**Rose, Antony; Sudheer, K. P.**

# Spatio-Temporal Modelling of Groundwater Quality Using Artificial Neural Network In Palar River Basin, Tamil Nadu

*Proceedings of ninth International Conference on Hydro-Science and Engineering (ICHE 2010), IIT Madras, Chennai, India.*
*2 - 5 August 2010*

# SPATIO-TEMPORAL MODELLING OF GROUNDWATER QUALITY USING ARTIFICIAL NEURAL NETWORK IN PALAR RIVER BASIN, TAMIL NADU

**Rose Antony[1], Sudheer, K.P[1]**

*Abstract: This paper proposes an application of a spatial analysis neural network (SANN) algorithm for regional characterization of water quality, and a case study on the Palar basin, Tamilnadu. The algorithm is based on non-parametric statistical analysis and the concepts of artificial neural networks. The SANN classifies the normalized and standardized data into certain degrees of hazard severity according to pre-defined truncation levels. Then posterior probabilities of hazard severity at any given point in the region are determined and the point is assigned a Bayesian hazard severity index. Using these indices, hazard severity maps of the region are prepared. The maps are useful in visualizing the spatial pattern of water quality hazard and based on these indices other properties such as duration, frequency etc. can be derived. The results of the study are encouraging and would help the decision makers to develop better water quality management plans for a given basin.*

## INTRODUCTION

The study of hydro-geochemical evolution in complex aquifers requires manipulation of a wide range of data of diverse origin. The physiochemical parameters indicate the diversity of ground water and various possible processes that take place through the aquifer. Due to growing concerns about water quality and pollution problems in the biosphere these studies are of great importance for a more reliable design of ground water protection schemes, optimal remediation strategies of contaminated aquifers and more adequate preventive measures for vulnerable sites. Field measurements of water quality parameters exhibit a large degree of spatial as well as temporal variability. The presence of a spatial correlation is often used to interpolate the data in space in order to establish a map of pollution, and it also provides useful insights on the structure of the water quality patterns. At the same time, when an assessment of epidemiological consequences of environmental factors is needed, decision makers will be less interested in obtaining an estimation of a pollutant's concentration at unsampled locations than in the probability of the pollutant's concentration to exceed a certain threshold. Gelher (1986) argues that the observed spatial variability suggest that it may be useful to describe such parameters in a stochastic context rather than in the traditional deterministic one. The traditional deterministic approach in which the aquifer properties are represented as a unique local parameter throughout the region is not realistic in many geological settings. On the contrary, their properties usually vary in a discrete or continuous manner on a multiplicity of scales from one location to another. Thus, often-encountered random spatial fluctuations cannot be adequately described by smooth deterministic functions (Sudicky, 1986), rather these parameters are uncertain. This uncertainty

[1] Department of Civil Engineering, Indian Institute of Technology Madras, Chennai-600 036.
rosejoycester@gmail.com, sudheer@iitm.ac.in

is due to the fact that the parameters are measured only at some sampled locations such as selected well locations and depth intervals, which are often sparse and/or due to the intrinsic complexity of the geological processes that causes the variability.
Different methods have been proposed from time to time to characterize the water quality in spatial as well as temporal domain. Kriging (Journel and Huijbegts, 1978) is an interpolating method of pattern completion that is widely used in the geo-hydrologic community. Rizzo and Dougherty (1994) criticizes this parametric method, as it needs re-estimation of variogram and other parameters, when more information is made available, and advocates the use of nonparametric methods for spatial analysis. Non-parametric methods make no assumptions about the functional form of the data distribution (Weiss and Kulikowski, 1997).
The main objective of the current paper is to present an approach to analyze and quantify the spatial and temporal patterns of ground water quality based on water quality data. By using a nonparametric spatial analysis neural network algorithm, the normalized and standardized water quality data are classified into certain classes (for example, extreme, severe, mild, and non-hazardous) based on a number of truncation levels corresponding to specified quantiles of the standard normal distribution (based on the BIS standards). The proposed water quality analysis approach was applied to the Palar River Basin in Tamilnadu, and the hazardous maps for calcium were prepared for the entire region.

## METHODOLOGY AND THERETICAL BACKGROUND

As discussed earlier, the paper presents an application of neural network approach to regional characterization of water quality. For instance, consider an area with 'S' sampling points of water quality, where observed data is available for N years. We consider the historical data represented by $Q_t(k)$, t=1,2,3…N and k=1,2,3…S, where N is the length of the record, k is a specific site and S is the total number of sites .Note that N covers the period from the earliest to the latest year of all the records, so the actual record length for a given site k may be shorter than N. Also site k is located at point $X_t(k)$, which is a two dimensional coordinate vector. The historical data $Q_t(k)$ for each site k are normalized and standardized so as to make the data at stations comparable (on the basis of the standard normal distribution). Such normalized and standardized water quality data, denoted by $Z_t(k)$ are statistically analyzed to obtain the conditional point estimate $z_t(x)$ at any point x in the study area and other spatial statistics such as the standard error of the estimate. For this purpose we use a spatial analysis neural network (SANN) algorithm the overall procedure is schematically summarized in Fig 1.

To obtain the water quality behaviour at a point, the value $z_t(x)$ is classified as belonging to any of the classes $\{C^1, C^2, C^3, C^4, \ldots..\text{etc.}\}$ that are associated with extreme ($C^4$), severe ($C^3$), mild ($C^2$)and non hazardous ($C^1$) area respectively, depending on whether $z_t(x)$ falls on a certain probability range of the standard normal distribution. Then for each year, t, and for equally spaced grid points $x_m$ (m=1,….M, where M is the number of grids), the conditional mean $z(x_m)$ and the posterior probabilities are estimated which is then used in preparing the water quality maps for each parameter.

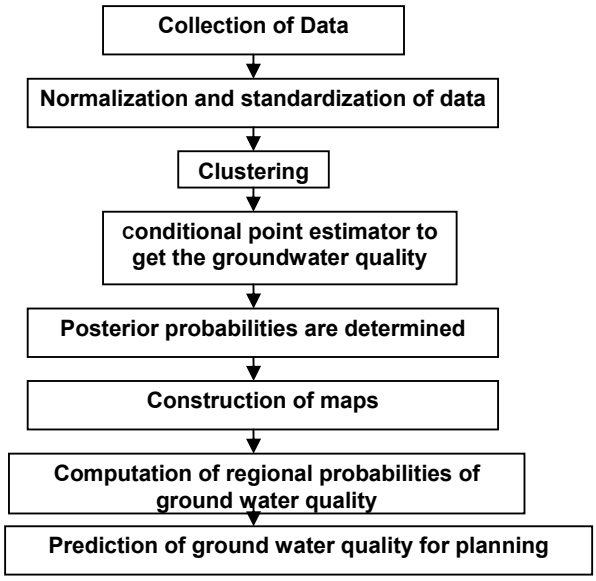


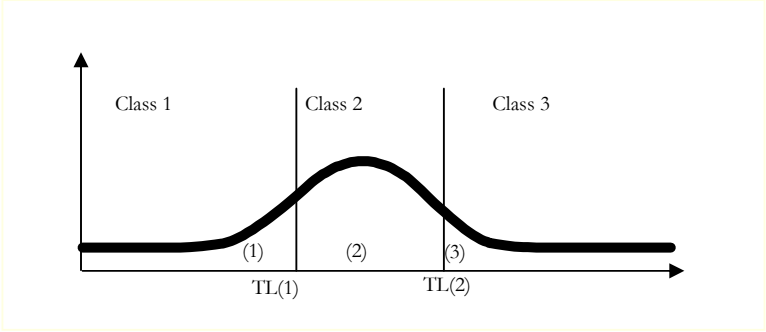

**Fig1. Flow chart of the spatial characterization procedure**

**Normalization, Standardization, and Truncation levels**

The first step in the analysis is to normalize and standardize the historical water quality data. The power and logarithmic transformations has been used to transform the historical data into normal. Let $Q_t(k)$ be the water quality data after transformation. These data are further standardized as:

$$Z_t(k) = \frac{Q_t(k) - \overline{Q}_t(k)}{S_Q(k)} \qquad (1)$$

where $\overline{Q}_t(k)$ and $S_Q(k)$ = mean and standard deviation of the $Q_t(k)$, respectively. Hence $Z_t(k)$ is assumed to be normally distributed, with a mean of zero and standard deviation of unity.Constant truncation levels are selected over the region and throughout the historical data to identify the area under threat.




$$\text{Area of (1)} : P[Z(\mathrm{x}) \le TL(1) \mid \mathrm{x}] = P[C^1 \mid \mathrm{x}]$$
$$\text{Area of (2)} : P[TL(1) < Z(\mathrm{x}) \le TL(2) \mid \mathrm{x}] = P[C^2 \mid \mathrm{x}]$$
$$\text{Area of (3)} : P[Z(\mathrm{x}) > TL(3) \mid \mathrm{x}] = P[C^3 \mid \mathrm{x}]$$

**Fig 2 Truncation levels used for classification of water quality**

Here the point falling below truncation level 1 is considered to belonging to the class 1 and so on. The present study has been completed for calcium and the probability distribution function (pdf) of the data has been generated and the truncation levels has been set according to the drinking quality standards recommended by the Bureau of Indian standards (BIS).The desirable and permissible limits for calcium in drinking water are 75mg/l and 200mg/l respectively (IS-10500:1991).

Conditional point estimator

A nonparametric spatial analysis algorithm for point estimation and classification of spatial data was developed. The analysis is based on Parzen's nonparametric point density estimators (Parzen 1962) and the Bayesian classifier (Bishop 1995). Consider a spatial variable $z$(x), which is available in a two-dimensional domain, i.e., x = [$x$, $y$]. Here, such spatial variable is the water quality data after normalization and standardization, also denoted as $Z_t$ ($k$), where $t$ is time in years and $k$ represents the site or gauging station. We want to estimate some spatial statistics at any (unknown) point x, such as the point conditional mean estimator $z^*$(x), the posterior probability $P$ [$C^j$|x] of each class severity at a point.

Suppose that in any given year $t$ water quality data is measured at $S$ sites in the region and the observation set is denoted as {$X_t$ ($k$), $Z_t$ ($k$)|$k$ = 1, . . . , $S$}. For estimating the posterior probabilities and Bayesian classification, the historical data $Z_t$ ($k$) is classified into one of the four classes, $C^1$, $C^2$, $C^3$, and $C^4$, that are associated with extreme, severe, mild, and non-hazardous, respectively. The classes are defined by the truncation levels $TL$(1), $TL$(2) etc. as indicated above. Then, the observation set {$X_t$ ($k$), $Z_t$ ($k$)|$k$ = 1, . . . , $S$ } is denoted as {$X_t$ ($i$,$j$), $Z_t$ ($i$,$j$)|i = 1, . . . , $S^j$ and $j$ = 1, . . . , 4}, where $i$ denotes a specific site in each class $C^j$, and $S^j$ is the number of sites belonging to class $C^j$.

Following Parzen (1962), Spetch (1991), and Bishop (1995), the conditional expectation of $Z$ given x may be written as

(1 )
$$z(x) = \frac{\sum_{j=1}^{4}\sum_{i=1}^{s_j}\left(\frac{1}{\sigma_X^2(i,j)}\right)Z_t(i,j)a_x(i,j)}{\sum_{j=1}^{4}\sum_{i=1}^{s_j}\left(\frac{1}{\sigma_X^2(i,j)}\right)a_x(i,j)} = \frac{\sum_2}{\sum_1}$$

where $a_x(i,j)$ = Gaussian kernel function (GKF) that acts as a transfer (activation) function. It is given by

(2)
$$a_X(i,j) = \exp\left[-\frac{D_X^2(i,j)}{2\sigma_X^2(i,j)}\right]$$

where $D_x(i, j)$ = Euclidean distance between the input vector x and the $i$th centre $X_{n-1}$ ($i$,$j$) in class $j$ and $\sigma_x$ is the width of the GKF node.To find the width of the GKF node the root-mean-square distance (*RMSD*) between centre $X_{n-1}$ ($i$, $j$) and its $P$-nearest neighbours is determined as

**(3)**
$$RMSD(i,j) = \sqrt{\frac{1}{P}\sum_{l=1}^{p}\left(Y_n - Y_{(n-1)}\right)^2 + \left(X_n - X_{(n-1)}\right)^2}$$

where $X_n$= nth nearest neighbour point from the center $X_{(n-1)}$ $(i,j)$ (of the $i$th GKF node in class unit $j$). Then $\sigma_x$ $(i,j) = RMSD(i,j)/F$, where $F$ is a control factor. Saha and Keeler (1990) stated that just one nearest neighbour, i.e., $P = 1$, can produce the desired performance. The biggest receptive field around a centre arises when $F = 1$. Musavi et al. (1992) stated that a reasonable separation between two centres is to cut the distance in half (i.e., $F = 2$) if uniform density is desired. Likewise, the posterior probabilities $P[C^j|\mathrm{x}], j = 1,2,3,4$ are determined by

**(4)**

$$P(C^j \mid x) = \frac{\sum_{i=1}^{S^j} \left( \frac{1}{\sigma_X^2(i,j)} \right) a_{x(i,j)}}{\sum_{j=1}^{4} \sum_{i=1}^{S^j} \left( \frac{1}{\sigma_X^2(i,j)} \right) a_{x(i,j)}} = \frac{\sum_{3,j}}{\sum_1}, \quad j = 1,2,3,4$$

For spatial analysis of the variable $z$ on a given region $R$, the Bayesian classifier provides a rule for assigning each point $x$ to one of the four classes. The region $R$ is regarded as being divided into sub regions $R_1$, $R_2$, $R_3$, $R_4$ so that a point falling in region $R_r$ is assigned to class $C^r$. The boundaries between the sub regions can be determined by using the Bayesian classifier (Cain 1990), as follows:

If max $\{P(C^1 \mid x), P(C^2 \mid x), P(C^3 \mid x), P(C^4 \mid x) = P(C^r \mid x)$ (5)

then d(x)= r

where r=0,1,2,3 and d(x) is the class indicator

**Implementation in Neural Network**

A neural network is a data processing system consisting of a large number of simple, highly interconnected processing elements in an architecture inspired by the structure of the cerebral cortex of the brain (Tsoukalas and Uhrig, 1997). In comparison to conventional computation techniques that employ complicated sets of equations to solve a complex problem, the ANN method uses very simple computational operations, such as addition, multiplication, and fundamental logical elements, to solve complex, mathematically ill-defined problems. In addition, the ANN technique possesses a self-organizing feature, thereby allowing it to solve a wide range of problems. In the proposed approach, feed forward neural network architecture is employed, and as it used for the spatial characterization, it is addressed as spatial analysis neural network (SANN)**.**

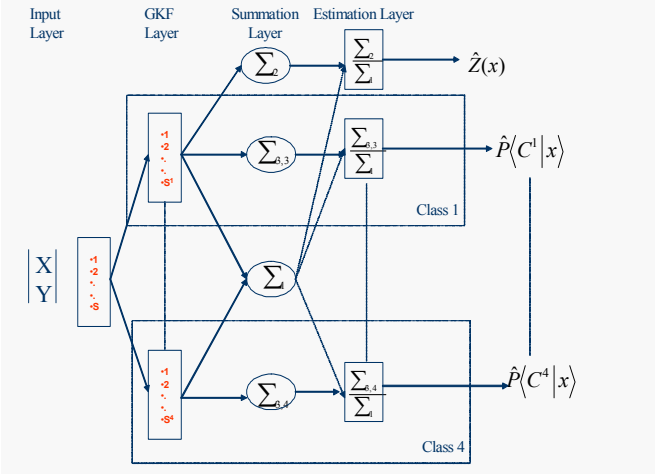



Fig 3 Spatial Analysis Neural Network

The SANN (Fig 3) consists of four layers, in which the neurons between layers are interconnected successively by the feedforward direction. The layers are called the input layer, GKF layer, summation layer, and the estimation layer. The input layer has $S$ nodes, each representing the x- and y-coordinates of an observation site, i. e. the vector x = [$x, y$]. This input coordinate vector is passed to the GKF layer with unit weights. In this layer, the observed set $\{X_t(k), Z_t(k) \mid k = 1,\ldots S\}$ is classified into four classes becoming $\{X_t(i,j), Z_t(i,j) \mid i = 1,\ldots S^j$ and $j = 1,..,4\}$, as stated earlier. Thus the GKF layer also consists of $S$ nodes, with $S^j$ nodes in each class $j$ such that $S^1 + S^2 + S^3 + S^4 = S$. The point $X^t(i,j)$ is then located at the center of the $i^{th}$ GKF node in class $j$ and the activation function of the GKF node $a_x(i,j)$ is determined. Thus, each GKF nodes has as internal parameters the center of the GKF node $X^t(i,j)$ and the smoothing parameter, $\sigma_x(i,j)$. The output from the GKF node is a function of the Euclidian distance between the center $X^t(i,j)$ and the input point x, and each GKF node only responds (or activates) when the input pattern falls within the receptive field, i. e. within the width of the GKF node $\sigma_x(i,j)$ (Poggio and Girosi, 1990). When the input vector x is placed at the center of the GKF node, the activation function becomes unity and it decreases exponentially as the input vector is farther from the center.

The outputs of the GKF nodes are passed to the summation layer with weighted connections and this layer estimates and gives the output, which in turn is passed to the estimation layer with unit weights. At this layer, the conditional mean and the posterior probabilities are determined. The SANN consists of two operation modes: training, and interpolation and classification. The training is carried out by a unsupervised training (Bishop, 1995), in that the input data set $X_t(i,j)$ alone is used to determine the parameters of the GKF. While this may not be optimal this gives a fairly good result (Shin and Salas, 2000). Once the training is completed, the interpolation and classification can be made. They essentially involve entering the coordinates for a number of $M$ points (unknown) points considered in the region of interest, and obtaining the interpolated values of conditional mean estimator and the posterior probabilities representing each class for the point. Thus, any point in the region of interest has a probability of belonging to a certain degree of hazard severity, so that the Bayesian hazard severity at a point x is that which has a maximum probability. This information is subjected to further analysis to characterize the water quality.

**Study Area and Data**

The Palar River Basin is a crystalline rock region in North Arcot District (Tamil Nadu).It lies between latitudes $12^0$14’40" N and $13^0$37’00" N and longitudes -$77^0$48’40" E and $80^0$14’40" E. The areal extent of the basin is 18,300 Km$^2$ (10,910.67 km$^2$,in TN). Palar River rises in the Eastern Ghats near Coimbatore, runs through Vellore and Chingelpattu districts of Tamil Nadu and terminates into the Bay of Bengal near Caturangapattinam. Paddy is the major crop in the basin followed by groundnut. The annual rainfall of the region is 1039mm (SW: 458 mm; NE: 461 mm). There are around 250,000 wells mostly dug wells.

The Palar River used to supply good drinking water to 30 towns on its banks and 50 villages surrounding it. The river water was also used by the villagers to cultivate their land. Now, there are a number of tanneries on the banks of the river Palar. They let out the effluents into it. As a

result the river water has been polluted and is not suitable for drinking or agricultural purposes. Due to pollution, the people are suffering from a number of diseases like asthma, skin disease and stomach ailment, etc. Thousands of acres of fertile land have become wasteland and it is not used for cultivation.

The water quality data for the study area is available for the year from 1973 to 2009.The data pertains to the following parameters viz: Calcium, Magnesium, Sodium, Potassium, Bicarbonate, Carbonate, Chloride, Nitrate, Total hardness, Total dissolved solids and Electrical conductivity.

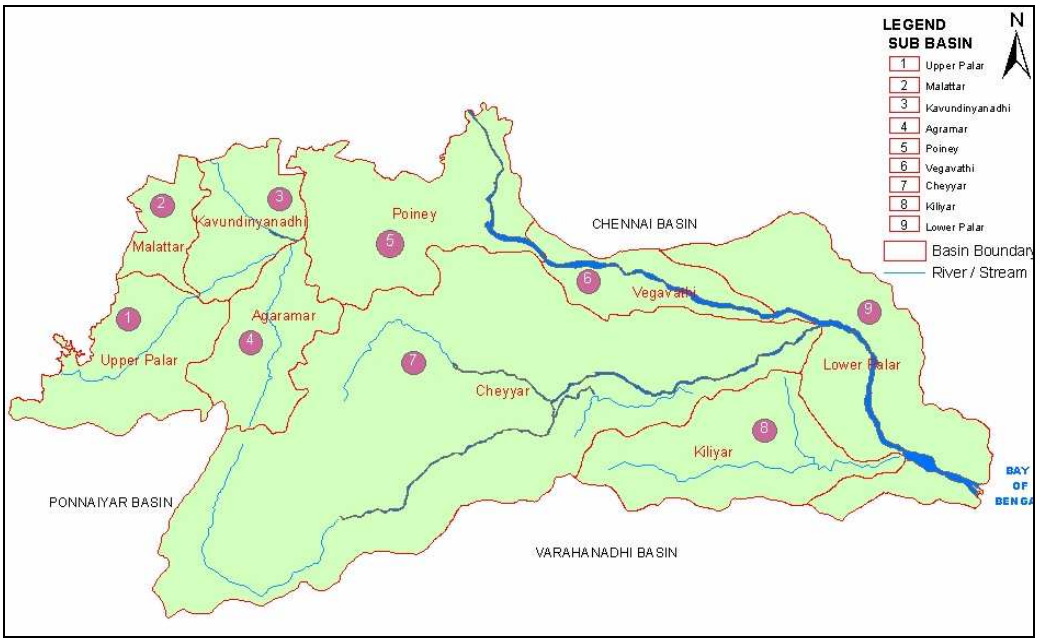



**Fig 4. The Palar River Basin**

## Result and Discussion

The SANN algorithm is applied on a yearly basis using the 126 well location data. After training, the (normalized standardized) water quality parameter is estimated for an area between $12^0$ and $13^0\,30'$ N latitude and $77^0$ and $80^0\,02'$ E longitude, at a grid system consisting of more than 200 sub areas, each having $0.1^0$ quadrangle of latitude and longitude. In addition, each sub area is assumed to be hydrologically homogenous. This assumption implies that every point in the sub area has approximately same water quality. All points in the grid system have been used as interpolation points. The interpolation is performed on a yearly basis for the entire study period of 1974-2009. Fig. 5 shows the calcium hazardous map of Palar River Basin for the year 1973 and 2008, where the areal distribution of the normalized standardized water content indicates that good portion of the region is having no calcium hazard(less than 40 mg/l).

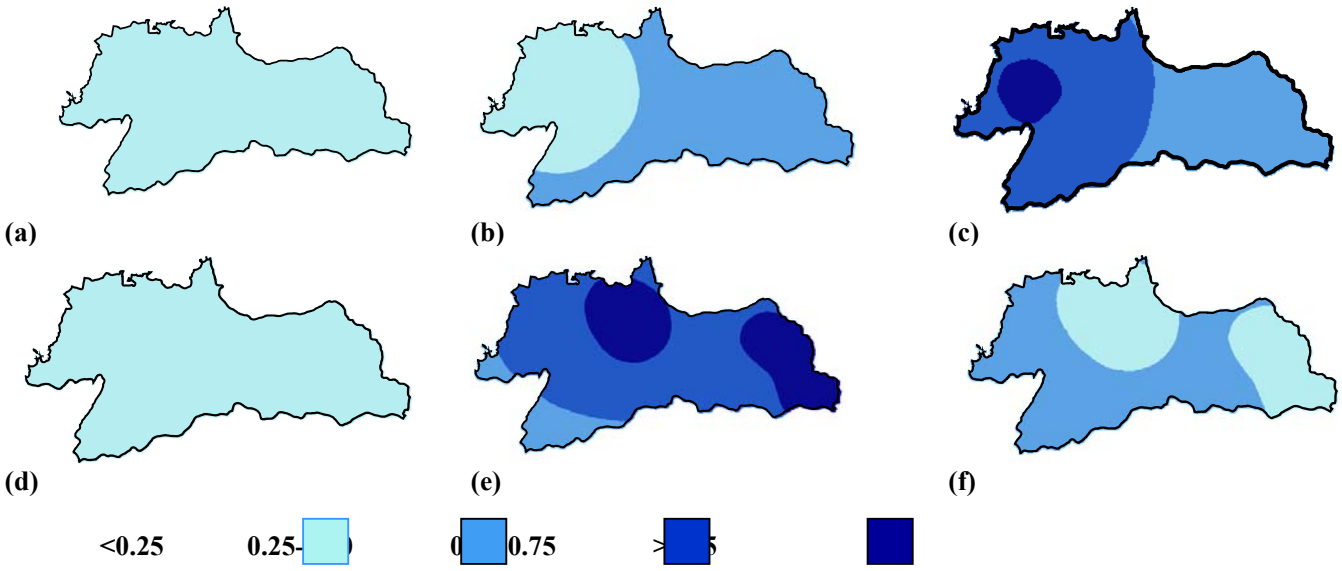



**Fig 5.Calcium hazardous map of Palar River Basin for the year 1973 (a,b,c) and 2008 (d,e,f) (a,d)Non-polluted (b,e) Mild calcium hazard (c,f) Severe calcium hazard**

The posterior probabilities defining various degrees of severity of the hazard are presented in figures 5(a-f) for the years 1973 and 2008. During the year 1973 the probability of the observation point to fall under severe hazard was high but in 2008 it drifted towards mild. For both the years the probability of falling under the non polluted class is very less. Severity of pollution remains high in the southern part of the region throughout the years. According to the water quality standards the southern part continues to lie within desirable and permissible limits (i.e 75mg/l to 200mg/l) of calcium in ground water. In 2008 the degree of severity has decreased, simultaneously there are more chances for the area to fall under mild calcium hazard. The maps signify that there is a drift of the calcium hazard towards east and there are also few places west of the river basin which are prone to pollution, but the degree of severity is localized. It was more concentrated near the Ambur region in the earlier years while it shifted to Arcot in the year 2008.Thus, the probability of getting ground water polluted is not steady, varies from place to place. It is observed that there are some leather factories in these areas which might have resulted for poor water quality. However, concrete studies are required to investigate the cause for this contamination. Further it can be seen that the spread of the contamination is not having any pattern in the basin, rather the contamination is localized, and most part of the basin is out of danger of a possible calcium contamination. The maps display the spatial distribution of the water quality over the basin and are useful for making probabilistic statement of hazard occurrence and severity at any arbitrary point or area.

## SUMMARY AND CONCLUSIONS

A method for spatial characterization of water quality in a basin based on the concepts of nonparametric statistics and neural networks is developed. A spatial analysis neural network is

described, implemented and applied to develop maps of discrete spatially distributed water quality variable. The SANN is data driven and requires no estimate of a covariance function unlike ordinary Kriging method. The maps that are prepared based on the proposed approach helps in identifying potential water contamination zones in the area. These maps are useful in visualizing the spatial pattern of water quality hazard and based on these indexes other properties such as duration, frequency etc. are derived. It is anticipated that the scientific views expressed and the information provided in this paper would aid to understand better the water quality characteristics, and for better management plans.

## ACKNOWLEDGEMENT

The project is partially supported by the Ministry of Water Resources, Government of India under the grant # R&D 528-537 in the year 2009.

## REFERENCES

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000a), "Artificial neural networks in hydrology I: Preliminary concepts", J. Hydrol. Engg., ASCE, 5 (2), 115-123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000b), "Artificial neural networks in hydrology II: Hydrologic applications" J. Hydrol. Engg., ASCE, 5 (2), 124-137.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). "Supervised and unsupervised discretization of continuous features". Proc of the 12th Int. Conference on Machine Learning, pp. 194-202. Tahoe, Morgan Kaufmann, California.

Duda, R, and Hart, P. (1973). Pattern classification and scene analysis. John Wiley & Sons, New York.

Gelher, L. W. (1986). "Stochastic subsurface hydrology: from theory to applications". Water Resources Research, 22(9): 135S-145S.

Hornik K, Stinchcombe M, White H. (1989). "Multilayer feed forward networks are universal approximators", Neural Networks, 2: 359–366.

Jha, R., Jain, C. K., and Bhatia, K. K. S. (2002) "ANN Applications for water quality modeling", Report of National Institute of Hydrology, Roorkee (in press).

Journel, A. G., and Huijbregts, C. J. (1978). Mining Geostatistics. Academic, San Diego, California.

Maier, H. R. and Dandy, G. C. (1996). "The use of artificial neural networks for the prediction of water quality parameters". Water Resources Research, 32(4): 1013-1022.

Raman, H. and Sunil Kumar, N. (1995). "Multivariate modeling of water resources time series using artificial neural networks", Journal of hydrological sciences, 40:145-163.

Ray, C., and Klindworth, K. K. (1996). ''Use of artificial neural networks for agricultural chemical assessment of rural private wells.'' Proc., North Am. Water and Envir. Conf., ASCE, New York, 1687–1692.

Rizzo, D. M., and Dougherty, D. E. (1994). "Characterization of aquifer properties using artificial neural networks: Neural Kriging". Water Resources Research, 30(2): 483-497.

Starrett, S. K., Najjar, Y. M., and Hill, J. C. (1996). ''Neural networks predict pesticide leaching.'' Proc., Am. Water and Envir. Conf., ASCE, New York, 1693–1698.

Sudheer, K. P., Nayak, P. C., and Ramasastri, K. S. (2003). "Improving Peak Flow Estimates in Artificial Neural Network River Flow Models". Hydrological Processes, 17(1): 677-686.

Sudicky. (1986). "A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in disperse process". Water Resources Research, 22(13): 2069-2082.

Tokar, S., and Markus, M. (2000). "Precipitation-runoff modeling using artificial neural networks and conceptual models". Journal of hydrologic engineering, ASCE, 5(2): 156-161.

Weiss, S., and Kulikowski, C. (1991). Computer systems that learn. Morgan Kaufmann, San Mateo, California.