Journal of Urban and Environmental Engineering, v.8, n.2, p.243-249

ISSN 1982-3932
doi: 10.4090/juee.2014.v8n2.243249

Journal of Urban and Environmental Engineering

www.journal-uee.org

DETERMINATION OF CONTAMINATATED WELLS TO NO_3-N: A NOVEL VULNERABILITY ASSESSMENT TOOL

Pijush Samui[1] * and Barnali Dixon[2]

[1]*Centre for Disaster Mitigation and Management, VIT University, India.*
[2]*Department of Environmental Science, University of South Florida, USA.*

Received 20 August 2014; received in revised form 13 October 2014; accepted 14 October 2014

Abstract: Contamination of well with nitrate-N (NO_3-N) possess various threats to human health. This problem becomes even more critical when these wells serve as source of drinking water as in the case of many rural parts of USA. This article employs Relevance Vector Machine (RVM) for determination of non-contaminated and contaminated well with nitrate-N (NO_3-N) in Polk County, Florida (USA). This research will provide a regional scale integrated GIS-based modeling approach to predict NO3-N contamination of ground water in a cost effective way. This approach also allows for higher true positive results (TPR) with fewer variables when data are imprecise and full of uncertainty which is common with available regional scale data). RVM technique is a Bayesian extension of the Support Vector Machine (SVM). Here, the RVM has been used as a classification tool. Well water quality data (nitrate-N) from 6,917 wells provided by Florida Department of Environmental Protection (USA) has been used to develop the RVM model. An equation has been also presented from the developed RVM model. The developed RVM has been compared with the Artificial Neural Network (ANN) and SVM models. This study shows that the developed RVM produces promising result for prediction of non-contaminated and contaminated well with N. The model is important because its real world applications enable water managers to more effectively manage contaminant levels within specific watersheds.

Keywords: Contaminated well, Relevance Vector Machine, Artificial Neural Network, support Vector Machine.

* Correspondence to: Celso A.G. Santos, Tel.: +55 83 3216 7684 Ext 27; Fax: +55 83 3216 7684 Ext 23.
E-mail: celso@ct.ufpb.br

INTRODUCTION

Groundwater is a major source of freshwater that is critical for human sustenance. Contamination of wells has direct and indirect consequences to human health. The determination of contaminated wells accurately and cost effectively is key to the successful management of this valuable fresh water resource and protection of human health. Contamination of wells with nitrate-N (NO_3-N) possess disproportionately higher health risk to rural population as oppose to city water consumers because in many rural areas these wells serve as source of drinking water.

In the United States 98% of the domestic water use (i.e. people who supply their own water) came from ground water (Kenny *et al.*, 2009). Although majority of USA water is provided by public suppliers in 2005, about 42.9 million people, or 14 percent of the total US population, supplied their own water for domestic use in 2005 (Kenny *et al.*, 2009). About 1.8 million people or 10% of Florida population is considered domestic self-supplied population in 2005 (http://ga.water.usgs.gov/edu/wudo.html). Florida is estimated to have estimated 940 000 domestic or private wells.

Additionally, according to the Department of Health, there are more than 2.6 million septic sewage systems in Florida serving about a third of the state's population. The presence of septic tanks makes the problem of groundwater contamination and associated health risk even more critical. Therefore, there is a need to develop methods to predict the risk of NO_3-N contamination in rural wells.

The purpose of the research is to develop and apply an innovative modeling method using relevance vector machine (RVM) to determine contamination levels in the wells based on soil, hydrogeological and land use parameters. This modeling approach can be used to identify wells with higher potential for vulnerability and the resulting maps can be used to monitor wells with high risk of contamination. Further, a map showing various level of contamination potential (such as low, moderate high) can facilitate identification of wells and families supplied by those wells to conduct rural health studies involving cohorts of people in rural areas to NO_3-N exposure from well water. Epidemiologists can use the predictions of well contamination potential to select wells and families using these wells for their studies to infer the risk of adverse health outcomes.

The determination of contaminated well is a challenging task and full of uncertainties (Dixon, 2009). Researches have used different methods for determination of contaminated well (Wagner, 1992; Scott *et al.*, 1998; Lin *et al.*, 1999; Hassan & Hamed, 2001; Kunstmann *et al.*, 2001). According to Scott *et al.* (1998) and Lin *et al.* (1999), statistical correlation between the contributing factors of contaminated well and the degree of contamination often does not produce reliable results.

Deterministic solute transport models have been also adopted for determination of contaminated well, however, this approach can computationally expensive at the watershed scale (Wagner, 1992; Hassan & Hamed, 2001; Kunstmann *et al.*, 2001). Therefore, there is a need to develop methods that are cost effective in contrast to site specific deterministic models but also reliable (Dixon, 2009).

Recently, Dixon (2009) successfully used Support Vector Machine (SVM), and Artificial Neural Network (ANN) for determination of wells contaminated with nitrate-N (NO_3-N) in Polk County, Florida (USA). In this research Dixon (2009) loosely coupled Geographic Information Systems (GIS) with ANN and SVM to relate all hydrogeological, soils and land use factors to well contamination data in a spatially explicit way.

This study examines application of Relevance Vector Machine (RVM) for determination of wells contaminated with nitrate-N (NO_3-N) with the same spatially explicit database for the Polk County, Florida (USA). RVM is a Bayesian extension of SVM (Vapnik, 1998). The goal is to improve current methodologies and develop new ones to improve our ability to predict groundwater contamination accurately and cost effectively where identification of true positive results (TPR, i.e. contaminated wells are identified as contaminated wells after using the classifiers) are increased significantly.

RVM has been used to classify contaminated and non-contaminated well and expected to increase TPR significantly as compared to previous work reported by Dixon (2009). Introduced by Tipping (2000), the RVM is based on Bayesian estimation theory, which can be applied for both classification and regression problems. Researchers have successfully used RVM for solving different problems (Tripathi & Govindaraju, 2007; Samui, 2007; Ghosh & Mujumdar, 2008). This study has the following aims:

- To examine the applicability of RVM in prediction of wells contaminated with nitrate-N (NO_3-N) in Polk County, Florida (USA)
- To develop an equation for determination of the contaminated well
- To conduct comparative accuracy assessments amongst the RVM model developed for this research and ANN as well as SVM model developed by Dixon (2009).

RVM Background

This section will describe the structure of RVM for classification problem. Let us consider a set of example of input vectors $\left\{x_i\right\}_{i=1}^{N}$ is given along with a corresponding set of targets $t=\left\{t_i\right\}_{i=1}^{N}$. For classification problem, t_i should be 0 for "Non-contaminated well" and +1 for "Contaminated well". The RVM constructs a logistic regression model based on a set of sequence features derived from the input patterns, i.e.

$$p(C_1/x)\approx\sigma\{y(x;w)\}$$

$$\text{with } y(x;w)=\sum_{i=1}^{N} w_i\Phi_i(x)+w_0 \qquad (1)$$

where Φ_i is an *i*th component of the basis vector function

$$\Phi(x)=\left(\Phi_1(x),\Phi_2(x),\ldots,\Phi_N(X)\right)^T=\begin{bmatrix}1,K(x_i,x_1),\\ K(x_i,x_2),\\ \ldots,K(x_i,x_N)\end{bmatrix}^T$$

, $w=\left(w_0,\ldots,w_N\right)^T$ are a vector of weights, $\sigma\{y\}=\left(1+\exp\{-y\}\right)^{-1}$ is the logistic sigmoid link function, T is transpose and $K\left(x_i,x_j\right)_{j=1}^{N}$ are kernel terms. Assuming a Bernoulii distribution for $P(t/x)$, the likelihood is written as (Tipping, 2001):

$$P(t/w)=\prod_{i=1}^{N}\sigma\left\{y\left(x_i;w\right)\right\}^{t_i}\left[1-\sigma\left\{y\left(x_i;w\right)\right\}\right]^{1-t_i} \quad (2)$$

We cannot integrate the weights analytically. The RVM adopts the following separable Gaussian prior, with a distinct hyper-parameter, α_i, for each weight,

$$p(w/\alpha)=\prod_{i=1}^{N}N\left(w_i/0,\alpha_i^{-1}\right) \qquad (3)$$

The optimal parameters of the model are then derived by minimizing the penalized negative log-likelihood,

$$\log\{P(t/w)p(w/\alpha)\}=\sum_{i=1}^{N}\begin{bmatrix}t_i\log y_i+\\ (1-t_i)\log(1-y_i)\end{bmatrix}-\frac{1}{2}w^TAw \quad (4)$$

If we differentiate twice **Eq. (2)**, the expression is given below:

$$\nabla w\nabla w\log p(w/t,\alpha)=-\left(\Phi^TB\Phi+A\right) \quad (5)$$

Where $B=diag(\beta_1,\ldots,\beta_N)$ is a diagonal matrix with $\beta_n=\sigma\{y(x_n)\}[1-\sigma\{y(x_n)\}]$

The following **Eq. (6)** has been used for updating hyper-parameter

$$\alpha_i^{new}=\frac{1-\alpha_i\Sigma_{ii}}{\mu_i^2} \qquad (6)$$

where μ_i is the ith posterior mean weight, $\sum_{ii}$ is the ith diagonal element of the posterior weight covariance. The property of this optimization problem is that the value of many w will be zero. The nonzero weights are called relevance vectors.

METHOD

Study Area and Data

This article aims to identify the contaminated well in Polk County, Florida (see **Fig. 1**).

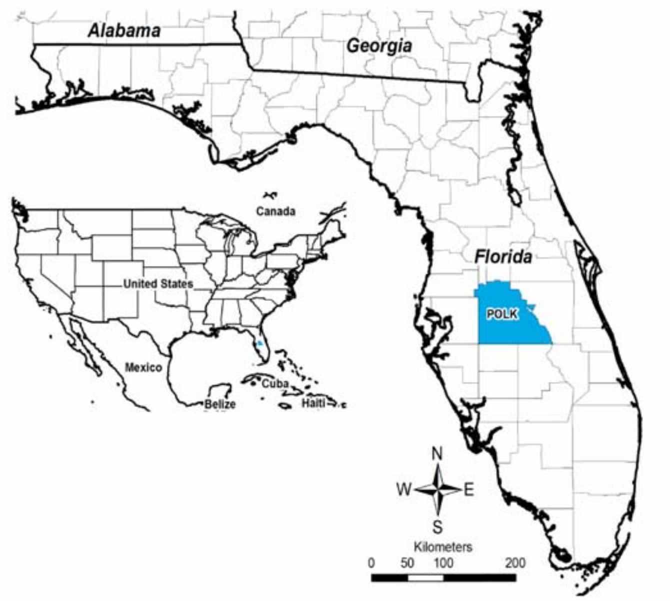

Fig. 1 Location of the study area.

The well data (nitrate-N) were collected with the help of Florida Department of Environmental Protection (FDEP) as a part of Water Supply Restoration Program (WSRP). **Figure 2** depicts the location of contaminated and non-contaminated wells in Polk County, Florida. 6917 wells data have utilized to develop the RVM model.

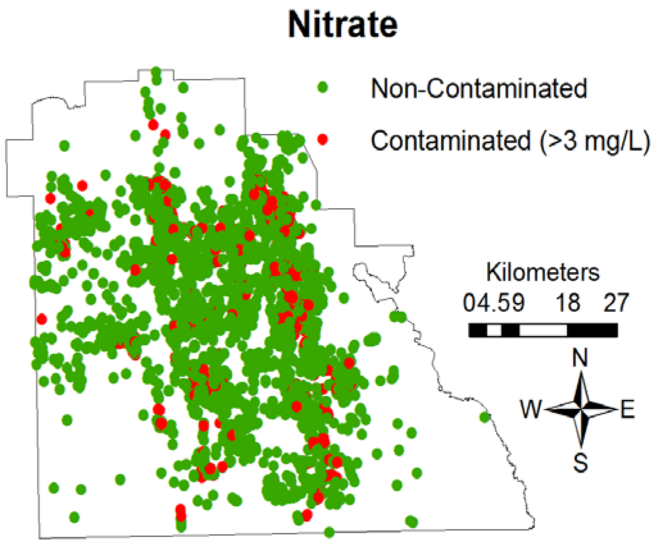

Fig. 2 Nitrate-N concentration in the wells in Polk County.

This article treats 933 wells as contaminated well and 5984 wells as non-contaminated well.

Development of RVM

The main aim of this study is the use of the above RVM model to classify between contaminated well with nitrate (N) and non-contaminated well in Polk County, Florida (USA). This study uses the spatially integrated database used by Dixon (2009). The GIS dataset contains information about depth to groundwater (D), recharge of aquifer (R), aquifer media (A), soil media (S), topography (T), impact of vadose zone (I), hydraulic conductivity (C), LULC (landuse), pedality, drainage, hydrologic group (hydrogrp), pH, organic matter (OM) and bulk density (BD) with respect to the location of 6917 wells. So, the input variables of RVM are D, R, A, S, T, I, C, LULC, pedality, drainage, hydrologic group (hydrogrp), pH, OM and BD.

Spatial extent of these variable in a GIS map format can be found in Dixon (2009). The target output for this RVM classifier is to predict wells that are contaminated or non-contaminated with NO_3.-N (Figure 2). Although the range of NO_3.-N concentration vary among these wells, for this classification purpose a Boolean approach was adopted, i.e. a value of 1 is assigned to contaminated wells while a value of 0 is assigned to non-contaminated wells. Any well that has NO_3-N concentration over 3 mg/L (**Table 1**) was considered as contaminated since this level of concentration indicates above background level (Moore *et al.,* 1986; USGS, 2004; Kelly *et al.,* 2005; Spechler & Kroening, 2007).

Table 1. Different concentration levels of the dataset

Concentration	Number of data
>10mg/L	396
3-10mg/L	537
<3mg/L	5,984

This study uses same training and testing dataset as used by Dixon (2009). Radial basis function ($K(x, x_k) = \exp\left\{-\frac{(x_k - x)^T (x_k - x)}{2s^2}\right\}$, where s is the width of the radial basis function) has been used as kernel function for RVM model. The program of RVM has been constructed by using MATLAB.

Performance Assessment

The accuracy of the developed RVM model has been determined by using the following equation (Dixon, 2009).

$$\text{Accuracy} = \frac{(\text{TPR} + \text{TNR})}{(\text{TPR} + \text{FPR} + \text{TNR} + \text{FNR})} \tag{7}$$

Where TPR is True positive rate, proportion of contaminated wells correctly detected, TNR is True negative rate, proportion of non-contaminated wells correctly detected, FPR is False positive rate, proportion of non-contaminated wells wrongly classified as contaminated and FNR is False negative rate, proportion of contaminated wells wrongly classified as non-contaminated. This study also adopts confusion matrix, Receiver Operating Characteristics (ROC) curves and Area Under the Curve (AUC) statistic to assess the performance of the RVM model. The details of confusion matrix, ROC and AUC are available in Dixon (2009).

RESULTS AND DISCUSSION

The design value of s has been determined by trial and error approach during training of RVM. The design value of s is 2. Number of relevance vector is 1580. The performance of training and testing dataset has been determined by using the design value of s. Hence, it has good generalization capability. The following equation has been developed from the RVM model.

$$y = \sum_{k=1}^{3328} w_k \exp\left\{-\frac{(x_k - x)^T (x_k - x)}{8}\right\} \qquad (8)$$

where x_k is input of the training dataset, x is the input of the data whose output is to be determined and T is transpose. Input variables are D, R, A, S, T, I, C, LULC, pedality, drainage, hydrologic group (hydrogrp), pH, OM and BD. **Figure 3** shows the value of w. User should use the value of w for using the **Eq. (8)**.

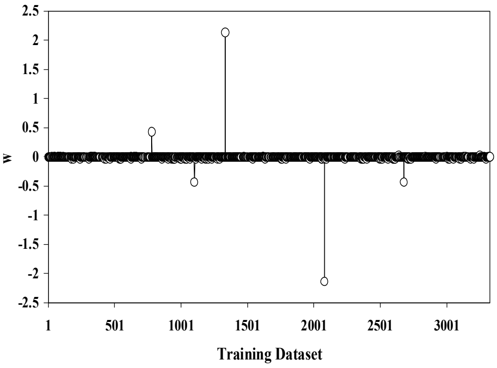

Fig .3 Values of w.

Figures 4 and **5** depict ROC curves of training and testing dataset respectively.

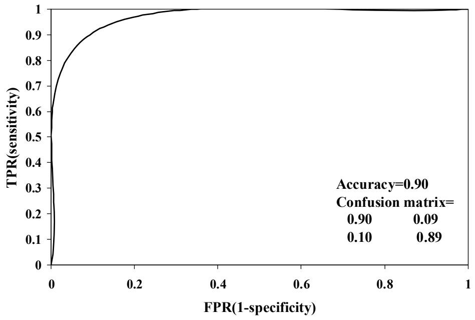

Fig. 4 ROC curve for training dataset.

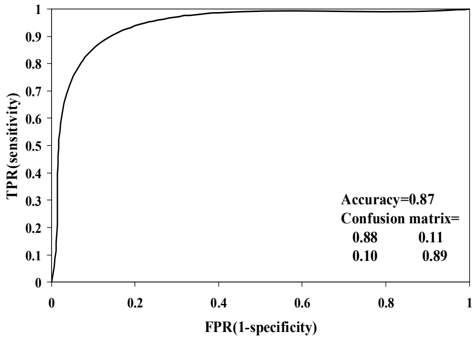

Fig. 5 ROC curve for testing dataset.

It is observed from **Figs 4** and **5** that the performance of RVM model is encouraging. **Figures 4** and **5** also show that the performance of training and testing dataset is almost same. So, the developed RVM model does not exhibit any overtraining phenomena. It is observed from **Figs 4** and **5** that the value of AUC is 0.969 and 0.924 for training and testing dataset respectively.

The value of AUS is close to one for training as well as testing dataset. So, the developed RVM has ability to predict contaminated well. The information about confusion matrix has been illustrated in **Figs 4** and **5**. **Table 2** shows the performance of the developed RVM. The results of RVM model have been compared with the ANN and SVM developed by Dixon (2009). **Figures 6** and **7** illustrate the bar chart of AUS and testing accuracy respectively.

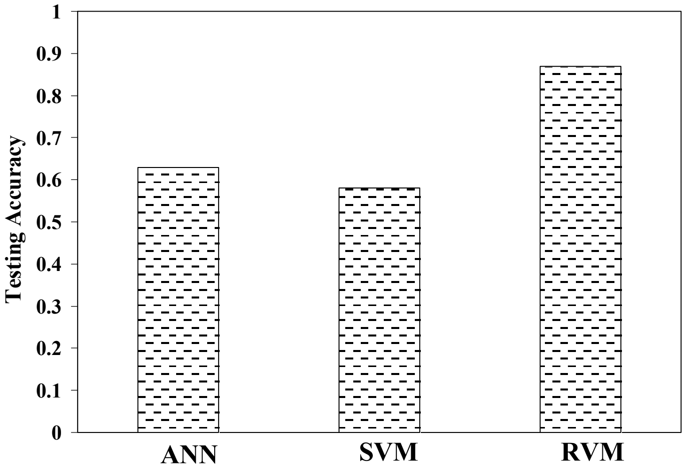

Fig. 6 Comparison between ANN, SVM and RVM in terms of Testing Accuracy.

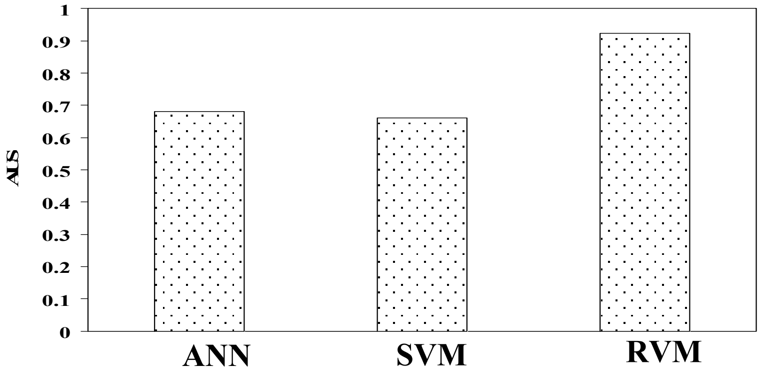

Fig. 7 Comparison between ANN, SVM and RVM in terms of AUS.

Table 2. performance of the RVM model

Dataset	TPR	TNR	FPR	FNR	Accuracy	Confusion matrix
Training	0.9084	0.8992	0.1007	0.0915	0.903893	0.90 – 0.09 0.10 – 0.89
Testing	0.8869	0.8728	0.1271	0.113	0.879943	0.88 – 0.11 0.10 – 0.89

It is observed from **Figs 6** and **7** the developed RVM model outperforms the ANN and SVM models. The developed RVM model only uses relevance vector for final prediction. Therefore, the developed RVM produces sparse solution. Sparseness means that a significant number of the weights are zero (or effectively zero), which has the consequence of producing compact, computationally efficient models, which in addition are simple and therefore produce smooth functions. ANN model uses many controlling parameters such as number of hidden layers, number of hidden nodes, learning rate, momentum term, number of training epochs, transfer functions, weight initialization methods, etc. SVM model uses two parameters (Capacity Factor (C) and kernel parameter), whereas, the developed RVM uses only one kernel parameter as a tuning parameter.

CONCLUSION

This study successfully applied RVM for classifying contaminated and non-contaminated well in Polk County, Florida (USA). In this study, 6917 data have been utilized to construct the RVM model. Of these wells, 933 were identified as contaminated (i.e. had concentration greater than 3 mg/L NO_3-N). Out of 933 contaminated wells, RVM correctly classified 793. The developed RVM produces sparse solution. It also shows good generalization capability. User can use the developed equation for determination of contaminated and non-contaminated well. The performance of RVM is better than the ANN and SVM models. RVM uses less tuning parameters compare to ANN and SVM models. RVM and SVM produce sparse solution, whereas ANN does not produce any sparse solution. This article shows that RVM is a reliable technique and has the potential for use in classifying between contaminated and non-contaminated well.

In summary, if RVM is used to assess regional scale vulnerability of groundwater, contamination of wells with NO_3-N can be identified reliably and cost effectively. User can use the developed equation for classifying contaminated and non-contaminated wells. SVM and ANN models did not give any equation. Integration of RVM with a GIS to predict contaminated wells can be a valuable tool when limited resources are available that restricts managers and health officials from sampling and analyzing all wells within their jurisdictions, and are forced to prioritize sampling strategies. The RVM model shows promising results in identifying contaminated wells. Hence, real-world applications of this methodology will enable water managers to more effectively monitor and manage contaminated wells within their jurisdictions.

REFERENCES

Dixon, B. (2009) A case study using support vector machines, neural networks and logistic regression in a GIS to identify wells contaminated with nitrate-N. *Hydrogeology.* **17**: 1507–1520.

Ghosh, S. & Mujumdar, P.P. (2008) *Statistical downscaling of GCM simulations to streamflow using relevance vector machine.* Advances in Water Resource. **31**(1): 132-146.

Hassan, A. & Hamed, K.H. (2001) *Prediction of plume migration in heterogeneous media using artificial neural networks.* Water Resour Res. **37**(3): 605–623.

Kenny, J.F., Barber, N.L., Hutson, S.S., Linsey, K.S., Lovelace, J.K. & Maupin, M.A. (2009) *Estimated use of water in the United States in 2005*: U.S. Geological Survey Circular. **1344**: 52.

Kunstmann, H., Kinzelbach, W. & Siegfried, T. (2001) Conditional first order second moment method and its application to the quantification of uncertainty in groundwater modeling. *Water Resource Res.* **38**(4): 1035

Lin, H.S., McInnes, K.J., Wilding, L.P. & Hallmark, C.T. (1999) Effects of soil morphology on hydraulic properties: I. quantification of soil morphology. *J. Soil Sci. Soc. Am.* **63**: 948–953.

Moore, D.D.,Martin, D.W.,Walker, S.T.,Rauch, J.T. & Jones, G.W. (1986) *Initial sampling results of an ambient background ground-water quality monitor network in the Southwest Florida Water Management District, FL.* Southwest Florida Water Management District Brooksville,393.

Pannol, S.V., Kelly, W.R., Martinsek, A.T. & Hackley, K.C. (2005) *Estimating background and threshold nitrate concentrations in a karst aquifer using probability graphs.* Geol Soc Am Abst Prog. **37**(7): 325.

Samui, P. (2007) *Seismic Liquefaction Potential Assessment by Using Relevance Vector Machine*. Earthquake Engineering and Engineering Vibration. **6**(4): 331–336.

Scott, H.D., Griffis, C.L. & Udouj, T.H. (1998) *Well contamination and related spatial attributes.* International Water Resources Engineering Conference Memphis TN.1200–1205.

Spechler, R.M. & Kroening, S.E. (2007) Hydrology of Polk County, Florida. US Geol Surv Sci Invest Rep. **114**: 2006–5320.

Tipping, M.E. (2000) *The relevance vector machine*. Advances in Neural Information Processing Systems. **12**: 625-658.

Tipping, M.E. (2001) *Sparse Bayesian leaning and relevance vector machine. J. Machine Learning Research.* **1**: 211–244.

Tripathi, S. & Govindaraju, S. (2007) *On selection of kernel parametes in relevance vector machines for hydrologic application.* Stoch Environ Res Risk Assess. **21**: 747–764.

US Geological Survey, (2004).USGS Circular 1207, US Geological Survey, Reston, VA [cited 14 Jan 2009].

Vapnik, V.N . (1998) *Statistical learning theory*, Wiley, New York.

Wagner, B.N. (1992) *Simultaneous parameter estimation and contaminant source characterization for couples groundwater flow and contaminant transport modeling. J. Hydrological*. **135**: 275.