APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Sept. 2005, p. 5244–5253 Vol. 71, No. 9
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.9.5244–5253.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Artificial Neural Network Prediction of Viruses in Shellfish

Gail Brion,[1]* Chandramouli Viswanathan,[2] T. R. Neelakantan,[3] Srinivasa Lingireddy,[1] Rosina Girones,[4] David Lees,[5] Annika Allard,[6] and Apostolos Vantarakis[7]

Department of Civil Engineering, University of Kentucky, Lexington, Kentucky 40506[1]; Indian Institute of Technology, Guwahati, Assam, India[2]; School of Civil Engineering, SASTRA Deemed University, Thanjavur 613402, India[3]; Biology School, University of Barcelona, Barcelona, Spain[4]; Centre for Environment, Fisheries and Aquaculture Science, Weymouth, United Kingdom[5]; Umeå University Hospital, Umeå, Sweden[6]; and School of Medicine, University of Patras, Patras, Greece[7]

Received 7 August 2004/Accepted 30 March 2005

A database was probed with artificial neural network (ANN) and multivariate logistic regression (MLR) models to investigate the efficacy of predicting PCR-identified human adenovirus (ADV), Norwalk-like virus (NLV), and enterovirus (EV) presence or absence in shellfish harvested from diverse countries in Europe (Spain, Sweden, Greece, and the United Kingdom). The relative importance of numerical and heuristic input variables to the ANN model for each country and for the combined data was analyzed with a newly defined relative strength effect, which illuminated the importance of bacteriophages as potential viral indicators. The results of this analysis showed that ANN models predicted all types of viral presence and absence in shellfish with better precision than MLR models for a multicountry database. For overall presence/absence classification accuracy, ANN modeling had a performance rate of 95.9%, 98.9%, and 95.7% versus 60.5%, 75.0%, and 64.6% for the MLR for ADV, NLV, and EV, respectively. The selectivity (prediction of viral negatives) was greater than the sensitivity (prediction of viral positives) for both models and with all virus types, with the ANN model performing with greater sensitivity than the MLR. ANN models were able to illuminate site-specific relationships between microbial indicators chosen as model inputs and human virus presence. A validation study on ADV demonstrated that the MLR and ANN models differed in sensitivity and selectivity, with the ANN model correctly identifying ADV presence with greater precision.

Health risks associated with the consumption of virally contaminated shellfish are well documented, as is the need for a more reliable viral indicator system by the industry (14, 17). Interdisciplinary studies are needed to define the underlying relationships between harvest area water quality, shellfish type, treatment processes, and viral presence, particularly with the advent of advanced detection and modeling methods. New understandings can be obtained with the application of new data-driven, fuzzy-logic-based models that can handle multiple, interrelated inputs and learn complex relationships. However, for the application of these new models, large, robust, multivariable, complete, and well-controlled datasets need to be created. A multicountry study in Europe has collected vital data in an effort to relate the viral contamination of shellfish with potential indicators. Analysis of these results has been reported earlier by Formiga-Cruz et al. (6, 7), and the database consisted of 468 individual observations from geographically diverse areas collected over 18 months. The resultant database was provided to a team of engineers and modeling experts for further probing with new artificial neural network (ANN) modeling tools under the hypothesis that these new modeling tools would be able to better define the relationships between viral presence/absence and potential water quality indicators than multivariate logistic regression (MLR) and provide more precise predictions with ANN models.

* Corresponding author. Mailing address: Dept. of Civil Engineering, University of Kentucky, 161 Raymond Bldg., Lexington, KY 40506-0281. Phone: (859) 257-4467. Fax: (859) 257-4404. E-mail: gbrion@engr.uky.edu.

Neural network models. Multilayered feed-forward networks have proven to be very powerful computational tools that excel in pattern recognition and function approximation. The general structure of a feed-forward neural network is shown in Fig. 1. Neurons, which are activation functions, are arranged in different horizontal layers, with multiple vertical layers possible. The nodes in the input layer receive the inputs of the model, and they flow through the hidden layer(s) internal to the network and produce outputs at nodes in the output layer. The working principle of feed-forward neural network is available elsewhere (15). Mathematically, a three-layer neural network with I input nodes, J hidden nodes in a hidden layer, and K output nodes, can be expressed as follows:

$$O_k = f_1\left[\sum_{j=1}^{J} w_{jk}^{ho} f_2\left(\sum_{i=1}^{I} w_{ij}^{ih} x_i + b_j^h\right) + b_k^o\right]$$

where O_k is the output from the kth node of the output layer, x_i is the input to the network at node i of the input layer, w_{ij}^{ih} is the weight between the ith node of the input layer and the jth node of the hidden layer, b_j^h is the bias term added to the jth hidden node, w_{jk}^{ho} is the weight between the jth node of the hidden layer and kth node of the output layer, and b_k^o is the bias term added to the kth output node. The architecture of the network shown in Fig. 1 can be summarized as 3:3:1 for three input nodes, three hidden nodes in a single hidden layer, and one output node. Each node has a function assigned to it, and the optimization of an ANN model often involves selecting the optimal combination of architecture (number of input and hidden nodes) and node functions (sigmoidal, hyperbolic, lin-

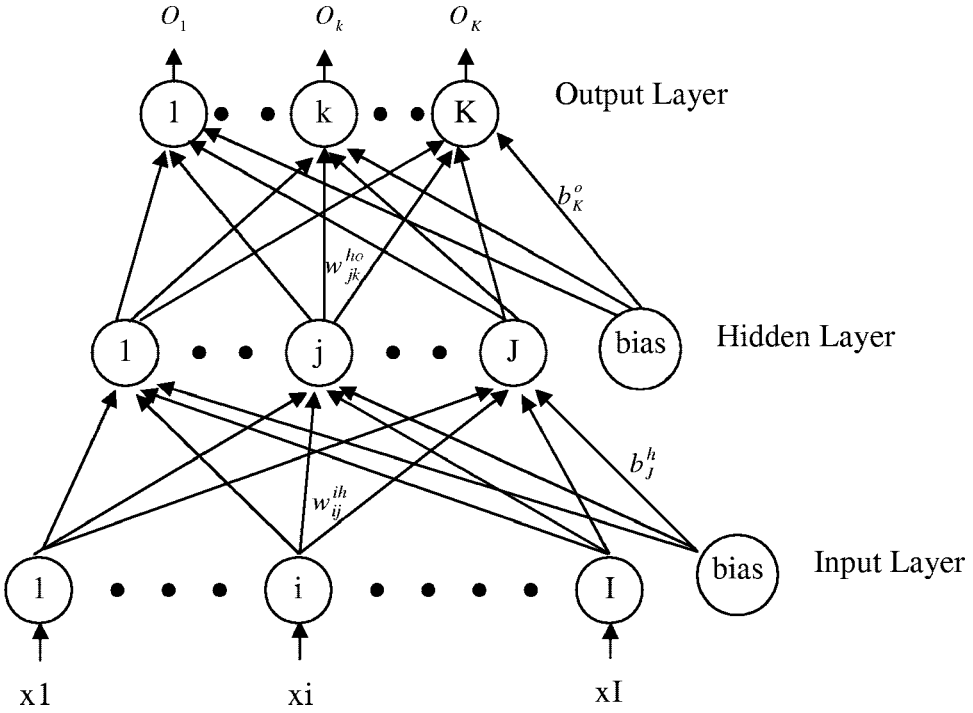

FIG. 1. Feed-forward neural network model.

ear, etc.) as well as selecting the correct input parameters. Neural networks, due to this complex interlocking structure, are excellent at applications where they are applied as universal function approximators for complex nonlinear relationships (9). Neural networks also have benefit in that they can learn the underlying functional relationships without being hampered by the distribution and independence issues common in environmental data.

Feed-forward neural networks are most commonly trained using a back-propagation algorithm. The back-propagation algorithm uses a gradient descent method for error minimization (18) on a randomized sort of the database of observations. In training, the weights on each connection are adjusted to yield the minimum error between the computed output pattern and the desired output pattern based on the method of least squares. The basic procedure used to train the network is embodied in the following steps: (i) apply input observations from training set to the network and calculate the corresponding output values, (ii) compare the computed output with the known output values and determine a measure of the error, (iii) determine corrections (increase or decrease) to the connection weights, (iv) apply the corrections to the weights, and (v) repeat items i through v with all the training vectors until the error for all vectors in the training set is reduced to an acceptable value.

RSE. In this paper, an approach using a first-order method based upon the relative strength effect (RSE) to evaluate the relative importance of input variables to the prediction of viral presence is used. This approach is based upon the work of Kim et al. (13), who proposed the RSE as a means to differentiate the relative influence of different input variables. They defined the RSE as the partial derivative of the output variable y_k, $\frac{\partial y_k}{\partial x_i}$. The RSE could be used to measure the relative importance of inputs in contributing to predict outputs. When $\frac{\partial y_k}{\partial x_i}$ is positive, an increase in input increases the output, and if it is negative, an increase in input causes a fall in output. Among the estimated RSE values of different inputs, the absolute maximum RSE value is used for normalizing the RSE values of all the inputs. Hence, for a considered data set, the RSE value would be either $+1$ or -1 for one value and for all other inputs, it will be in between $+1$ and -1. For basic screening, the average RSE value of an input is considered, i.e., if we considered p data sets, the RSE value for each input will be the average of RSE for that input in p data sets. The larger the absolute value of the RSE, the greater the contribution of that input variable is.

MATERIALS AND METHODS

Sampling and microbial assay. Detailed information on the sampling and assay methods is reported in Formiga-Cruz et al. (6, 7). Briefly, bivalve molluscan shellfish were collected monthly from 20 sites with different levels of fecal pollution over an 18-month period with analysis of the meat for human enteric virus presence (adenovirus [ADV], Norwalk-like viruses [NLV], and enterovirus [EV]) by nested PCR. Other microbial indicators in the shellfish tissue (*Escherichia coli*, somatic coliphages, F-specific RNA coliphages, *Bacteroides fragilis* phages) were quantified. Information on the country, area, season, mollusk type, temperature, and depuration status were added to the microbial results, and the entire resultant data set consisted of 468 individual observations.

Once collected, shellfish were shipped directly to each laboratory via cold storage within a 24-h period where *Escherichia coli* and bacteriophage groups were determined immediately. The sampling regime also included paired samples before and after depuration. To analyze for somatic coliphages and for bacteriophages infecting *Bacteroides fragilis*, shellfish flesh and liquor were collected into a sterile beaker with glycine buffer, pH 10 (1:5, wt/vol). For F-specific RNA bacteriophages (F-specific coliphages), peptone water (1:2, wt/vol) was added to the shellfish meat/liquor mixture. After the elution solutions were

added, the shellfish was homogenized with a blender and stirred for a 15-min contact time and then pH adjusted to 7.2 ± 0.2. The homogenate was then centrifuged at 2,170 × *g* for 15 min at 4°C. Phages contained in the supernatant were quantified by the double-agar-layer method with appropriate hosts (*E. coli* WG5-SP, *Salmonella enterica* serovar Typhimurium WG49, *B. fragilis* RYC2056-BP). Standardized protocol for all phage assays was used (10–12). *E. coli* was, with little modification, assayed by most probable number (MPN) as described by Donovan et al. (4), which consisted of a two-stage, five-tube, three-dilution MPN method. In brief, it required initial inoculation in mineral-modified glutamate broth and further confirmation by subculturing positive tubes onto a chromogenic agar to detect β-glucuronidase activity. A program for quality assurance and control for all phage types and *E. coli* was followed to ensure interlaboratory consistency. Human enteric viruses (ADV, NLV, EV) were detected by nested PCR after elution from tissue and liquor with glycine buffer 0.25 N at pH 10 (1:5, wt/vol) as described by Formiga-Cruz et al. (7) and then stored at −70 ± 10°C until the detection assay.

MLR modeling. Details of the prior regressions between indicators and viral presence/absence can be found in the work by Formiga-Cruz et al. (6). Microbial concentration values for *E. coli*, F-specific coliphages, somatic coliphages, and *B. fragilis* phages were transformed by the $\log_{10}(x + 1)$ function before fitting for the presence/absence of individual virus types by MLR run on the statistical software SPSS 10.0.7. The six input parameters used for the MLRs included *E. coli*, somatic coliphages, F-specific coliphages, *B. fragilis* phages, mollusk type, and country. An additional MLR run on Excel with all nine input parameters, but utilizing only a subset of the data to fit the model, was run for the purpose of verifying ADV presence.

Artificial neural network modeling. The same database used by the MLR done by Formiga-Cruz et al. (6) was used for ANN modeling efforts. Before applying the ANN model, the microbial input data were transformed using the $\log_{10}(x + 1)$ and then normalized by dividing the actual data value by 1.2 times the maximum value found in the input field. The microbial data for predicting NLV presence underwent a second transformation by the square root before normalization. The normalization was done to provide an equivalent numerical basis for judging the RSE of the numerical, microbial input parameters (*E. coli*, F-specific coliphages, somatic coliphages, and *B. fragilis* phages). In addition to these four microbially based numerical inputs, five heuristic knowledge inputs of mollusk type, area, month, and depuration status were used. In total, there were as many as nine and as few as six inputs for the models, each model using some combination of these inputs found to be optimal by initial ANN training attempts.

The heuristic input variable area was modified from the one reported by Formiga-Cruz et al. (6) that referred to areas classified as A or B relative to the ability to consume shellfish directly or after depuration. The variable area as used in this study reflected the relative level of fecal contamination at the sampling site on the day of observation rather than an average classification. The variable area was defined as one of four classifications corresponding to the value of the sum of the *E. coli* and somatic coliphages concentrations for the individual observation. If the *E. coli* and somatic coliphages concentration sum was <1,200, then the area coding was 1 for that observation. If the sum was between 1,201 and 12,000, then the area coding was 2. If the sum was between 12,001 and 60,000, then the area was coded as 3. If the sum was >60,000, then the area coding was 4. This classification scheme captured the relationship between somatic bacteriophage and potential host bacteria. The classification scheme created a way to relate diverse geographic sites based upon indicator-estimated fecal loadings within the shellfish. In addition, the input parameter date was split into 12 classifications, with January assigned a value of 1 and December a value of 12.

Separate ANN models were built to predict ADV, NLV, and EV presence/absence with Norwalk-like viruses of genogroups I and II lumped together into a single presence signal for NLV. Lumping the two groups of NLV together was done to provide more NLV-positive results for the purposes of training the ANN to avoid the phenomenon of memorization (overfitting) that can occur with limited observation and complex model structures. From prior experience, databases used for ANN classification modeling should contain more than 100 observations split evenly between outcomes to minimize memorization and emphasize generalization. There are several theory-based approaches outlined by Sarle (19) that provide guidelines for avoiding overfitting. One of the simplest is to maximize the number of data observations used, using between 30 and 5 times as many training cases as there are weights in the network, with fewer observations required as noise in the data decreases. The output of the ANN model was coded 0 for virus presence and 1 for virus absence, with 0.5 serving as the breakpoint between classifications per convention. The ANN model used for each individual type of human virus presence/absence prediction was a feed-forward ANN model with back propagation training developed using the software Neurosort VerII (16) created by some of the authors (3) and recently modified to contain a new calculation of the RSE. Modeling was done first for all countries' data combined to establish the optimum model architecture and node functions for each virus type, and then the established ANN model for each virus type was applied on a country-by-country data split basis to examine geographical differences in the relationships between indicators and the specified viral presence by RSE. The input variables for each of these individual models are presented in Table 1. The architecture and node functions for the combined data models are presented in Table 2. For direct comparison, RSE values were normalized by division with the sum of the absolute value of all input RSE values calculated for the trained ANN model.

TABLE 1. Input parameters used for viral presence modeling by ANN

Type and input parameter	All countries combined	Spain	Sweden	Greece	United Kingdom
Heuristic knowledge					
Country	X				
Month	X	X	X	X	X
Area	X	X	X	X	X
Mollusk	X	X			X
Depuration	X	X			X
Numerical knowledge					
E. coli	X	X	X	X	X
Somatic coliphages	X	X	X	X	X
F-specific coliphages	X	X	X	X	X
B. fragilis phages	X	X	X	X	X

For the comparison results presented in this paper, all of the data observations were used to train the ANN models. The MLR models constructed by the authors (for selected viral groups and input parameter selections not addressed by Formiga-Cruz et al. (6) also used all of the observations for performance testing. This was done so that direct comparisons of accuracy and generalization between the models could be made on equal datasets handled similarly. Accuracy was computed by dividing the number of correct predictions in a classification by the total number of observations available in that classification for viral presence (sensitivity), viral absence (selectivity), and combined accuracy [(number of viral presence correct + number of viral absence correct)/total number of observations)]. With the limited number of virus-positive samples in the groups identified, only the combined ADV data set was able to support ANN modeling and validation with split datasets, where the model was verified on data observations not used for training.

RESULTS

Modeling of the combined database. As can be seen in Table 3, The most frequently identified virus group in shellfish was adenovirus (39.1%). ADV was found at twice the frequency of NLV and EV groups. Yet the overall prevalence of isolating viruses from the shellfish was low, resulting in a smaller presence training, or fitting for MLR data set for the models rather than for the absence of viruses. The smallest number of positive samples was shown for NLV, with only 69 positive events from 468 observations for the ANN and MLR models to train and fit upon. As noted by Formiga-Cruz et al. (7), the greater

TABLE 2. Architecture and node functions for ANN modeling of virus presence on combined country dataset

Virus type	Optimal architecture	Node function	
		Hidden layer	Output layer
Adenovirus	9:20:1	Sigmoidal	Sigmoidal
Norwalk-like viruses	9:16:1	Hyperbolic	Sigmoidal
Enterovirus	9:20:1	Sigmoidal	Sigmoidal

TABLE 3. Frequency of virus presence and absence

Virus group	Total no. of samples (% of total)		No. of positive samples (% in country)			
	Positive	Negative	Spain	Sweden	Greece	United Kingdom
Adenovirus	182 (40)	284 (61)	37 (36)	46 (85)	20 (14)	79 (46)
Norwalk-like viruses	69 (15)	398 (85)	24 (23)	17 (32)	5 (4)	23 (13)
Enterovirus	86 (18)	380 (82)	26 (25)	13 (24)	22 (15)	25 (14)

prevalence of NLV was found in three countries, Spain, the United Kingdom, and Sweden, with less than 5% prevalence in Greece shellfish samples. This means that the models for NLV were basically training or fitting on only three countries' inputs. This geography-specific relationship for a particular virus type has bearing upon the relative importance of input parameters to the ANN and leads to questions about the appropriateness of combining all results into a single data set for modeling.

The descriptive power, or performance, of MLR and ANN modeling for the presence and absence of human viruses on the combined database is shown in Table 4. ANN clearly outperforms the MLR modeling, as demonstrated with more than 95% total accuracy for each virus group. The samples where virus was not detected were predicted with greater precision than for those where virus was present for both types of modeling efforts. Enterovirus presence was least well predicted by ANN (76.4%) but still was an improvement over the MLR results reported by (54%) Formiga-Cruz et al. (6). The ANN approach was able to learn the complex relationships between multiple input parameters and the desired output to a greater degree than the published MLR models.

Even with the least number of positive samples, NLV presence was best predicted by ANN of the three viral groups studied, with 100% accuracy on predicting shellfish samples that did not contain NLV genetic material. Previous ANN modeling for combined type I and II NLV presence (2) reported a correct NLV presence prediction of only 73.9%. This was improved in the current study to 92.8% by using a different codification for the input variable area and different architecture, numerical input transformation, and node functions. Yet there were a few errors in predicting NLV-positive observations for Greece, Sweden, and the United Kingdom, with 2 of 5, 2 of 17, and 1 of 23 mispredicted, respectively. Attempts to fit an MLR model to a combined NLVI and NLVII data set refused to converge; therefore, the separate MLR models published by Formiga-Cruz et al. (6) were cited for comparison.

TABLE 4. Relative prediction performance of MLR versus artificial neural network on combined all-country database

Virus group	Total percent accuracy (% positives correct versus % negatives correct)	
	MLR prediction	ANN prediction
Adenovirus	60.5 (46:70)	95.9 (91:99)
Norwalk-like viruses	75.1[a] (71[a]:76[a])	98.9 (93:100)
Enterovirus	64.6 (54:67)	95.7 (76:100)

[a] Estimated as an average of the two MLRs for NLVI and NLVII presented by Formiga-Cruz et al. (6).

For all the viral groups, the ANN model was learning a pattern between the numerical and heuristic inputs that was quite distinct and resulted in wide separation of predictions. Frequency graphs that show this separation for each virus type are shown in Fig. 2. The majority of the ANN data predictions are tightly clustered around the extreme ends of the range, close to either 0 or 1. Only for EV are there any output values in the middle range, close to 0.5. This type of linear cluster analysis on prediction values indicates that the underlying functions between virus presence/absence and the indicators selected is very specific.

Using a normalized relative strength effect (NRSE) to investigate the relative influence of the numerical and heuristic input parameters on ANN modeling, it can be seen (Table 5) that for predicting virus presence, bacteriophages are important numerical input parameters, with somatic coliphages ranked highest in all cases, with more than 25% of the strength of prediction based upon these values. In overall NRSE rank for predicting NLV and EV presence, the important microbial input parameters are the same with descending order of importance concentrations of somatic coliphages, F-specific coliphages, and *B. fragilis* phages. Except for ADV, values calculated for the bacteriophage groups measured in the shellfish were of more importance than concentrations of *E. coli* by more than twice. Only for predicting the ADV in shellfish are *E. coli* concentrations of equal numerical importance to somatic coliphages for the ANN model. It would seem that for predictions of different groups of human viruses, there are differences in the importance of the indicators selected for use as inputs to the models, but viable somatic coliphages are highly correlated with genetically detected human virus presence.

For the heuristic input parameters, to predict NLV, the time of year was as important as the country the data originated from. This is to be expected, as the distribution of NLV detection in shellfish was not evenly distributed across all countries and across all months. For ADV and EV, the type of mollusk was an important input, but depuration did not rise to a contribution larger than 10% of the prediction, as indicated by NRSE. The relative insignificance of depuration is in agreement with that cited by Formiga-Cruz et al. (6).

ANN modeling results produced utilizing the entire set of observations runs the risk of overfitting, or memorizing, and not generalizing the underlying pattern. Models should be evaluated on their ability not only to describe the relationship between inputs and outputs but to generalize that relationship to observations not known to the model during fitting or training. Therefore, it was important to verify ANN models with a separate modeling exercise where part of the data set was withheld from training for prediction verification. The only data set that had enough positive observations to run a classic training and verification exercise upon was that for ADV. Therefore, a separate ANN model with architecture ratio of 9:18:1, trained on all nine normalized input parameters, was developed on a randomly selected 400-observation subset of the total ADV database and then verified on the 68 verification observations withheld from training. An MLR was fit using the same input variables on the same datasets for direct compar-

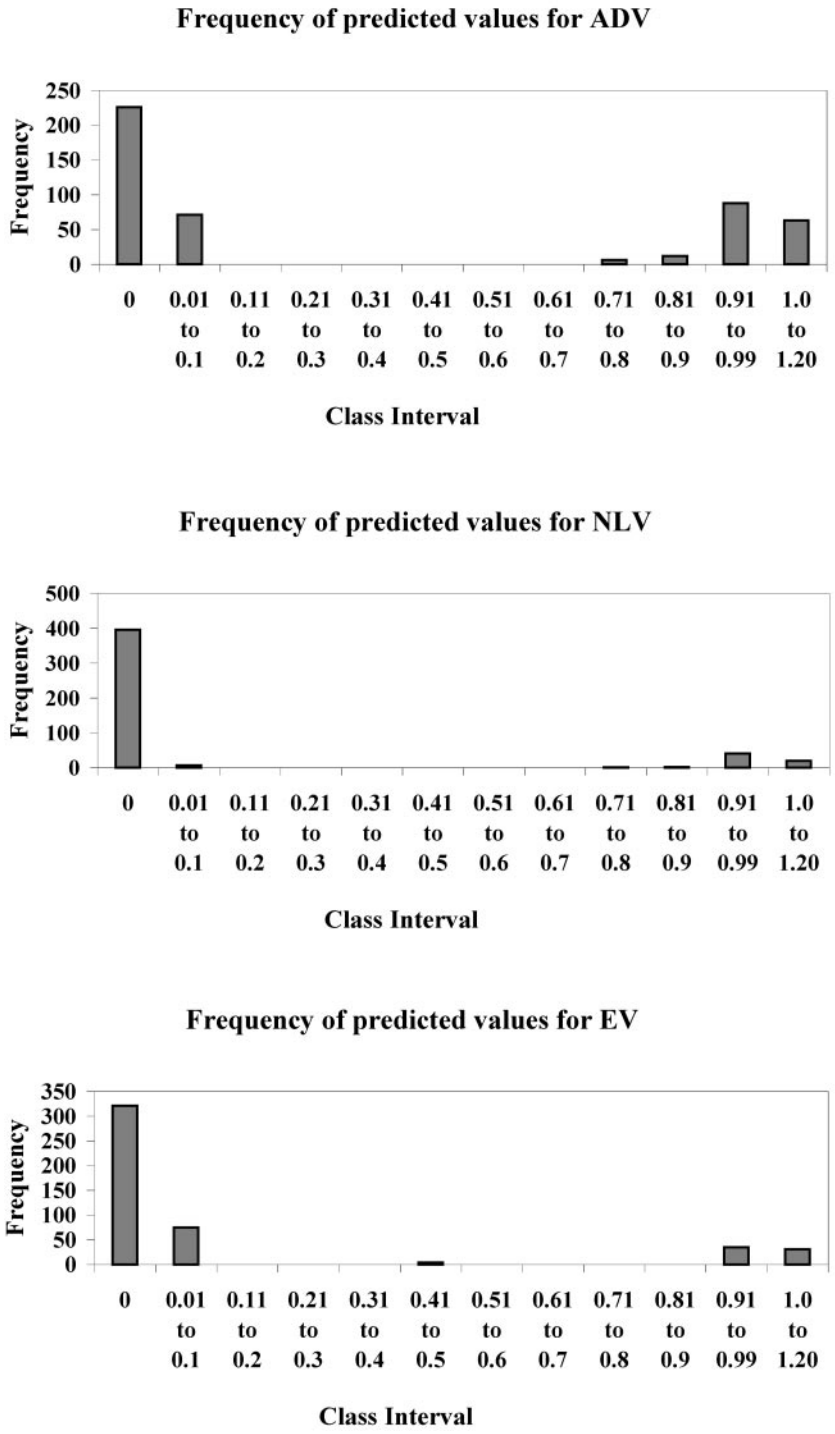

FIG. 2. Prediction frequency charts produced by ANN models for virus in shellfish.

ison. The results are presented in Table 6. Overall, the ANN model predicted 414 of 468 data observations correctly for a combined accuracy of 88.5% compared to MLR, whose combined accuracy was 66%. The MLR model verification results clearly indicate a model bias toward negative prediction and demonstrated poor sensitivity (21%). The ANN model had greater sensitivity on the validation set (41%) than the MLR, but accuracy was less than that desired for a useful commercial model. The ANN model predicted the absence of ADV genetic material with greater precision in the training data set slightly better than the MLR model.

Frequency graphs of prediction values for individual observations (Fig. 3) show that the majority of ANN predictions clustered toward the ends of the 0–1 prediction scale, with the

TABLE 5. Ranked importance of input parameters to ANN model based upon NRSE

Input parameter	Adenovirus (NRSE value)	Norwalk-like viruses (NRSE value)	Enterovirus (NRSE value)
Numerical			
Somatic coliphages	1 (0.29)	1 (0.28)	1 (0.25)
F-specific coliphages	3 (0.11)	2 (0.11)	2 (0.16)
B. fragilis phages	4 (0.01)	3 (0.08)	3 (0.14)
E. coli	2 (0.27)	4 (0.02)	4 (0.06)
Heuristic			
Area	2	4	1[a]
Month	3	2[a]	5
Mollusk	1[a]	3	2[a]
Country	4	1[a]	4
Depuration	5	5	3

[a] NRSE, >0.10.

majority of predictions less than 0.2 units from either extreme. This is in contrast to the frequency graphs for the MLR (Fig. 3) that show the majority of predictions for both testing and fitting to be clustered in the center of the range, with no predictions observed at the extremes of 0 or 1. It is clear by comparing the graphs that the ANN is providing more prediction separation for the presence or absence of ADV when trained on the same input variables than the MLR model.

Two-country comparison modeling. Applying a Kruskal-Wallis one-way analysis of variance on Ranks-log-transformed microbial data from each country provides information on the pattern of differences in the average microbial concentrations in shellfish between countries (Table 7). Spain and Greece have the least number of differences in average microbial indicator concentration, while the concentrations from the United Kingdom and Spain, and Sweden and Greece, were the most different from each other in pairwise analysis of variance comparisons ($P < 0.05$). Spain and Sweden had different average indicator concentrations in shellfish, with the exception of somatic coliphages. Spain and United Kingdom had significant differences in all of the average indicator concentrations. Sweden and United Kingdom had similar average concentrations for *E. coli* and F-specific coliphages but different concentrations for *B. fragilis* phages and somatic coliphages. Noting these differences led to questions about geographically induced differences in the underlying patterns between indicator concentration and viral presence/absence.

TABLE 6. Verification of ANN and MLR models for adenovirus prediction

Model used and dataset	Prediction results	
	Viral presence (no. predicted/total no.)	Viral absence (no. predicted/total no.)
ANN		
Training set	83% (127/153)	98% (243/247)
Verification set	41% (12/29)	82% (32/39)
MLR		
Fitting set	26% (39/153)	92% (228/247)
Verification set	21% (6/29)	95% (37/39)

In general, United Kingdom shellfish had higher coliform and bacteriophage concentrations than all other countries studied by Formiga-Cruz et al. (6, 7), with the exception of *B. fragilis* phages in Sweden. However, this observation was influenced by sampling site coverage. Sites were selected to cover a range of all the fecal contamination levels found in the diverse countries. The United Kingdom sampling plan included both moderately polluted EU class C (shellfish must have less than 46,000 *E. coli* cells per 100 g of mollusk flesh and intravalvular liquid and are subject to treatment before consumption) and prohibited shellfish harvesting sites in order to investigate virus occurrence in polluted sites restricted or prohibited for commercial exploitation. Spain ranked lowest in shellfish coliform and bacteriophage concentrations with the exception of somatic coliphages in Greece. This is to be expected, as the sampling sites selected for Greece and Spain were reported to have the highest incidence of sampling days where the potential *E. coli* host concentration was <230 MPN/100 g. It was questioned if statistical differences in the input indicator concentrations would result in differences in the underlying patterns between human virus and those indicators.

In order to investigate the country-specific differences that might be present, three separate ANNs were developed on the entire database to predict NLV presence in shellfish and the NRSE values for input parameters calculated individually for Spain, United Kingdom, and Sweden compared. Observations from Greece were not included in this analysis because of the paucity of positive observations. These ANN models all achieved more than 97% prediction accuracy, with only the model for Spain mispredicting any NLV-positive events (3 of 24). While these ANN models were developed on a suboptimal number of observations which can lead to overtraining, some trends can be noted. Sweden used less heuristic input parameters than the United Kingdom or Spain, comparing the overall rank and NRSE values for the input variables used by each individual ANN shows that the relative importance of inputs differs for each country (Table 8). Comparing Spain and the United Kingdom, the relative importance of the time of year is very clear. For Spain, time of year was the most influential input variable, but this input contributed least to the prediction of NLV in the United Kingdom. In the United Kingdom, Sweden, and Spain, somatic and F-specific coliphages were above an NRSE of 0.10, but only in Sweden were concentrations of *B. fragilis* phages of relatively equal value to the coliphages groups. Indeed, all three types of bacteriophage were equally important to predicting the presence of NLV in Sweden, while Spain and the United Kingdom model relied primarily upon the coliphage groups, somatic coliphages, and F-specific coliphages. In all three countries, *E. coli* is relied upon for less than 10% of the prediction. While Sweden shares with Spain strong reliance upon time of year, this input was not important to prediction of NLV presence in United Kingdom shellfish. The input variable area, which represented the normalized sum of somatic coliphages and their potential hosts, *E. coli* bacteria, helped further define the observations in each country. Looking at just the numerical indicator organism input NRSE, the rank order is the same for the United Kingdom and Spain, suggesting that these databases could be merged. Sweden has a very different pattern underlying the presence of NLV in shellfish, and this is borne out by the observation that

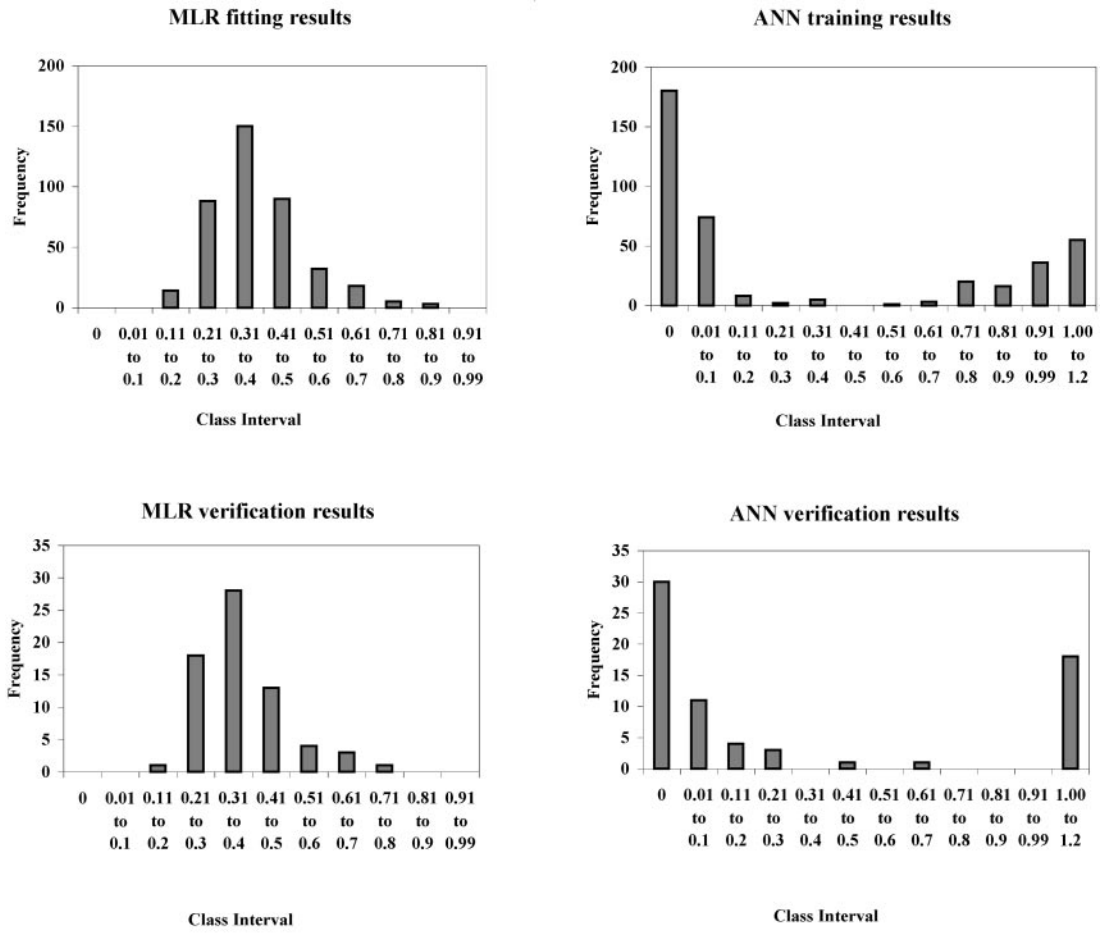

FIG. 3. Prediction frequency curves for ADV validation study comparing MLR to ANN models using nine input variables.

of the five positive observations mispredicted by the combined, all-country ANN model described prior, two of these were in Sweden. Spain and the United Kingdom modeled best in the combined ANN exercise, with only a single misprediction of NLV presence in the United Kingdom database. Greece was not expected to model well, as it did not have a sufficient number of NLV positive observations to train an ANN model upon.

DISCUSSION

Classification confidence. With regards to the levels of confidence in the dichotomous classification values for the data used in this study, all positive values were confirmed independently, primarily to eliminate the potential for false-positive results as a result of cross-contamination. So the classification of viral presence as indicated by genetic material could be ascribed great confidence. Negative values, since the numbers of viruses may be near the detection limit at times, do not have the same confidence level, and there are assumed to be false negatives, since not finding virus when there was virus in the shellfish was coded into the database. There is an assumption that an indefinable number of the negatives are in truth negative. So in the classification distribution, which is conceptually binomial, there could be some observations that with repeated testing would come to be represented by an average value falling between 0 and 1, with the shape of that peak determined by the underlying distribution of viruses in the sample and the number of replicate tests performed on a negative sample.

TABLE 7. Significant differences in microbial indicator concentrations by pairwise country comparison[a]

Country	United Kingdom	Sweden	Greece
United Kingdom			
Sweden	SP, BP		
Greece	EC, SP, FP	EC, SP, FP, BP	
Spain	EC, SP, FP, BP	EC, FP, BP	SP

[a] EC, *E. coli*, SP, somatic coliphages; FP, F-specific coliphages; BP, *B. fragilis* phages.

ANN versus MLR modeling. Predictions of viral presence in shellfish require models and indicator system that are capable of precision and accuracy in predicting viral presence for the protection of public health without undue burden on the shellfish industry. The large amount of uncertainty that exists around simple linear regressions obtained from single indicator systems cannot be tolerated for these types of public health policy decisions. The degree of uncertainty must be reduced,

TABLE 8. Country-specific differences in RSE values for ANN prediction of NLV[a]

RSE rank for input	Spain (NRSE)	United Kingdom (NRSE)	Sweden (NRSE)
1	Month (0.24)	F-specific coliphages (0.29)	Month (0.30)
2	Mollusk (0.15)	Somatic coliphages (0.19)	*B. fragilis* phages (0.17)
3	Somatic coliphages (0.15)	Area (0.12)	F-specific coliphages (0.17)
4	Depuration (0.14)	Mollusk (0.12)	Somatic coliphages (0.17)
5	Area (0.14)	Depuration (0.11)	*E. coli* (0.10)
6	F-specific coliphages (0.10)	*E. coli* (0.06)	Area (0.08)
7	*E. coli* (0.05)	*B. fragilis* phages (0.06)	
8	*B. fragilis* phages (0.02)	Month (0.05)	

[a] NRSE is the normalized input of the absolute RSE value.

and MLR modeling has been a large step forward in this goal. However, the underlying patterns between multiple, often interrelated indicators and pathogen risk are very complex and require a modeling system that can correctly capture this complexity without losing precision, precisely the attributes ANN models have been designed for. In this study, ANN modeling was clearly superior to MLR modeling, capturing the pattern between multiple indicators and genetic viral presence with greater precision and accuracy. The performance of the logistic function for modeling a nonlinear relationship is amplified by the additional dimensionality introduced by the additional structure of the ANN architecture. The fault does not lie with the logistic regression function as the majority of ANN models used for this study had in their hidden inner nodes, an MLR calculating a result that is passed on to the next node with a weighting factor for further processing. The ability of the ANN to accurately describe convoluted functional surfaces that exist between the input parameters and the output variable is due to a matrix of weighted MLR equations all feeding into a final MLR model. This interlocking complexity allows for the ANN model to learn multiple paths to the same answer and create different paths for shifts in input conditions that may occur without negatively impacting the accuracy of output classifications.

There are several things to consider when comparing conceptually dichotomous classification models. One is performance (correct prediction) that demonstrates how the model is capturing the description of the relationship between the inputs and the defined output. The other is generalization (validation), which measures the strength of the descriptive model to accurately predict a known outcome for an observation not seen during the fitting or training processes. Both of these must be evaluated with respects to the overall correct predictions and correct predictions within a class (sensitivity and selectivity) and with an understanding of the confidence one has in the correctness of the data classifications. Overall performance and total numbers of correct prediction may be misleading, as a correctly performing model should strike a balance between sensitivity (correctly identifying a positive response) and selectivity (correctly predicting a negative response). With unequal numbers of positive and negative observations, it is possible to have a high overall correct percentage but very poor performance for one of the classifications. In the data provided, ADV had a good split between the proportion of positive and negative samples (40:60) while NLV and EV were skewed with more negative findings than positive (85% and 82%, respectively). Therefore, for these virus types, it was important to evaluate not only the overall correct classification but the selectivity and sensitivity when evaluating performance.

The ANN models demonstrated superior performance in comparison to the MLR models repeatedly. In total, there were nine performance comparisons that could be made between the MLR and ANN models (Table 4) where the combined database and all data were used to fit or train the respective models for three different virus groups. When measuring the sensitivity (number of observations where virus presence was correctly predicted) and the selectivity (number of observations where virus absence was correctly predicted), the ANN model was superior to the MLR six of six times. When looking at the total number of correct predictions for each model per virus group, regardless of classification class, the ANN model was superior three of three times. ANN clearly outperformed the MLR modeling with more than 95% total accuracy for each virus group. The samples where virus was not detected were predicted with greater precision than for where virus was present for both modeling efforts. Enterovirus presence was least well predicted by ANN (76.4%) but still was an improvement over the MLR results reported by (54%) Formiga-Cruz et al. (6). The ANN approach was able to learn the complex relationships to a greater degree than MLR.

With regards to comparing generalization between the models, the validation study provided an opportunity to evaluate if the fit or trained model could accurately predict the classification of observations not known to the model. There are six ways to compare the MLR to the ANN model results for ADV classification presented in Table 6, and the ANN model was superior in terms of absolute accuracy for five of those six. First, the results of the fitting and training sets show that the ANN model was more accurate in prediction in terms of sensitivity, selectivity, and total correct predictions than the MLR model. Of note is the fact that the MLR model was unable to identify the most confident classification type, ADV presence, with much greater accuracy (83% versus 26%). This was not due to the phenomenon of classification skew that has been observed by us to occur in MLR models, as there were 40% of the total observations that were positive, and then confirmed as positive. The MLR model was unable to pick up this very strong signal in the fitting of the model, and this was repeated in the accuracy results for the validation set where only 26% of the confirmed positives are correctly identified. If one evaluated only the overall accuracy of the validation set, with the

ANN model providing 65% accuracy compared to the MLR model's 63%, this distinction would be lost.

The expanded dimensionality of the ANN model allows the use of inputs that might not appear significant to simpler models and provides a basis to recommend multiphage assay. The ANN modeling found value in input parameters that MLR neglected to find of significance. Of particular interest are the differences, and similarities, in the importance each modeling approach assigned to the three phage groups. The MLRs reported by Formiga-Cruz et al. (6) did not find somatic coliphages to be significant for ADV or NLV type I prediction, but ANN found the somatic coliphages to be the most significant input parameter for all virus groups with the largest numerical impact upon the output prediction. F-specific coliphages were found by both ANN and MLR models to be linked to human virus presence, but the *B. fragilis* phages were more important to the ANN model than to the MLR models. For predicting NLV, the *B. fragilis* phages were significant for the ANN, as significant numerically as concentrations of F-specific coliphages, but *B. fragilis* phages were not found to be significant for the NLVI and NLVII MLR models. Of the human viruses, only ADV was not significantly related to *B. fragilis* phages by ANN modeling, a finding that was in agreement with the prior MLR modeling results. The phages that infect *B. fragilis* have been promoted by other researchers (21) as reliable indicators for human wastes, and our results show a strong tie between their presence and NLV and EV presence. It appears that one cannot choose between these indicator phage groups when designing a shellfish study, as they appear to be related differently to the human viruses of concern.

Of as large an import as the differences between the relative significance of the input parameters are the similarities that were found between the modeling studies with regards to the significance of depuration to predicting viral presence. The insignificance of depuration as an input parameter is supported by the study by Formiga-Cruz et al. (6) that found that depuration as currently commercially practiced was shown not to appreciably reduce either the levels of F-RNA bacteriophages, phages of *B. fragilis*, and somatic coliphages or the occurrence of human pathogenic viruses in any of the countries shellfish. The insignificance of depuration to the modeling efforts is supported by the very low NRSE values and low ranking for depuration as an input parameter for the ANN modeling done in this study. Clearly, it made little difference to the ANN model if depuration was practiced, and this agrees with the lack of phage and viral clearing found by Formiga-Cruz et al. (6) and by other researchers (1, 5, 20). The relative unimportance of *E. coli* as an input parameter adds strength to the argument that reductions of *E. coli* cannot be relied upon to determine the duration or effectiveness of depuration.

The application of ANN modeling for pathogen prediction can provide a larger margin of safety around risk classifications and can allow researchers to see if a strong pattern underlies the data. The predictions produced by the ANN model separate the acceptance range for prediction values with greater distance than that found for MLR producing clusters of observations at the ends of the 0–1 range. This type of linear cluster analysis on the prediction frequencies is one way that a researcher can verify the existence of a strong pattern between the inputs and desired output classification. It is a visual tool that provides a means to check the strength of a model for dichotomous output classification by ANN or MLR modeling techniques.

It has been said that ANN models are relatively insensitive to the underlying distribution of the data, and often prediction efficiencies cannot be improved by additional data transformation. However, in this study the prediction values for previously reported ANN modeling of NLV (2) were improved by applying a second transformation by square root before normalization and by modifying the node activation function from the sigmoidal to the hyperbolic in the hidden layer (Table 2). The ANN model was improved to the point where prediction of all known observations was nearly perfect, and the tendency is to drive the fit toward perfection. However, care should be applied to prevent overtraining when applying ANN models and a balance must be struck between obtaining a perfect fit by memorizing specific paths to the established outputs and generalizing the underlying pattern between the inputs and the outputs with acceptable precision. Because of the ability of ANN models to memorize, it is imperative that models be fit to subsets of the data, and then their performance verified on data not seen during training, when adequate numbers of observations are available.

Because ANN models are data dependent, requiring more individual observations than normally required by simpler modeling and statistical methods, research projects that choose to apply this technique, or those that may provide a database for future mining should be designed appropriately and the potential impact of additional input parameters carefully considered. The creation of an ANN database for modeling can get expensive, especially if a number of different potential input parameters are being measured. However, the ability of ANN models to capture changes in complex environmental systems between a few strongly related inputs and the modeled output that may be significantly modified by parameters that other modeling techniques find insignificant, or worse that introduce lack of discrimination, has the potential to deepen our understanding of the relationships between pathogens and their indicators. Funding agencies need to be aware of the need to provide long-term support to build the potentially expensive databases that will be useful to applications of superior ANN modeling techniques.

Individual country ANN comparisons. The presence of NLV in mussels from Sweden appears to rely more heavily upon temporally associated input parameters than for shellfish from Spain and the United Kingdom, with the *B. fragilis* phages serving as a significant input for NLV presence prediction by ANN models. The reliance upon time of year is in agreement with Hernroth et al. (8) who noted the effect of spring thawing and runoff on the prevalence of human viruses from all Swedish harvesting areas tested. Comparing the MLR for Sweden done by Hernroth et al. (8) to that of Formiga-Cruz et al. (6) on the combined country data set, the relative importance of *B. fragilis* phages is reduced, with only F-specific coliphage showing significance in the combined regression results. The flood and thaw event that happened in Sweden during the time of study had a unique influence on the underlying pattern between indicators and this human virus that does not appear when the data from multiple countries are combined.

Differences between the underlying patterns between indi-

cators and pathogens between different countries support the idea that there is no ideal model that can be exported blindly from one area to another, but that a localized approach be proposed and verified. There are many factors that may individualize the underlying patterns between the pathogen response to be modeled and the input variables. Some countries may wish to include variables that are of local import. In the original database from the study by Formiga-Cruz et al. (6), there was information gathered on other potential input parameters (pH, water temperature, salinity, oxygen content, and turbidity), which while not found to be useful for prediction of viral presence in their study, may be significant locally. The impact of the aforementioned flood event in Sweden is a good example of the potential of some of these parameters to impact prediction models and river flow, or changes in river flow, could have been a valuable input parameter for the harvest beds under study. The inclusion of the input variable month in the ANN that resulted in improved prediction of NLV presence supports this idea. Individual countries should develop monitoring based ANN models that utilize the indicators most linked to the pathogens in their environments, and that requires funding of intense localized study as well as large-scale collaborations so discoveries can be made, and compared, on both scales. Since larger databases, obtained by combining data from multiple areas, allow researchers and policy analysts to expand their understanding of general indicator pathogen relationships, ANN modeling could be applied as a new way to evaluate if databases from geographically separate areas should be combined, rather than relying upon statistical methods that are very sensitive to the underlying data distribution.

Conclusions. ANN modeling can provide insight into the relationships between viral pathogens and their indicators. Analysis of different groups of bacteriophage and the bacteria they infect may yet provide the basis for viral shellfish quality control, especially when used in a combined indicator system that is attuned to unique geographic and temporal characteristics through the application of ANN and other advanced modeling techniques. The ideal set of indicators and input parameters for modeling has yet to be defined, and is likely subject to some geographical differences, but this study shows that ANN models can provide improved description and more accurate prediction of viral presence than MLR models on the same set of input parameters where the number of data observations is adequate for their training. In the same way that the number of samples are built into the sampling scheme for research utilizing traditional statistical methods, studies planning on applying ANN models must assure that enough observations are obtained to support training and validation studies so that model performance is evaluated on generalization as well as overall accuracy, sensitivity, and selectivity. The findings of this study, and our experience with other studies utilizing microbial databases, suggest that the utility of ANNs be more widely explored, in concert with traditional statistical methods, to obtain the most benefit from environmental studies.

ACKNOWLEDGMENTS

This research was supported by Commission of the European Communities, Agriculture and Fisheries (FAIR) project CT98-4039, and United States Environmental Protection Agency (STAR) project R830376.

REFERENCES

1. **Abad, F. X., R. M. Pinto, R. Gajardo, and A. Bosch.** 1997. Viruses in mussels: public health implications and depuration. J. Food Prot. **60:**677–681.
2. **Brion, G. M., S. Lingireddy, T. R. Neelakantan, M. Wang, R. Girones, D. Lees, A. Allard, and A. Vantarakis.** 2004. Probing Norwalk-like virus presence in shellfish with artificial neural networks. Water Sci. Technol. **50**(1)**:** 125–129.
3. **Brion, G. M., T. R. Neelakantan, and S. Lingireddy.** 2002. A neural network based classification scheme for sorting sources and ages of fecal contamination in water. Water Res. **36:**3765–3774.
4. **Donovan, T. J., S. Gallacher, N. J. Andrews, M. H. Greenwod, J. Graham, J. E. Russell, D. Roberts, and R. Lee.** 1998. Modification of the standard method used in the United Kingdom for counting *Escherichia coli* in live bivalve molluscs. Commun. Dis. Public Health **1:**188–196.
5. **Doré, W. J., and D. Lees.** 1995. Behavior of *Escherichia coli* and male-specific bacteriophage in environmentally contaminated bivalve molluscs before and after depuration. Appl. Environ. Microbiol. **61:**2830–2834.
6. **Formiga-Cruz, M., A. K. Allard, A. C. Conden-Hansson, K. Henshilwood, B. E. Hernroth, J. Jofre, D. N. Lees, F. Lucena, M. Papapetropoulou, R. E. Rangdale, A. Tsibouxi, A. Vantarakis, and R. Girones.** 2003. Evaluation of potential indicators of viral contamination in shellfish and their applicability to diverse geographical areas. Appl. Environ. Microbiol. **69:**1556–1563.
7. **Formiga-Cruz, M., G. Tofino-Quesada, S. Bofill-Mas, D. N. Lees, K. Henshilwood, A. K. Allard, A. C. Conden-Hansson, B. E. Hernroth, A. Vantarakis, A. Tsibouxi, M. Papapetropoulou, M. D. Furones, and R. Girones.** 2002. Distribution of human viral contamination in shellfish from different growing areas in Greece, Spain, Sweden, and the United Kingdom. Appl. Environ. Microbiol. **68:**5990–5998.
8. **Hernroth, B. E., A.-C. Conden-Hansson, A.-S. Rehnstam-Holm, R. Girones, and A. K. Allard.** 2002. Environmental factors influencing human viral pathogens and their potential indicator organisms in the blue mussel, *Mytilus edulis*: the first Scandiavian report. Appl. Environ. Microbiol. **68:**4523–4533.
9. **Hopfield, J. J.** 1982. Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA **79:**2554–2558.
10. **International Organization for Standardization.** 1996. Water quality. Detection and enumeration of bacteriophages, part 1: enumeration of F-specific RNA bacteriophages. ISO 10705-1. International Organization for Standardization, Geneva, Switzerland.
11. **International Organization for Standardization.** 1999. Water quality. Detection and enumeration of bacteriophages, part 2: enumeration of somatic coliphages. ISO/FDIS 10705-2. International Organization for Standardization, Geneva, Switzerland.
12. **International Organization for Standardization.** 1999. Water quality. Detection and enumeration of bacteriophages, part 4: enumeration of bacteriophages infecting Bacteroides fragilis. ISO/DIS 10705-4. International Organization for Standardization, Geneva, Switzerland.
13. **Kim, C. Y., G. J. Bae, S. W. Hong, C. H. Park, H. K. Moon, and H. S. Shin.** 2001. Neural network based prediction of ground surface settlements due to tunneling. Comput. Geotechnics **28:**517–547.
14. **Lee, R. J., and O. C. Morgan.** 2003. Environmental factors influencing the microbiological contamination of commercially harvested shellfish. Water Sci. Technol. **47(3):**65–70.
15. **Masters, T.** 1993. Practical neural network recipes in C++. Academic Press, Boston, Mass.
16. **Neurosort Ver II.** 2004. Department of Civil Engineering, University of Kentucky, Lexington.
17. **Rippey, S. R.** 1994. Infectious diseases associated with molluscan shellfish consumption. Clin. Microbiol. Rev. **7:**419–425.
18. **Rumelhart, D. E., G. E. Hinton, and R. J. Williams.** 1986. Learning internal representation by error propagation, p. 318–362. *In* D. E. Rumelhart and J. L. McClelland (ed.), Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. MIT Press, Cambridge, Mass.
19. **Sarle, W. S. (ed.).** 1997. Neural Network FAQ, part 3 of 7: generalization. Periodic posting to the Usenet newsgroup comp.ai.neural-nets, ftp://ftp.sas.com/pub/neural/FAQ3.html#questions.
20. **Schwab K. J., F. H. Neill, M. K. Estes, T. G. Metcalf, and R. L. Atmar.** 1998. Distribution of Norwalk virus within shellfish following bioaccumulation and subsequent depuration by detection using RT-PCR. J. Food Prot. **61:**1674–1680.
21. **Scott, T. M., J. B. Rose, T. M. Jenkins, S. R. Farrah, and J. Lukasik.** 2002. Microbial source tracking: current methodology and future directions. Appl. Environ. Microbiol. **68:**5796–5803.