7	5
Flowing agricultural	60	44	16	8	5
Total	296	134	66	15	10

It must be mentioned that other studies involving water quality and ANNs have concentrated primarily on predicting the quantities of chemicals in waters, not microbial fingerprints [1–7]. Even the work by Stanley et al. [16] chose to predict the color of the North Saskatchewan River and then tried to relate that measure to protozoan concentrations, rather than predict the microbial concentrations directly, and with good reason. The precision with which a parameter can be measured in water has a bearing on how accurately it can be predicted, as does the underlying distribution. Salinity, nutrients, color, and THM concentrations can all be measured with greater precision than microbes and are generally normally distributed within a sample. Chemical parameters can be measured easier, often on-line, and the resultant databases are large with many observations for training. Microbial populations are not so easily measured. They are often not normally distributed, even after data transformation as was the case for this study, and the inherent variability in the analytical techniques creates very large standard deviations around mean concentrations (Table 2). In addition to the inherent variability of microbial quantification the additional impacts that temperature and sunlight have on survival, multiple sources, and the interdependency of the microbial communities, the accomplishment of ANN to sort source (sewage versus runoff) and age (impounded runoff versus fresh) with >90% accuracy for any single day's observation on 7 simple parameters from 4 test methods is put into better perspective.

After running ANNs for Steps 1 and 2 of Fig. 2, the database has been winnowed down to observations that are the most difficult to separate from each other, flowing animal-impacted runoff. The first (source) sort identified most risky source of potential fecal contamination and human pathogens, human sewage. The second (age) sort identified the least risky, under dry weather conditions when holding times exceed a month, impounded runoff and supports findings reported previously [8], but with data from two different years of study. The last sort attempts to classify both land-use-associated source, and relative fecal age, of runoff carried actively in small creeks within the watershed.

3.4. Sorting the land-use-associated source of animal-impacted runoff

Looking at the ANOVA results presented in Table 3, the two classifications that have the least statistical difference on the input parameters are FS and FA. Only the measurements for TC and turbidity are significantly different. From this, it can be predicted that an ANN sort of these classifications would be difficult and prone to misclassification. As shown in Table 7, the model mispredicted the land-use-associated fecal source in 25 out of 296 observations (9%), evenly split between the two classifications. Eighteen of these 25 misclassifications were associated with colder weather as was noted for the previous runoff results. Yet, with only 2 significant differences in the input parameter between FS and FA, as compared to 5 for sorting based on age between IS and FS, the overall rate of misclassification is similar.

4. Summary and conclusion

This paper reports a successful attempt to distinguish between human sewage and animal-impacted runoff, fresh runoff from aged, and agricultural land-use-associated fresh runoff from that of suburban land-use-associated runoff using separate ANNs for data classification from a multi-year monitoring of a well-defined watershed. The ANNs were applied in a cascading, or hierarchical manner trained upon seven simple, easy to measure, water quality, or watershed parameters. The investigation of the ANN performance was measured in two ways: (1) training and (2) testing. An ANN was able to sort sewage from runoff with less than 1% error overall, and never misclassified sewage or aged animal runoff. Turbidity was found to be relatively unimportant for the classification of human sewage, while different, and sometimes gross, measurements of gram-negative and gram-positive bacteria were required. ANNs were able to produce accurate predictions on the fecal source, age, and land-associated animal fecal source from microbial counts with large standard deviations about the mean with great accuracy and

negligible errors. Predictions clustered tightly around the known values. From the results obtained by this study, ANN models for fecal source and age classification appear to be a useful tool for future water-borne pathogen risk identification, quantification, and development of early warning systems for surface water quality. There is room for improvement in parameter selection, model architecture design, and application, but the potential of ANN for sorting fecal source and runoff age on commonly measured water quality parameters has been demonstrated on a local scale.

References

- [1] Rogers LL, Dowla FU. Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling. *Water Resour Res* 1994;30(2):457–81.
- [2] Basheer IA, Najjar YM. Designing and analyzing fixed-bed adsorption systems with artificial neural networks. *J Environ Syst* 1995;23(3):291–312.
- [3] Maier HR, Dandy GC. The use of artificial neural networks for the prediction of water quality parameters. *Water Resour Res* 1996;32(4):1013–22.
- [4] Hutton PH, Sandhu N, Chung FI. Predicting THM formation with artificial neural networks. *Proceedings of the North American Water and Environment Conference*. New York: ASCE, 1996. p. 3556–7.
- [5] Sandhu N, Finch R. Emulation of DWRDSM using artificial neural networks and estimation of Sacramento River flow from salinity. *Proceedings of the North American Water and Environment Conference*. New York: ASCE, 1996. p. 4335–40.
- [6] Starrett SK, Najjar YM, Hill JC. Neural networks predict pesticide leaching. *Proceedings of the American Water and Environment Conference*. New York: ASCE, 1996. p. 1693–8.
- [7] Lek S, Guiresse M, Giraudel JL. Predicting stream nitrogen concentration from watershed features using neural networks. *Water Res* 1999;33(16):3469–78.
- [8] Brion GM, Lingireddy S. A neural network approach to identify non-point sources of microbial contamination. *Water Res* 1999;33(14):3099–106.
- [9] APHA, Standard methods for the examination of water and wastewater, 17th and 19th eds. Washington, DC: APHA, AWWA and WEF, 1989.
- [10] Brion GM, Mao HH. Use of total coliform test for watershed monitoring with respects to atypicals. *ASCE J Environ Eng* 2000;126(2):175–81.
- [11] APHA, Standard methods for the examination of water and wastewater, 17th and 19th eds. Washington, DC: APHA, AWWA and WEF, 1995.
- [12] Neelakantan TR, Lingireddy S, Brion GM. Effectiveness of different ANN training algorithms in predicting protozoa risks in surface waters. *J Environ Eng, ASCE*, 2002, in press.
- [13] Freeman JA, Skapura DM. *Neural networks: algorithms, applications and programming techniques*. New York, USA: Addison-Wesley, 1991.
- [14] Masters T. *Practical neural network recipes in C++*. USA: Academic Press, 1993.
- [15] Toranzas GA, McFeters GA. Detection of indicator microorganisms in environmental freshwaters and drinking waters. In: Hurst CJ, Knudsen GR, McInerney MJ, Stetzenbach LD, Walter MV (editors). *Manual of environmental microbiology*. Cleveland, OH: ASM, 1997. p. 184–94, ISBN 1-55581-087-X.
- [16] Stanley SJ, Zhang Q, Cudrak A. Early warning systems: a case study using artificial intelligence modeling techniques. *Proceedings of the Source Water Protection Symposium: A Focus on Waterborne Pathogens*. San Francisco, CA: AWWA/IWSA/USDA/USEPA, October 1998.
- [17] Brion GM, Neelakantan TR, Lingireddy S. Using neural networks to predict peak *Cryptosporidium* concentrations. *J Am Water Works Assoc* 2001;93(1):99–105.