JOURNAL OF
ENVIRONMENTAL HYDROLOGY

The Electronic Journal of the International Association for Environmental Hydrology
On the World Wide Web at http://www.hydroweb.com

VOLUME 5 1997

USE OF THE SPATIAL ANALYSIS NEURAL NETWORK (SANN) METHOD FOR REGIONAL GROUNDWATER CONTAMINATION DECISION-MAKING

Hyun-Suk Shin
Hydrologic Science and Engineering Program
Colorado State University
Fort Collins, Colorado, USA

Spatial Analysis Neural Network (SANN) is a specified neural network for conducting the spatial analysis of any type of variable. It provides a nonparametric mean estimator and also estimators of higher order statistics such as standard deviation and skewness. In addition, it provides a decision-making tool, including an estimator of posterior probability that a spatial variable at a given point will belong to various classes representing the severity of the problem of interest, and a Bayesian classifier to define the boundaries of subregions belonging to the classes. In this paper, the use of SANN as a decision-making tool to investigate an area contaminated by viruses in a groundwater system is illustrated. SANN provides two pieces of information; the contamination probability that the virus decay rate at a given point is less than a predefined threshold value, and the classification map defining contaminated and non-contaminated regions. The method is applied to several cases with varying threshold levels of the observed virus decay rate values, and the results show graphically the extent of the contaminated region and the change of the contamination probabilities.

INTRODUCTION

Knowing the uncertainty of hydrological and environmental data such as precipitation, soil properties, and groundwater contaminant concentration is of a great necessity and importance for solving various problems related to water resources planning and management, groundwater contamination, and water quality control. A number of methods such as kriging have been suggested in the literature for hydrological and environmental data (Bras and Rodrigues-Iturbe, 1985). However, these methods are limited for analysis of complex natural phenomena, because of the assumptions of stationarity and normality of the underlying variables, and the drawbacks in structural analysis such as the shadow effect, anisotropic data, nested structure, and the hole effect (ASCE, 1990). Shin and Salas (1997) introduced an alternative method, called Spatial Analysis Neural Network (SANN) which has the following characteristics: (1) non parametric estimators of the conditional mean and higher order moments such as standard deviation and skewness coefficient; (2) the estimator of the point posterior probability estimator for some classes is predefined and the Bayesian classifier assigns a class to an arbitrary spatial point. Recently, neural networks have been successfully used to solve some complex hydrological and environmental problems such as river flow prediction (Markus et al., 1995), activated sludge prediction (Novotny et al., 1991), determination of aquifer parameters (Rashid et al., 1992), and groundwater reclamation (Rogers and Dowla, 1994). The proposed estimators are implemented into a specified multi-layer feed-forward neural network structure to achieve computational efficiency based on parallel system modeling, and its structure and operation scheme is summarized briefly in this paper.

Viruses in groundwater are of interest to the public because many waterborne diseases are caused by the contamination of drinking water from groundwater viruses (Craun and Knox, 1985). In estimating the impact of viruses on groundwater contamination, it has been common practice to use virus decay rate (or virus inactivation rate). This is the slope of the linear regression curve which is constructed with time (day) and the number of the infective virus particles remaining after a corresponding time transformed to a logarithmic base. Then, the unit is described as $-\log_{10}$ (virus particles) / day (Yates and Yates, 1989). A small rate means that the viruses can remain in the groundwater for a long period. In groundwater contamination problems caused by viruses, the area which has the smaller virus decay rate is considered as the more seriously contaminated area. In these cases, one may wish to determine the probability that the virus decay rate will be less than a given truncation level. The truncation level might be a standard level associated with the virus decay rate. In addition, one may want to know how the contaminated area changes over the region according to varying truncation levels. For solving those problems related to decision-making of future groundwater remedial action, we demonstrate the use of the posterior probability estimator and Bayesian classifier provided by SANN as a decision-making tool for various cases with varying truncation levels.

SANN MODEL DESCRIPTIONS

In this section, we describe the structure and operation of SANN, which has been developed for analyzing any type of spatial variables, based on a multi-layer feed-forward neural network form. The more theoretical derivations and illustrations related to various estimators can be found in Shin (1997) and Shin and Salas (1997). Suppose that a spatial variable z of interest exists, for which measurements are available in a two-dimensional domain, i.e. $\mathbf{x} = [x, y]$. We want to obtain some spatial information at an unknown or unmeasured point **x,** such as estimations of conditional mean $\hat{z}$ (**x**), its standard deviation $\hat{s}$ (**x**), posterior probability $P[C^j \mid \mathbf{x}]$ of each class, and class indicator d(**x**). We have

N sample observations in the region which are denoted by the observation set $\{\mathbf{X}_n, Z_n \mid n = 1, \ldots, N\}$. For conducting the estimation of posterior probabilities and classification, suppose that a spatial variable $z(\mathbf{x})$ is classified into N_c classes, $C^1, C^2, \ldots, C^{Nc}$ where C^j denotes the j-th arbitrary class. The classes are defined by the truncation levels $TL(j)$, $j = 0, 1, \ldots, N_c$ in which $TL(0) = -\infty$ and $TL(N_c) = \infty$. Based on the definition of the classes, the observation set $\{\mathbf{X}_n, Z_n \mid n=1,\ldots, N\}$ can be classified as $\{\mathbf{X}_{(k,j)}, Z_{(k,j)} \mid k=1,\ldots,N^j$ and $j=1,\ldots,N_c\}$ where k denotes the observed point in each class C^j, and N^j is the number of the observed points belonging to class C^j.

For this purpose, SANN is structured as shown in Figure 1. It consists of four layers, in which the neurons or nodes between layers are interconnected successively by feed-forward direction as shown in the Figure. The four layers are called: *input layer*, *GKF layer*, *summation layer*, and *estimator layer.*

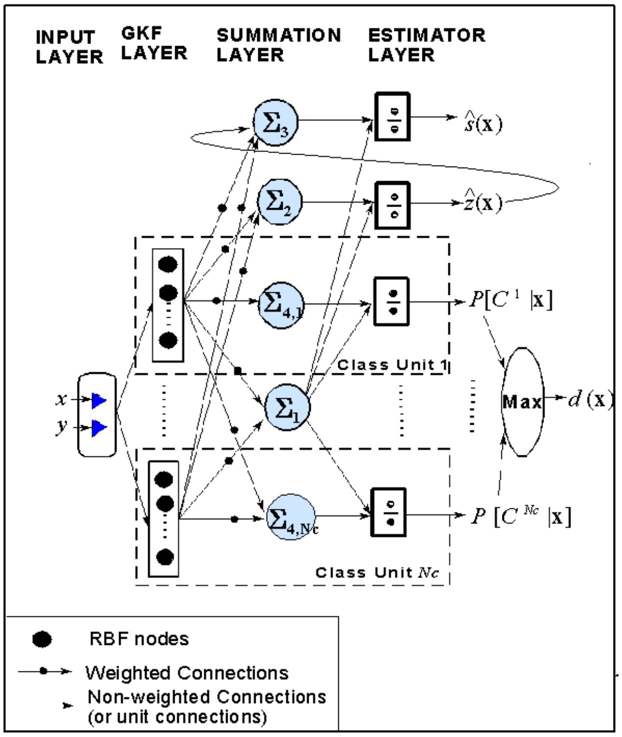


Figure 1. The structure of SANN.

In the following paragraphs, the function and the connection mechanism of each layer will be explained in detail. Considering a two-dimensional domain, the *input layer* has two nodes which represent the x and y coordinates, i.e. the vector $\mathbf{x} = [x, y]$. The connections of the input layer implement a pass of the input coordinate vector $\mathbf{x} = [x, y]$ to the *GKF layer*, and those are not weighted. The GKF layer consists of N Gaussian Kernel Function (GKF) nodes. To determine the posterior probability estimator and the Bayesian classifier, the GKF nodes must be divided up into N_c class units as shown in Figure 1. For doing this, the observed set $\{\mathbf{X}_n, Z_n \mid n=1,\ldots,N\}$ is rearranged as $\{\mathbf{X}_{(k,j)}, Z_{(k,j)} \mid k=1,\ldots,N^j$ and $j=1,\ldots,N_c\}$. $\mathbf{X}_{(k,j)}$ is then located at the center of the k-th GKF node in class unit j in which the number of the GKF nodes is N^j. Then, the transfer or activation functions of the k-th GKF node in class unit j are expressed as:

$$\alpha_{(k,j)} = \exp\left[-\frac{D^2_{x(k,j)}}{2\sigma^2_{x(k,j)}}\right] \tag{1}$$

where $=\alpha_{(k,j)}$ the GKF node output from the k^{th} node in class unit j; $D_{x(k,j)}$ = the Euclidean distance between the input vector **x** and the k^{th} *center* $\mathbf{X}_{(k,j)}$ in class unit j and the square of it is expressed as $D^2_{x(k,j)} = (\mathbf{x}-\mathbf{X}_{(k,j)})^{\mathrm{T}}(\mathbf{x}-\mathbf{X}_{(k,j)})$; and $\sigma_{x(k,j)}$ = the *width* for the k^{th} GKF node in class unit j. Each GKF node has the internal parameters; $\mathbf{X}_{(k,j)}$ = the position of the *center* of the GKF node in two-dimensional space, and $\sigma_{x(k,j)}$ = the smoothing parameter known as the *width* of the GKF nodes. The function of the GKF node may be summarized as: the output from each GKF node is a function of the Euclidean distance from the center $\mathbf{X}_{(k,j)}$ to the input point **x,** and each GKF node only responds (or activates) when the input pattern falls within its *receptive field* which is defined by the width of the GKF node $\sigma_{x(k,j)}$ (Poggio and Girosi, 1990). When the input vector **x** is placed at the center of the GKF node $\mathbf{X}_{(k,j)}$, the output (1) becomes the maximum value which is one. Otherwise, the magnitude of the GKF output decreases exponentially, as the input vector is farther from the center. The outputs of the GKF nodes in the GKF layer are passed to the *summation layer* with weighted connections as shown in Figure 2.

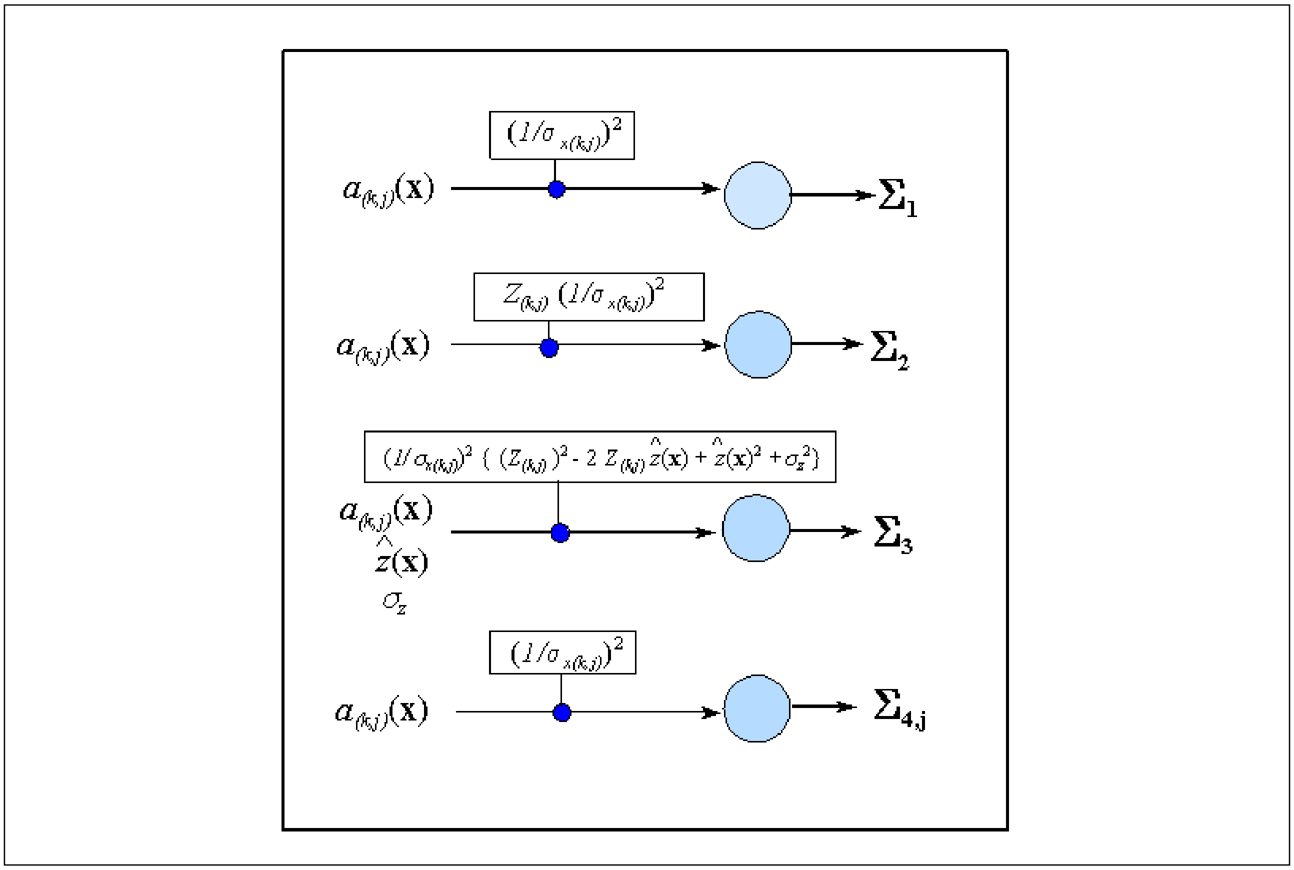


Figure 2. Weighted connections between the GKF nodes and the summation nodes.

Then, the summation layer provides the following outputs:

$$\sum{}_1 = \sum_{j=1}^{N_c}\sum_{k=1}^{N^j}\left(\frac{1}{\sigma^2_{x(k,j)}}\right)\alpha_{(k,j)} \tag{2.a}$$

$$\sum{}_2 = \sum_{j=1}^{N_c}\sum_{k=1}^{N^j}\left(\frac{1}{\sigma^2_{x(k,j)}}\right)\mathrm{Z}_{(k,j)}\alpha_{(k,j)} \tag{2.b}$$

$$\sum_{3} = \sum_{j=1}^{N_c}\sum_{k=1}^{N^j}\left(\frac{1}{\sigma^2_{x(k,j)}}\right)\left\{\left(\mathrm{Z}_{(k,j)}\right)^2 - 2\,\mathrm{Z}_{(k,j)}\hat{z}(\mathbf{x}) + \hat{z}(\mathbf{x})^2 + \sigma_z^2\right\}\alpha_{(k,j)} \tag{2.c}$$

$$\sum_{4,j} = \sum_{k=1}^{N^j}\left(\frac{1}{\sigma^2_{x(k,j)}}\right)\alpha_{(k,j)} \tag{2.d}$$

where $\mathrm{Z}_{(k,j)}$ = observed value corresponding to the *k-th* GKF node for class unit *j*, $\hat{z}$ (**x**) = estimated value at the point **x**, σ_z = the smoothing parameter or Gaussian kernel width associated with the spatial variable *z,* which must be defined. The outputs from the summation nodes are passed to the estimator nodes with unit weights. Then, as shown in Figure 1, the outputs of estimator nodes assign the estimations of the conditional mean $\hat{z}$ (**x**), its standard deviation $\hat{s}$ (**x**), and the posterior probability $P[C^j \,|\, \mathbf{x}\,]$ of each class, respectively, as:

$$\hat{z}(\mathbf{x}) = \frac{\sum_2}{\sum_1} \tag{3}$$

$$\hat{s}(\mathbf{x}) = \sqrt{\frac{\sum_3}{\sum_1}} \tag{4}$$

$$\hat{P}\left[C^j \middle| \mathbf{x}\right] = \frac{\sum_{4j}}{\sum_1} \tag{5}$$

Finally, the class indicator *d*(**x**) is determined by assigning the class with maximum posterior probability.

SANN consists of three operation modes, namely, a training mode, an interpolation mode, and a classification mode. In the training mode, the model structure is constructed according to the classes defined by the user as described above. In addition, the model parameters such as the centers and the widths for all GKF nodes must be determined by using sample observations. The training procedure can be summarized as:

(a) Prepare the *observation set* { $\mathbf{X}^n$, Z^n | n= 1,…, N } where N is the number of observations.

(b) Define the classes C^j = { C^1, C^2, …, C^{Nc}} with *truncation levels* { $TL(j)$ |j=1,…,N_c}. Based on the definition of the classes, classify the observation set into each class C^j with {$\mathbf{X}_{(k,j)}$, $\mathrm{Z}_{(k,j)}$ | k= 1, …, N^j ; j=1,…, N_c }.

(c) Set the centers of the GKF nodes with the observed coordinate vector $\mathbf{X}_{(k,j)}$. For instance, the center of the k th GKF node in class unit j is assigned to be $\mathbf{X}_{(k,j)}$. Here, the class layer is arranged with N_c class units as shown in Figure 3.

(d) Determine the widths $\sigma_{x(k,j)}$ of the GKF nodes. The widths represent the shape of the Gaussian kernel as well as the diameter of the receptive region. They have a profound effect upon the accuracy of the estimation (Haykin, 1994). To cover the whole input space as uniformly as possible, centers

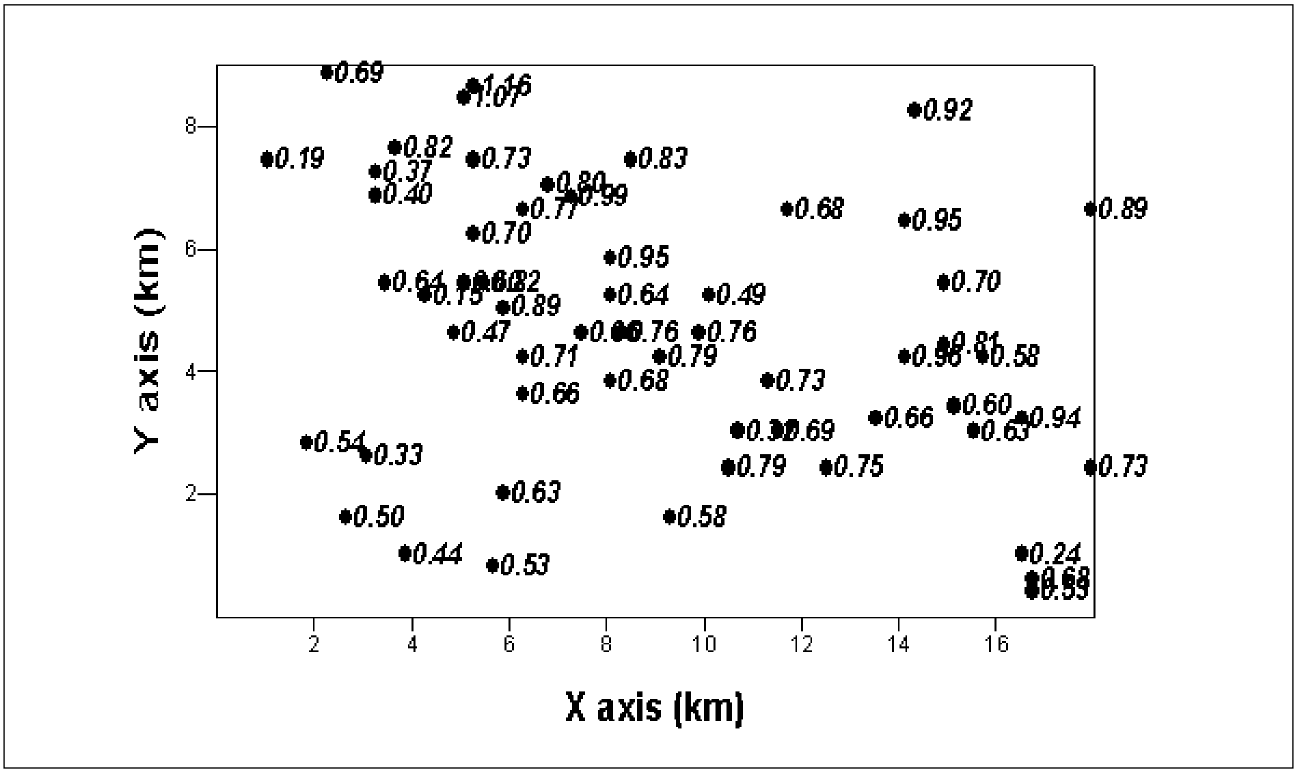


Figure 3. Location of groundwater sample collection sites and values of virus decay rate, Tucson Arizona (Yates and Yates, 1989).

are separated far away from each other. In this study, the *P-nearest neighbor method* (Moody and Darken, 1989) is applied to determine $\sigma_{x(k,j)}$ where P is the number of the nearest neighbor points. First, the root mean square distance (RMSD) between a center $\mathbf{X}_{(k,j)}$ and its P-nearest neighbors is determined for each GKF node:

$$RMSD_{(k,j)} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}\left|\mathbf{X}_i - \mathbf{X}_{(k,j)}\right|^2} = \sqrt{\frac{1}{P}\sum_{i-1}^{P}\left(\mathbf{X}_i - \mathbf{X}_{(k,j)}\right)^T\left(\mathbf{X}_i - \mathbf{X}_{k,j}\right)} \qquad (6)$$

where $\mathbf{X}_i$ is the i-th nearest neighbor point from the center $\mathbf{X}_{(k,j)}$of the *k-th* GKF node in class unit j. Then the width $\sigma_{x(k,j)}$ is given by $\sigma_{x(k,j)} = \mathrm{RMSD}_{(k,j)}/F$ where F is a control factor. Saha and Keeler (1990) stated that just one nearest neighbor, i.e. $P = 1$, can produce the desired performance.

(e) After setting the centers and the widths of the GKF nodes, the estimates at the observed points are obtained as $\hat{z}(\mathbf{X}_{(k,j)}) = \sum_2 / \sum_1$. Then, the root mean square error (RMSE) between the observed values $Z_{(k,j)} = Z(\mathbf{X}_{(k,j)})$ and the estimated values $\hat{z}(\mathbf{X}_{(k,j)})$ is determined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{Nc}\sum_{k=1}^{N^j}\left[\mathrm{Z}\left(\mathbf{X}_{(k,j)}\right) - \hat{z}\left(\mathbf{X}_{(k,j)}\right)\right]^2} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[\mathrm{Z}_n - \hat{z}(\mathbf{X}_n)\right]^2} \qquad (7)$$

Then, the width of the spatial variable z, σ_z is determined by $\sigma_z = \mathrm{RMSE}$.

Once the training is completed, the interpolation mode is performed as:

(a) Enter the set of spatial coordinate vectors $\{\mathbf{x}^m \mid m=1,\ldots,M\}$ where m is a given point in the region and M is the number of interpolation points.

(b) Obtain the interpolated value $\hat{z}$ ($\mathbf{x}^m$), the standard deviation of the estimate $\hat{s}$ ($\mathbf{x}^m$), the observation point density ρ ($\mathbf{x}^m$), and the posterior probability $P[\ C^j \mid \mathbf{x}^m]$ for each class.

After completing the interpolation mode, then the classification mode is accomplished by using the estimated posterior probabilities.

APPLICATION TO DECISION-MAKING OF REGIONAL GROUNDWATER CONTAMINATION

In this study, we used the virus decay rate data taken from Yates and Yates (1989), which were estimated at 57 pumping wells in the Tucson area, Arizona. Those well locations are indicated in Figure 3 and the basic statistics of the virus decay rate in this region are given in Table 1. SANN is applied here to identify the boundaries of contaminated areas where remedial actions for groundwater may be needed. The classification of the contaminated areas has been done considering different truncation levels.

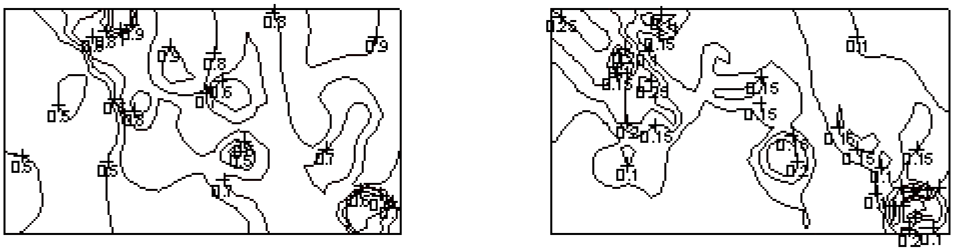
Table 1. Basic Statistics of Observed Virus Decay Rates

Mean (-$\log_{10}$ (virus particles) /day)	0.671
Standard Deviation	0.212
Coefficient of Variation	0.315
Skewness Coefficient	-0.310
Minimum	0.151
10 %	0.379
20 %	0.508
30 %	0.600
40 %	0.651
50%	0.684
70 %	0.771
90 %	0.946
Maximum	1.164

SANN was trained based on the virus decay rate data obtained from the 57 groundwater samples as above described. The control parameters were P =1 and F =1.3. The virus decay rate field $\hat{z}$ ($\mathbf{x}$) and the corresponding standard deviation field $\hat{s}$ ($\mathbf{x}$) were determined at 648 points on a 0.5 km x 0.5 km grid system. Figure 4 (a) shows the interpolated fields. The contour lines indicate that the western area has a small virus decay rate, which means that this area may be seriously contaminated by the virus. The corresponding standard deviations are shown in Figure 4(b). The values of the virus decay rate z were partitioned into two parts (cases) and the posterior probability that the virus decay rate at a point $\mathbf{x}$ belongs to a given class, $P(C^j|\mathbf{x})$, j=1,2 were determined by

$$\mathrm{P}\,(C^1 \mid \mathbf{x}) = \mathrm{P}\,[\ z(\mathbf{x}) \leq TL \mid \mathbf{x}]$$

= the probability that the virus decay rate at **x** is less or equal than the truncation level TL (**contamination probability**)



(a) Interpolated field

(b) Standard deviation field

Figure 4. Interpolated and standard deviation fields for virus decay rate sample, Tucson, Arizona.

$\mathrm{P}\,(C^{2} \mid \mathbf{x}) = \mathrm{P}\,[\,TL < z(\mathbf{x}) \mid \mathbf{x}\,]$

= the probability that the virus decay rate at **x** is larger than the truncation level *TL* (**non-contamination probability**)

where the variable *z* indicates virus decay rate and *TL* is a prescribed threshold level. In addition, the groundwater area has been classified by considering the following criteria:

Contaminated Area : if $\max(\mathrm{P}[C^{1} \mid \mathbf{x}], \mathrm{P}[C^{2} \mid \mathbf{x}]) = \mathrm{P}[C^{1} \mid \mathbf{x}]$

Non-Contaminated Area : if $\max(\mathrm{P}[C^{1} \mid \mathbf{x}], \mathrm{P}[C^{2} \mid \mathbf{x}]) = \mathrm{P}[C^{2} \mid \mathbf{x}]$

Four classification scenarios with varying truncation levels of 10 %, 20 %, 30 %, and 40 % of the observed virus decay rates were determined and the contaminant probability maps are shown in Figures 5 (a.1), (b.1), (c.1), and (d.1), respectively. The foregoing information may be useful for making probabilistic statements about whether a particular area is contaminated or not. For instance, the point located at coordinate (2 km, 6 km) has contamination probabilities of about 50 %, 70 %, 80 %, and 85 % for the truncation levels of 10 % (0.379), 20 % (0.508), 30 % (0.6), and 40% (0.651), respectively. Furthermore, the contaminated fields were identified with black as shown in Figures 5 (a.2), (b.2), (c.2), and (d.2) for the truncation levels of 10 %, 20 %, 30 %, and 40 % of the observed values, respectively. As expected, the contaminated area becomes larger as the truncation level increases. In all cases, the northwestern area appears to be the most seriously contaminated , and remedial action may be considered.

SUMMARY AND CONCLUSIONS

In this paper, the Spatial Analysis Neural Network Method has been applied to test its capability as a decision-making tool for a groundwater contamination problem. The variable that measures the severity of groundwater contamination was virus decay rate, and the 57 observations were taken from Yates and Yates (1989) over the specified region in Tucson, Arizona. For the groundwater contamination problems associated with the spatial variability of virus decay rate, two questions may often arise in the decision-making processes: (1) what is the probability that the virus decay rate at a given point is less than the standard level (contamination probability-of-occurrence); and (2) what are the boundaries dividing the contaminated and non-contaminated areas (optimal-classification problem). Those questions were answered by the posterior probability estimator and the Bayesian classifier provided by SANN. In this paper, we obtained the contamination probability maps and the classification maps with varying truncation levels of 10 %, 20 %, 30 %, and 40 %, respectively. How this spatial information can help one to make the decisions about the severity of groundwater contamination, as well as of the contaminated areas to be treated, was illustrated.

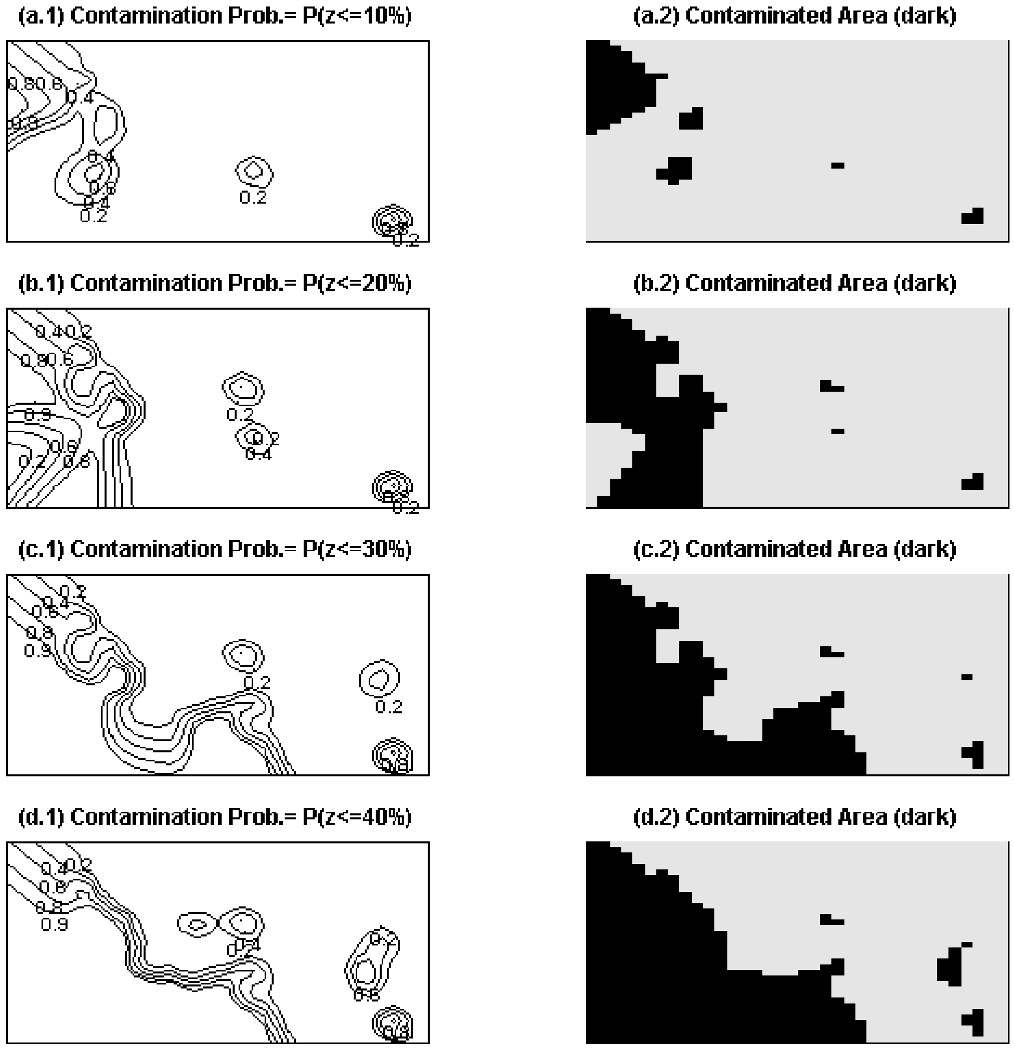

(a.1) Contamination Prob.= P(z<=10%)
(a.2) Contaminated Area (dark)
(b.1) Contamination Prob.= P(z<=20%)
(b.2) Contaminated Area (dark)
(c.1) Contamination Prob.= P(z<=30%)
(c.2) Contaminated Area (dark)
(d.1) Contamination Prob.= P(z<=40%)
(d.2) Contaminated Area (dark)

Figure 5. Contamination probability maps and contaminated area maps for truncation levels TL = 10%, 20%, 30%, and 40%, respectively.

REFERENCES

ASCE Task Committee on Geostatistical Techniques; Review of Geostatistics in Geohydrology: I. Basis Concepts, *J. of Hydraulic Engineering*, 116(5),612-632, 1990.

Bras R.L., I. Rodriguez-Iturbe; Random Functions and Hydrology, Addison-Wesley, 1985.

Craun L., R.C. Knox; Evaluation of Septic Tank System Effects on Ground Water Quality, *EPA-600/2-84-107*, Washington, D. C., 1984.

Haykin S., Neural Networks: A Comprehensive Foundation, Macmillan College Pub. Comp., Inc., 1994.

Markus M., H.-S. Shin, J.D. Salas; Predicting Streamflows based on Neural Networks, *1995 First International Conference on Water Resources Engineering*, ASCE, San Antonio, TX, 1995.

Moody J.E., D.J.Darken; Fast Learning in Networks of Locally-tuned Processing Units, *Neural computation 1*, 281-294, 1989.

Novotny V., H. Jones, X. Feng, A. Capodaglio; Time Series Analysis Models of Activated Sludge Plants, *Wat. Sci. Tech*., Vol. 23, 1107-1116, 1991.

Poggio T., F. Girosi; Regularization algorithms for learning that are equivalent to Multilayer Networks, *Science,* Vol. 247, 978-982, 1990.

Rashid A., A. Aziz, K.F. Wong; A Neural Network Approach to the Determinationof Aquifer Parameters, *Ground Water*, 30(2), 164-166, 1992.

Rogers L.L., F.U. Dowla; Optimization of Groundwater Remediation Using Artificial Neural Networks with Parallel Solute Transport Modeling*, Water Resour. Res*., 30(2), 4578-481, 1994.

Saha A., J.D. Keeler; In Algorithms for Better Representation and Faster Learning in Radial Basis Function Networks*. Advance in Neural Information Processing Systems 2*, Touretzky, D. S. et al., eds., 482-489, San Monteo, CA: Morgan Kaufmann, 1990.

Shin H.-S.; Uncertainty of Hydrological and Environmental Data Based on Spatial Analysis and Neural Network Modelling, Dissertation, Dep. of Civil Engr., Colorado State University, Ft. Collins, 1997.

Shin H.-S. J.D. Salas; Spatial Analysis Neural Network Model and its Applications to Hydrological and Environmental Data, *Water Resources Paper*, Dep. of Civil Engr., Colorado State University, Ft. Collins, 1997, submitted.

Yates M.V., S.R. Yates; Septic tank setback distance: A way to minimize virus contamination of drinking water, *Ground Water*, 27(2), 1989.

ADDRESS FOR CORRESPONDENCE

Hyun-Suk Shin
Engineering Research Center #220
Hydrologic Science and Engineering Program
Department of Civil Engineering
Colorado State University
Ft. Collins, CO 80523
U.S.A.

E-mail: hsshin@lamar.colostate.edu