Report for 2005FL100B: Development of an integrated methodology to assess vulnerability of groundwater to pathogen intrusion using GIS, remote sensing, neural networks and neuro-fuzzy methods

Publications

- Conference Proceedings:
  - Dixon, B, Candade, N. and Julie Earls. 2006. Development of an integrated methodology to assess vulnerability of ground water to pathogen intrusion using GIS, remote sensing, neural networks and neuro-fuzzy methods. American Association of Geographers, Annual Meeting, Chicago, IL, March.
  - Dixon, B. and Candade N. 2005. Integrated GIS and machine learning algorithms applied to ground water contamination mapping: a comparative study. Applied Geography Conference. Washington D.C. November.
  - Dixon, B. and Candade N. 2005. Groundwater Contamination Mapping Using Integrated GIS and Neural Networks: A Sensitivity Analysis. Presentation. International Conference on Environmental Science and Technology. January, New Orleans.
  - Candade, N. and B. Dixon. 2005. Effects of Training Sizes and Dimensionality on NN and SVM Performance: A Comparative Study. American Association of Geographers, Annual Meeting, Denver, CO, April.

Report Follows

Development of an integrated methodology to assess vulnerability of ground water to pathogen intrusion using GIS, remote sensing, neural networks and neuro-fuzzy methods

FINAL COMPLETION REPORT

Subcontract UF-EIES-0404012-USF (3/1/05 - 2/28/06)

PI: Barnali Dixon Ph.D.
University of South Florida St. Petersburg
Davis Hall 258
140 Seventh Ave South
St. Petersburg
FL 33701
Phone: 727 553 4025
Email: bdixon@stpt.usf.edu

COMPARISON OF NEURAL NETWORK AND NEURO-FUZZY TECHNIQUES IN GROUND WATER VULNERABILITY MAPPING: A CASE STUDY

Barnali Dixon[1]

1. INTRODUCTION

Contamination of surface and subsurface waters by anthropogenic activities has been a major concern of agencies involved with water management, water quality, water quantity and human health. Ground water (GW) accounts for 60% of the fresh water withdrawals in Florida and about 51% of this water is being used without further treatment or disinfection (Marella, 19999). Occurrence of well-drained sandy soils and karst features along with high rainfall makes Florida's GW, a major source of freshwater supply, vulnerable to contamination (Berndt et al, 1998; Purdum, 2002; Lee et. al, 2002). The proportion of outbreaks associated with groundwater sources in Florida increased 87% from the previous reporting period, and these outbreaks were primarily associated (60.7%) with consumption of untreated groundwater (Lee et. al, 2002). Close connection between ground and surface water, common in Florida, means that pathogens found in surface water may find their way into GW and vice versa. In recent years, Florida has been one of the most common relocation destinations in the US. Florida's population grew from 4 million in 1955 to 16 million in 2000, the highest growth rate in the nation. As a result, we have two inevitable problems throughout Florida i) increased amount of wastewater treatment and resultant sludge production, and (ii) increased number (and density) of septic systems. One of the dominant ways of sludge disposal is land application. Florida Department of health (FLDEP) has established detailed regulations for processing sludge before application and controlling the application of sludge to land. In 2003, 66% of the sludge was land applied in Florida, 17% were land-filled, and remaining 17% accounted for distribution and marketing. (http://www.dep.state.fl.us/water/wastewater/dom/reshome.htm). In Florida, parks and golf courses are common sites for Class A sludge application whereas many farmers apply Class B sludge to their pasture and farmland to reduce cost of fertilizer and lime. Since most processes used for complete pathogen/viral inactivation is not sufficient (EPA, 2003), landowners and the public as well as regulatory agencies are justifiably concerned about potential negative impacts of the potential spread of pathogens and resultant outbreaks. Therefore, there is a need to adopt waste application practices that take into consideration soil properties, hydrogeology, hydraulic loading and contaminant transport characteristics to minimize pathogen contamination risk (EPA, 2003). Additionally, in Florida, 31% of the population is served by estimated 2.3 million septic systems. These systems discharge over 426 million gallons of waste water per day into the subsurface soil environment (Florida Dept. of Health (DOH) http://www.doh.state.fl.us/environment/OSTDS/intro.html). Inadequately treated sewage from septic systems can lead to contamination of groundwater and poses a significant threat to drinking water and human health (http://www.epa.gov/owm/septic/pubs/homeowner_guide_long.pdf). There are no easy solutions to the sludge disposal or septic tank problems in Florida.

Traditionally state and county regulators used fixed setback distances for sludge application and septic tank locations for all geologic setting in their jurisdiction to protect our water resources (EPA, 2003). One approach in determining setback is to the use travel time using GW flow characteristics (Yeats and Yeats, 1987). A comprehensive study conducted by Matthesss et al (1984) that used the aforementioned approach showed that ground water flow-based 50 day residence time was not adequate for all of the sites for virus reduction. It takes longer and varied between 160 days and 270 days (EPA, 2003). Study conducted for Ground Water Rule showed that setback distances were found to be quite variable (EPA, 2000). Some distances were scientific and others

---

[1] Assistant Professor, University of South Florida, St. Petersburg. 140 seventh ave south, St. Petersburg, Fl 33701. Phone: 727 553 4025, bdixon@stpt.usf.edu

were holdovers from past practices (EPA, 2003). Very few of them considered preferential flow paths common in Florida karst. A table summarized by EPA (2003) listed critical factors that control pathogen/viral transport. They are: soil moisture content, type and depth of the soils, soil porperties such as organic matter and pH as well as hydraulic conditions to name a few. It is obvious that these factors vary over the landscape. Therefore, one size fits all mode of regulations for establishing setback distances might not be adequate. Conducting site- specific studies (on a case by case basis) to regulate entire Florida will clearly be cost prohibitive. Therefore, there is a need to strike a balance between expensive site-specific studies and broad-based one-size fits all policy. We are proposing to develop a spatially explicit method that will provide a vulnerability map for an area based on similar hydrogeological, topographical, climatological, soils, preferential flow pathways and landuse. This will be a useful environmental management tool to establish setback rules. This vulnerability map will have explicit representation of the possible level of risk associated with GW vulnerability to pathogens. In a gross sense, information from soil surveys, hydrologeological parameters, and landuse will be incorporated in a screening tool that will provide an indication of the level of risk a particular site may have to GW contamination by pathogen.

Coupling of neural networks (NN) and neuro-fuzzy models with a Geographic Information System (GIS) will facilitate vulnerability mapping of a complex system with enhanced spatial visualization capabilities of the models as suggested by Burrough (1996), Corwin et al., (1996). Integration with GIS will allow us to evaluate sensitivity of NN and neuro-fuzzy in a spatial context.

NN are multi-input, multi-output nonlinear models and can represent the complex interactions among the input/output parameters. In recent years NN has been successfully used in solving difficult hydrological and environmental problems. One major criticism is that it is not possible to determine how the solution was found due to the inherent black box nature of the NN. Also, it is also not possible to insert prior knowledge to a NN. The question is does NN need prior knowledge?? Also, how sensitive NNs are??

Incorporation of fuzzy logic with a GIS has shown to reduce error propagation (Wang et, al, 1990; Burrough et al 1992; De Gruizter, et al. 1997). Neuro-fuzzy modeling is an approach where the fusion of NN and Fuzzy Logic find their strengths and complement each other **(**Dixon, 2001, 2002, Khan, 1999, Nauck and Kruse, 1999). A key disadvantage of fuzzy logic based approach is inability to meet pre specified accuracy and lack of self-learning and generalization capability.

Neuro-fuzzy approach employs heuristic learning strategies derived from the domain of NN theory to support the development of a fuzzy system. A marriage between NN and fuzzy logic techniques should help overcome the shortcomings of both techniques discussed at length by Nauck and Kruse (1999). A neuro-fuzzy technique can learn a system's behavior from a sufficiently large data set and automatically generate fuzzy rules and fuzzy sets to a pre-specified accuracy level. They are capable of generalization, thus overcoming to the key disadvantages of fuzzy logic based approach. A fusion of NN and fuzzy logic provides a system that usually requires less computational power but has the ability to generalize and learn through the convergence of net. The research reports a case study of Polk County, Florida. This County was selected for its extensive agricultural landuse and the presence of underlying alluvial aquifer. This study aimed at using selected parameters from the DRASTIC model (Aller et al.,) with the NN and Neuro-fuzzy. Authors are aware of the strength and weaknesses of the DRASTIC model. The authors used parameters from the DRASTIC in this study because of readily available GIS layers for the Polk County. The intention of this study is not to promote or criticize DRASTIC.

2. OBJECTIVES

This study aims at comparing the vulnerability maps developed using NN and neuro-fuzzy methods. Specific objectives are to: i) compare the NN models with neuro-fuzzy and (ii) to check the sensitivity of NN and neuro-fuzzy models to training data.

3. METHODOLOGY

3.1 Digital Database

The DRASTIC model (Aller et al, 1987) is comprised of seven hydro geological variables: Depth to ground water (D), Recharge of aquifer (R), Aquifer media (A), Soil media (S), Topography/slope (T), Impact of vadose zone (I) and hydraulic Conductivity (C). Only four out of these seven variables were used as inputs for both NN and Neuro-fuzzy models. A, T and C were not used in the models due to lack of variability. All seven parameters were derived from the primary data layers in a GIS. The primary data layers used in this study are potentiometric surface, elevation, soils and geology. All of these primary data layers except soils were obtained from US Geological Survey (USGS). The soils data were obtained from the Natural Resource Conservation Service (NRCS). Digital Elevation Models (DEMs) with 30m resolution were used to generate elevation data for the study area. Potentiometric surface data provided by USGS was collected during the fall of 1996. This data was recorded in contour line with 20 ft interval. Potentiometric surface was generated using the GRASS command s.surf.tps. The data layer for D was generated by subtracting potentiometric surface from elevation. The data layer for net recharge was obtained from USGS. The net recharge was calculated based on the past behavior of the aquifer using MODFLOW at a one square mile cell resolution. The output from the MODFLOW was a site file. The site file was interpolated in GRASS to create the data layer R. The data layer for S was created through a multi-step process. Soil leaching index and soil pesticide leaching potential data were used to create the layer S. Soil leaching index layer was obtained from annual precipitation and soil hydrologic group. Soil pesticide leaching potential was created from soil attenuation, soil infiltration and soil permeability data layers. Soil attenuation information was generated from GLEAMS model. Please refer to Smith et al. (1994) for details. Thickness of the clay cap (I) is an important property since it influences the recharge to the aquifer and pesticide adsorption and degradation processes. GRASS command s.surf.tps was performed on the point data provided by USGS to create the interpolated surface. SSURGO data were used create maps for bulk density (BD), soils drainage class (D), soil Hydrologic group (H) (referred together as DH), soils structure or pedality. Landuse data we obtained from SWFWMD (1999).

Water quality data from 55 wells were used for validation of the models. The water quality data was provided by the Florida Department of Environmental Protection (FLDEP) in an excel spreadsheet containg well ID with locations of wells collected using a Global Positioning System (GPS). GRASS command s.menu was used to create site files for the wells. The wells then were reclassed into 2 categories: contaminated wells and non-contaminated wells. Single occurrence of E Coli was considered as contaminated well. (See Appendix A) for summery..

3.2 Coupling of NN and Neuro-fuzzy with GIS

Table 1. Example of Training Data used with NEFCLASS-J.

Nefclass J inputs				Output vulnerability			
D	R	S	I	Low	Mod low	Moderate	High
NF1							
3	1	1	5	1	0	0	0
7	3	9	7	0	0	1	0
7	1	5	8	0	0	0	1
9	1	2	5	0	1	0	0
NF2 & NF3							
63	1	1	35	1	0	0	0
23	3	9	15	0	0	1	0
23	1	5	10	0	0	0	1
10	1	2	35	0	1	0	0

Modeling of GW vulnerability was accomplished by loosely coupling GIS (GRASS 4.1) and NN software PREDICT (Neuralware, 2001, version 2.4) and the Neuro-fuzzy software Nefclass-J (Nauck and Kruse, 1999 version 1.0). The output function (single column n ASCII output) of NEFCLASS-J was modified to make files compatible with the GRASS. The NN software PREDICT has limitation of number of rows of data it can take. It can only take 132,000 of rows for each run. The application data for the Polk County consisted of 4,093,760 rows. So a custom code in VC++ was written to break down our application dataset in manageable size for the PREDICT. This custom software is available at the website:www.stpt.usf.edu/bdixon/gal/mainpage_final.html.

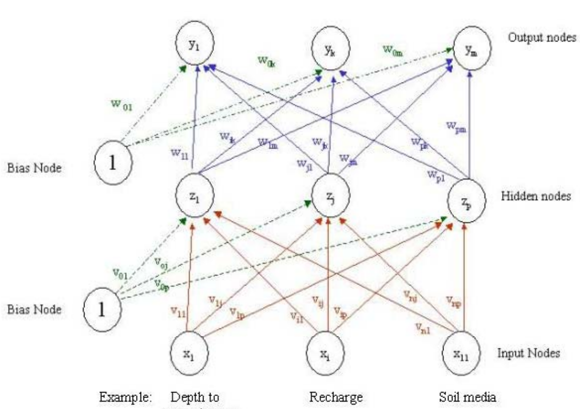
Table 2. Example of Training Data used with PREDICT.

NN inputs				Output
D	R	S	I	Vulnerability
NN1				
3	1	1	5	1 Low
7	3	9	7	3 Moderate
7	1	5	8	4 High
9	1	2	5	2 Moderately Low
NN2 & NN3				
63	1	1	35	1 Low
23	3	9	15	3 Moderate
23	1	5	10	4 High
10	1	2	35	2 Moderately Low

Use of NN and Neuro-fuzzy requires training data and application data. The training and application data for the NEFCLASS-J and PREDICT was obtained from the GIS. GRASS command r.stats was used to create training dataset. This GRASS command generated all possible combinations of D, R, S and I for the County. The training data consisted of 408 rows. Examples of training data are given in the Tables 1 and 2.

3.3 Development of Neural Network model

The Standard Back Propagation (SBP) architecture provided by PREDICT was used to perform classification. Figure 1 shows the Multi Layer Perceptron (MLP) network architecture. SBP is a method for training the MLP. It is a method for assigning responsibility for mismatches to each of the processing elements in the network; this is achieved by propagating the gradient of the objective function back through the network to the hidden units. Based on the degree of responsibility, the weights of each individual processing element are modified iteratively to improve the objective function.



Use of NN is a 3-step process: i) training, ii) testing and iii) application. The entire training data was divided into 2 groups training (286) and testing (214) data sets. Once the NN was trained and tested, application data consisting of 4,093,760 rows were used to generate ground water vulnerability maps. A batch file written in PREDICT automatically ran all the groups that were created, and produced the single column output. Examples of training parameter are presented in the Table 3.

Figure 1. Multi layer Perceptron Architecture

The training data for NN models was modified by adding a 0, 0, 0, 0 value to a total number of training pattern of 410. This was necessary because it was noted during simulation that an input of 0,0,0,0 in the application data resulted in a valid class of 1 (which indicates low vulnerability). But in reality these zero value represented data value out of the GIS mask but within the region. The following parameters were used with the PREDICT classifier:

Table 3: Parameters of NN Classifier

Learning Rate	Hidden layer=100
	Output layer=0.01
Learning Rule	Adaptive gradient
Variable selection model	Multiple regression
Training and testing	10-fold cross validation

The training data obtained from GRASS was inspected and classified based on expert's opinion according to the relationships between input parameter and the output vulnerability category. D and I are inversely related to the vulnerability categories whereas S and R are directly related. Examples of training rules are presented in the Table 4.

Table 4. Example of Rules used to Define Vulnerability.

Inputs [a]								Output vulnerability categories [b]
D	(ft)	R	(inch/yr)	S	(rate)	I	(ft)	
Low	(0-5)	Low	(0-1)	Mod low	(mod slow)	Low	(11-20)	Moderate
Low		Low		High	(rapid)	Low		High
Low		Mod low	(5-7)	Low	(slow)	High	(51-75)	Moderately low
Low		Mod low		High		Low		High
High	(51-75)	Mod low		Mod low		Mod	(31-50)	Low
High		Mod	(8-10)	Mod low		Mod low	(21-30)	Moderate
High		High	(20)	High		Low		High
High		High		High		High		Moderate

a = inputs obtained from raster data layers
b = output vulnerability categories: manually classified based on expert's opinion.

3.4 Development of Neuro-fuzzy model

Neuro-fuzzy model also uses a supervised learning-like algorithm based on fuzzy error back propagation (Figure 2). The learning procedure for the fuzzy sets is a simple heuristics. It results in shifting the membership functions and in making their supports larger or smaller (Nauck and Kruse, 1999). The adaptation of fuzzy sets is carried out by simply changing the parameters of its membership function in a way that the membership degree for the current feature value is increased or decreased respectively.

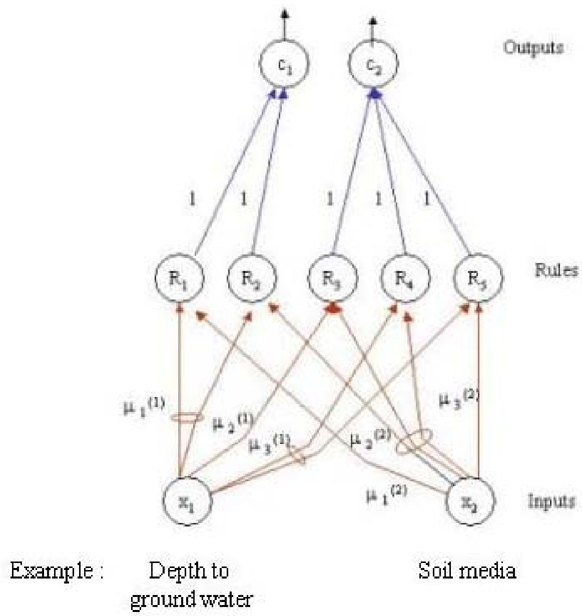


Figure 2: Neuro-Fuzzy architecture

In this research, the trapezoidal membership function was used since it was the most stable as compared to triangular and bell shaped membership functions (Dixon, 2001). The training parameter for the Neuro-fuzzy models are presented in the Table 5. Mathematically, trapezoidal membership function can be defined as follows:

$$(x : a, b, c, d) = \begin{cases} 0 & x < a \\ (x - a)/(b - a) & a <= x < b \\ 1 & b <= x < c \\ (d - x)/(d - c) & c <= x < d \\ 0 & x >= d \end{cases}$$

For detailed description of membership functions and Neuro-fuzzy architecture please refer to Nauck and Kruse (1999). Use of Neuro-fuzzy model is also a 3-step process: i) training, ii) testing/validation and iii) application. The testing/validation technique used with the neuro-fuzzy model is ‘cross validation’. This approach randomly divides the training data into the number of parts specified by the operator (10-fold for this project) and generalized errors are estimated from the results of the learning processes that are provided through a mean error and a confidence interval calculated at the 99% level. The same data that was used for training the NN was used here with the exception of two rows containing data value of 0,0,0,0. Thus the total number of patterns used for training was 408. NEFCLASS-J did not have problem in classifying data that were 0 (outside the GIS mask) as ‘0’ category representing not classified category.

Table 5: Parameters of Neuro-fuzzy classifier

Learning Rate	0.01
No. of fuzzy sets	4 TRAPEZOIDAL for each variable
Rule learning strategy	Best per class
Stop control	Maximum number of epochs=1000
	Minimum number if epochs= 100
Validation mode	10-fold cross validation

3.5 Sensitivity Analysis

Sensitivity of the training data set was analyzed by changing the training data. A total of 26 models were created using neural networks and neuro-fuzzy methods ((13 per method). All models were compared to DRASTIC model too. First set of simulations was run using D, R, S and I value as reflected by the weight of the parameters outlined by the DRASTIC model. For details on weight of the parameters please refer to Dixon et al., (2002). These models from here on will be referred to as NN1, NN2 and so on for neural netweoks models and NF1, NF2 and so on for neuro fuzzy based models (Table 6).

Table 6. Summary of Model Parameters and Model Names

Models	Neural networks	Neuro-fuzzy	Name of Model
Model-1	NN1	NF1	DRASTIC
Model-2	NN2	NF2	DRASTIC DH
Model-3	NN3	NF3	DRASTIC,DH,PED
Model-4	NN4	NF4	DRASTIC,DH,PED LULC
Model-5	NN5	NF5	DRSI
Model-6	NN6	NF6	DRSI DH
Model-7	NN7	NF7	DRSI DH PED
Model-8	NN8	NF8	DRSI DH PED LULC
Model-9	NN9	NF9	DH PED LULC
Model-10	NN10	NF10	PED_DH_BD
Model-11	NN11	NF11	PED_DH_BD_LULC
Model-12	NN12	NF12	pH_OM_BD
Model-13	NN13	NF13	DRASTIC, LULC
Model-14	NN14	NF14	DRSI, LULC

3.6 Coincidence reports

Once the vulnerability maps were generated using the above methods, field data were used to generate coincidence reports to evaluate their performances. Water quality data for E Coli for 55 wells were used in this study. The well water quality data we provided by FL Dept. of Environmental Protection (FLDEP). Sixteen out of 55 wells were contaminated with at least one occurrence ce of E Coli. It was assumes 1 E coli is an indicator of the presence of contamination movement pathways. A set of coincidence reports was generated between vulnerability maps and well contamination data to compare actual contamination with potential contamination (or vulnerability) generated by our models.

4. RESULTS AND DISCUSSION

4.1 Training of NN model

The Table 7 shows training results of the NN. Internally each category in the classification output is considered to be mutually exclusive and is assigned an output node in the neural net. The relative entropy measure ensures that the outputs of the NN enforce the mutual dependence of the outputs. It maximizes the probability of successful classification. Ideally a very low value of relative entropy indicates a good fit of the model to the data. The accuracy gives the fraction of records whose prediction is within a specified tolerance of the desired output. By default the accuracy tolerance is set to 20% of the range of the output.

Table 7. NN Models: Performance of Training Data.

Model name	Patterns	Training Set Accuracy	Test Set Accuracy	All Data Accuracy	Relative Entropy		
					Train	Test	All
DRASTIC DH	1733	0.86314	0.857692	0.861512	0.09899	0.108067	0.101714
DRASTIC DH PED	2675	0.873397	0.902864	0.882243	0.115352	0.125596	0.118427
DRASTIC DH PED LULC	7159	0.787867	0.799348	0.791312	0.143857	0.151916	0.146275
DRSI	55	1	0.945455	0.945455	0.003272	0.032348	0.032348
DRSI DH	502	0.766382	0.774834	0.768924	0.16442	0.16174	0.163614
DRSI DH PED	830	0.817241	0.8	0.812048	0.191069	0.188303	0.190236
DRSI DH PED LULC	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
DH PED LULC	160	0.873874	0.857143	0.86875	0.08823	0.121458	0.098406
PED_DH_BD	57	1	0.929825	0.929825	0.006575	0.016977	0.016977
PED_DH_BD_LULC	256	0.73743	0.688312	0.722656	0.206299	0.203072	0.205329
pH_OM_BD	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
DRASTIC, LULC	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
DRSI, LULC	187	0.938462	0.929825	0.935829	0.078415	0.077921	0.078264

4.2 Training Results of NF model

The Table 8 shows characteristics of training data used with the Neuro-fuzzy models. For the model NF1, the variable R showed maximum SD whereas for the models NF2 and NF3 variable D showed the maximum SD.

Table 8. Characteristics of Training Data for Neuro-fuzzy Models.

Variable	Name	mean	SD	minimum	Maximum	Missing
			NF 1			
Var 1	D	8.07	1.93	3	10	0
Var 2	R	3.6	2.79	1	10	0
Var 3	S	3.77	2.44	0	9	0
Var 4	I	7.09	1.82	3	10	0
			NF 2 and NF 3			
Var 1	D	16.92	14.90	3	63	0
Var 2	R	3.6	2.79	1	10	0

Var 3	S	3.77	2.44	0	9	0
Var 4	I	20.03	13.44	10	63	0

Table 9 shows performance of the training data used with Neruo-fuzzy models. NF2 showed highest number of misclassification as well as error. This error corresponds to the sum of squared differences between targets and outputs and is a measure of the ambiguity of classifications.

Table 9. Neuro-fuzzy Models: Performance of Training Data.

Name of Model	Patterns	Accuracy (%)	Misclassifications(%)	Error
DRASTIC DH	1733	62.84	37.16	1170.98
DRASTIC,DH,PED	2677	75.57	24.43	1233.6
DRASTIC,DH,PED LULC	7159	84.73	15.27	2397.67
DRSI	55	80	20	25
DRSI DH	502	58.96	41.04	394
DRSI DH PED	832	79.81	20.19	363.6
DRSI DH PED LULC	*NA*	*NA*	*NA*	*NA*
DH PED LULC	160	82.5	17.5	66.7
PED_DH_BD	57	71.93	28.07	24.45
PED_DH_BD_LULC	256	80	20	85
pH_OM_BD	*NA*	*NA*	*NA*	*NA*
DRASTIC, LULC	*NA*	*NA*	*NA*	*NA*
DRSI, LULC	183	90.16	9.84	41.1

4.3 Sensitivity Analysis of NN and NF

4.3.1 Vulnerability maps

The models described below were created in a GIS by incorporating the various permutation combination of 14 parameters D, R, A, S, T, I, C, BD, DH, LULC, Soils structure and pedality. The same training data sets were used for models are shown in (Tables 6 and 7). Spatial distribution of vulnerability category varied from models to models (Figures 3 and 4). Due to sheer number of we are going to summarize the most interesting simulations in the main document. Please find the comprehensive simulation results in the Appendix B. Table 10 shows summery of areal coverage for the selected models.

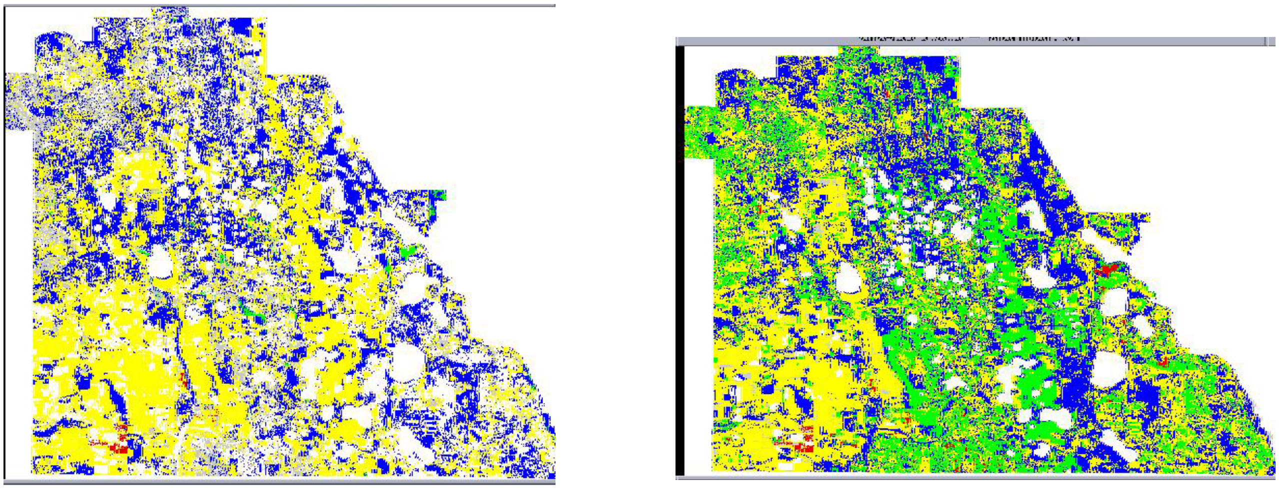


Figure 3. Spatial Distribution of vulnerability categories from the model using DH, ped, LULC parameters using NF (left); using NN(right)

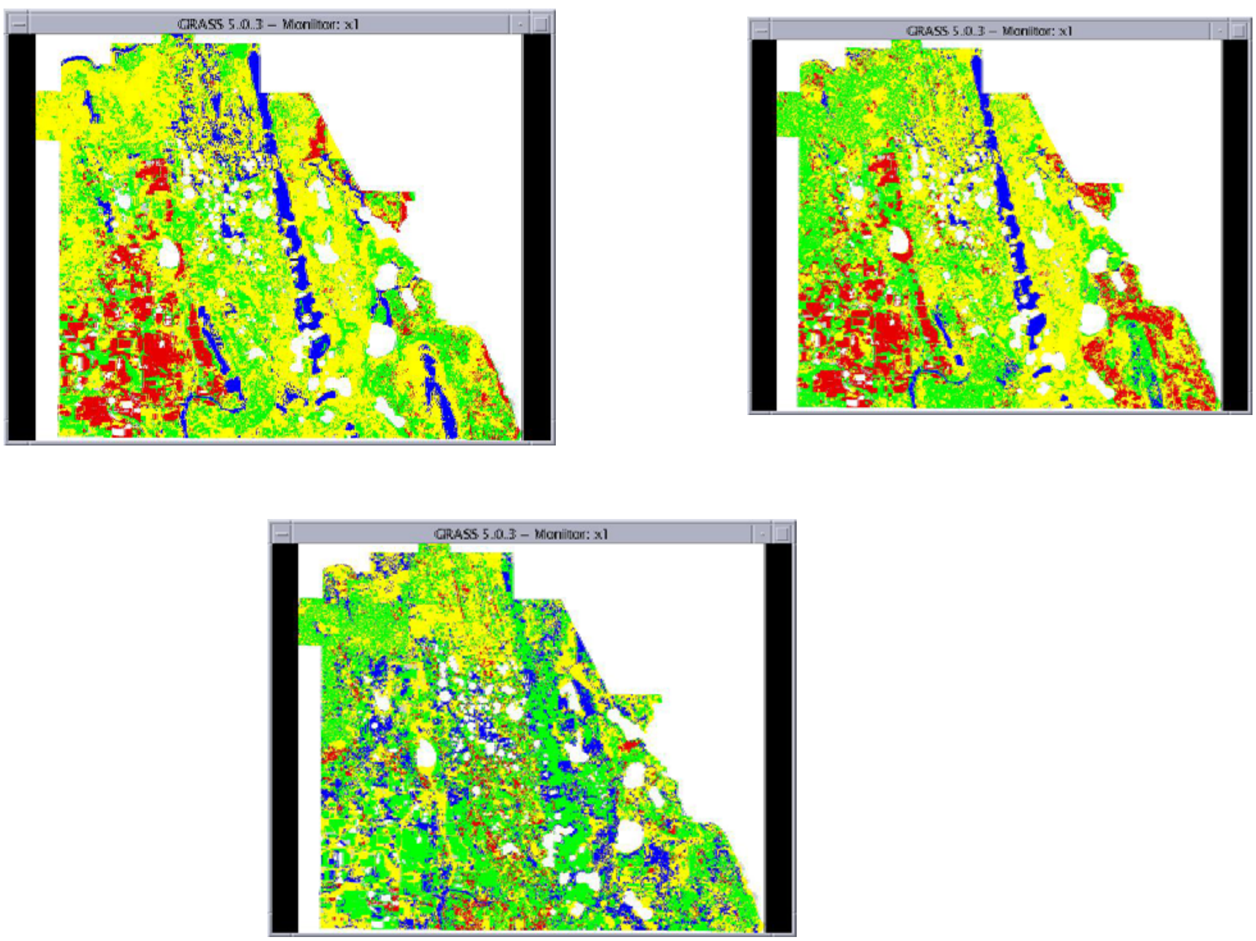


Figure 4. Spatial Distribution of Vulnerability from Neural networks models: (left NN2, right NN3, and bottom NN4). **Parameters and model names***: Drastic DH (NN2), DRASTIC DH, Ped (NN3), DRSTIC , DH, Ped, LULC (NN4)*

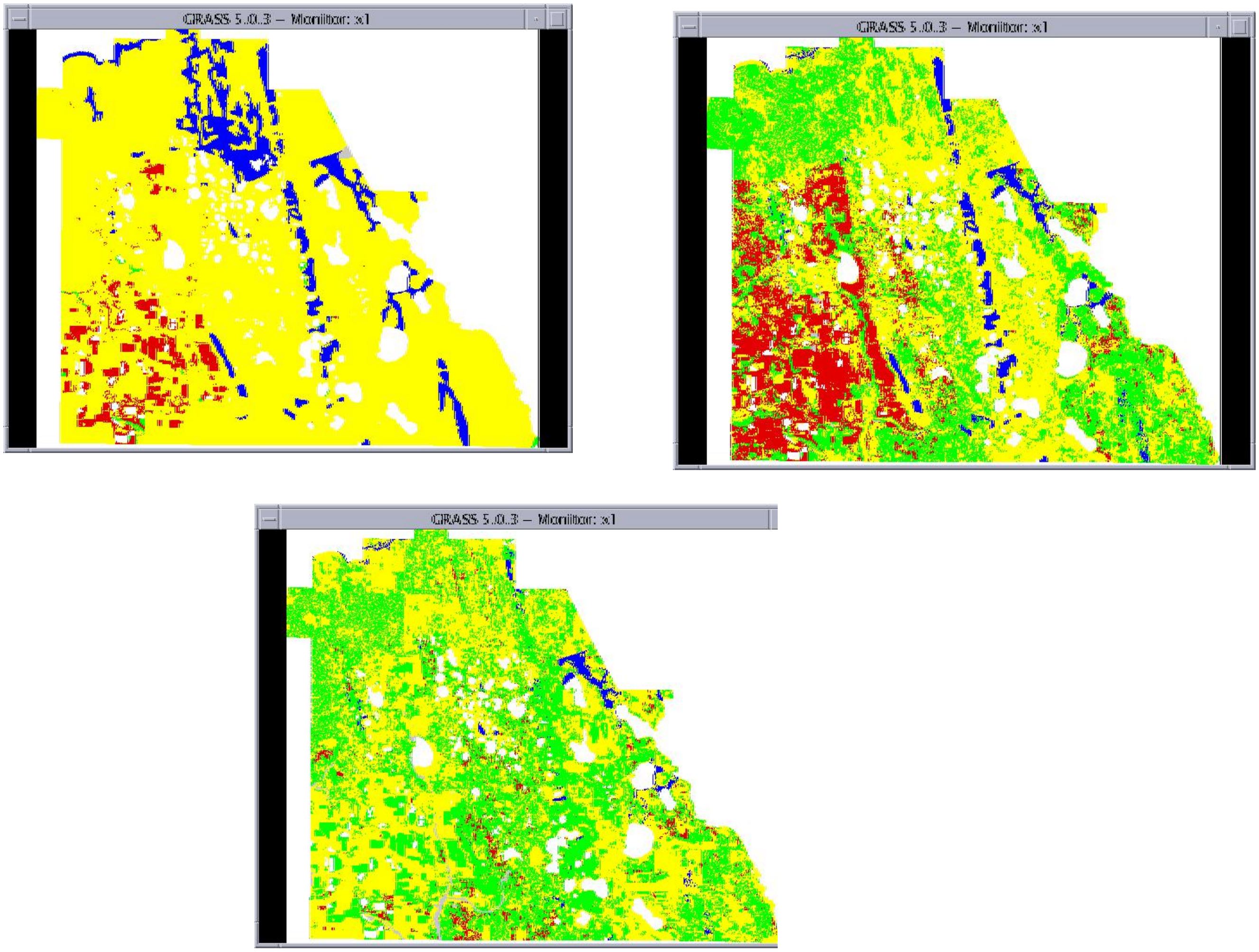


Figure 5. Spatial Distribution of Vulnerability from Neuro-fuzzy models: (left NF2, right NF3, and bottom NF4).). **Parameters and model names*: Drastic DH (NN2), DRASTIC DH, Ped (NN3), DRSTIC , DH, Ped, LULC (NN4)***

Table 10. Spatial Distribution of Vulnerability Categories in Percentage: Selected Models.

Vulnerability Categories	**Area Coverage by the Models (%)**					
	NN2	NF2	NN3	NF 3	NN4	NF4
Not classified/no data	2	0	2	0	2	1
Low	9	8	5	3	16	2
Moderate low	54	87	47	51	36	54
Moderate	24	0	30	29	42	41
High	11	5	15	16	5	2
Total	100	100	100	100	100	100

4.3.2 Coincidence Reports

Coincidence reports between vulnerability maps and well data were generated using GRASS raster files (Figures 6- 9). A total of 28 models were created using neural networks and neuro-fuzzy methods ((14 each). All models were compared to DRASTIC model too. NN2 performed reasonably good in predicting contaminated wells – it did not perform well while classifying non-contaminated wells. NN2 predicted 5 out of 7 contaminated wells as highly vulnerable category whereas 25 of the non-contaminated wells were classified as in the high category. A well performing NN or NF model should be able to classify contaminated wells in high category and non-contaminated wells in low or moderately low vulnerability category. The model NN5 (DRSI) predicted 8 contaminated wells as highly vulnerable and 1 well each in the low and moderately low vulnerability category, respectively. NN7 (DRSI,DH,ped) and model NN11 predicted equal number of contaminated wells in the highly vulnerable category. However, all these 3 models (NN5, NN7 and NN11) over predicted non-contaminated wells. The comparison of point data to a spatial data for accuracy assessment is not the ideal way. The coincidence analysis was performed to get an idea of how the models are performing in a relative sense. These values should not be used as absolute indicators of the suitability of models. Although the coincidence reports show similar trend for NF 7 and NN7, NF 5 showed drastically different results than NN5. In general, NF models showed higher uncertainty than the NN models while predicting contaminated wells.

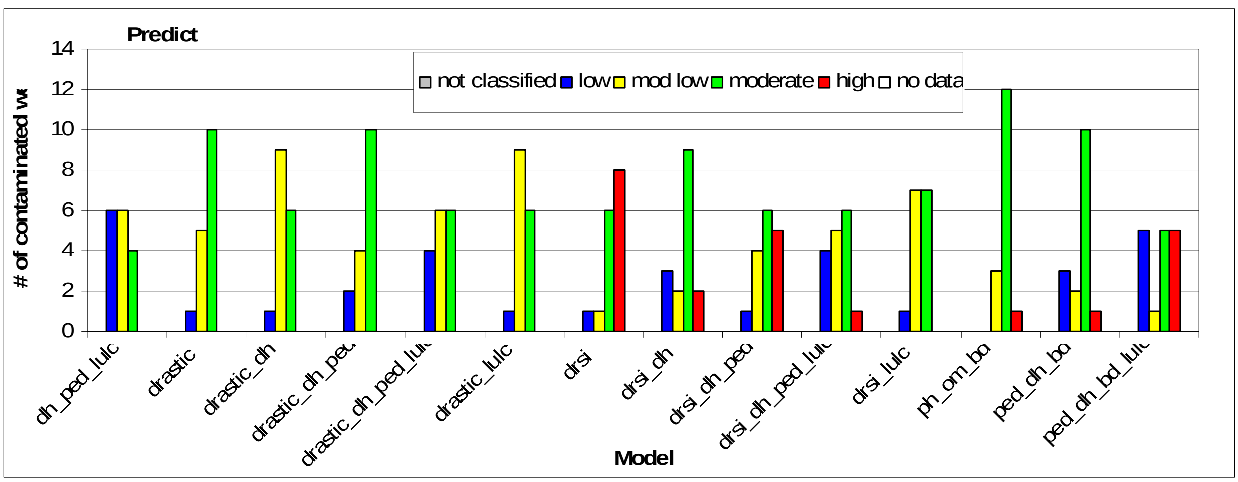


Figure 6. Coincidence report for NN models and well water quality data E Coli data

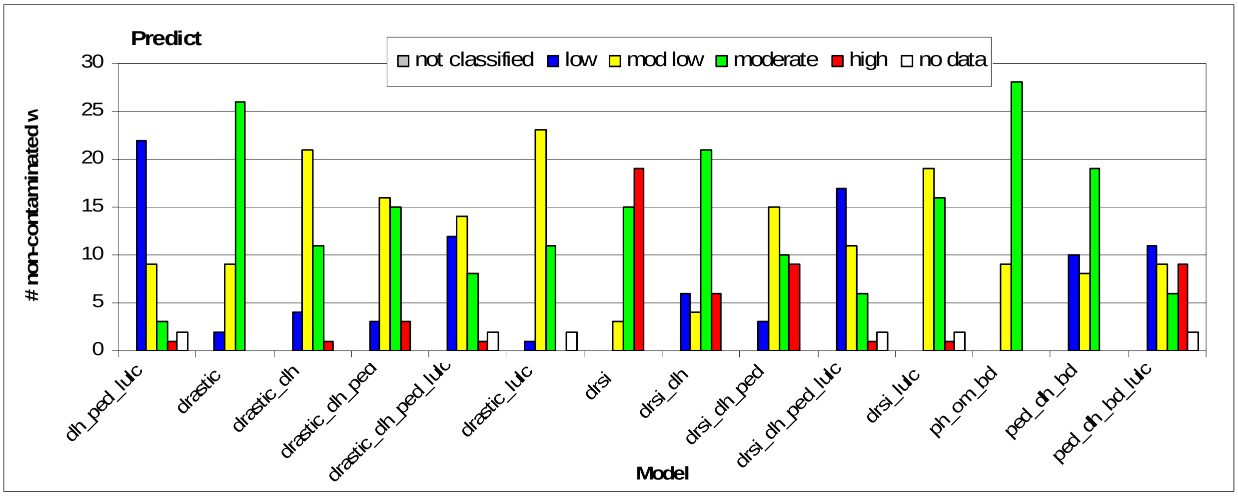


Figure 7. Coincidence report for NN models and non contaminated wells

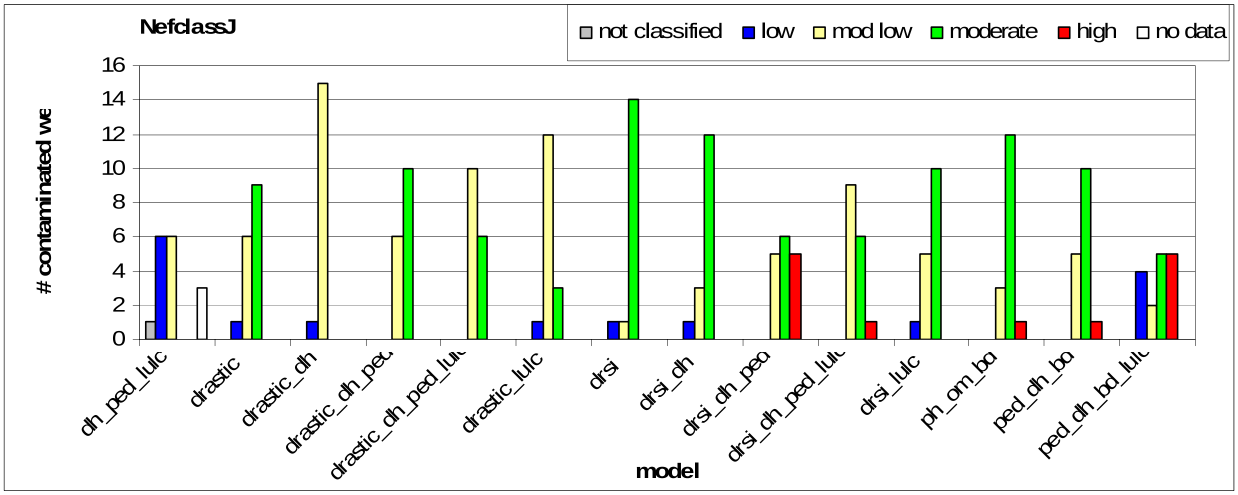


Figure 8. Coincidence between neuro-fuzzy models and contaminated wells (E Coli).

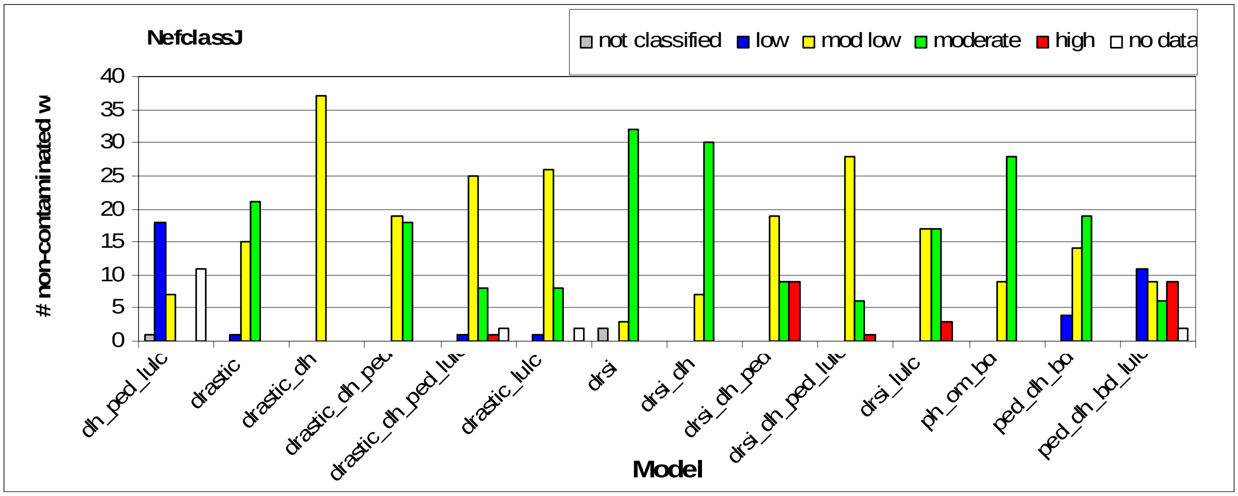


Figure 9. Coincidence report between neuro-fuzzy models and non contaminated wells.

5. CONCLUSION

Compared to Neuro-fuzzy models, NN models performed better with contradictory data. Coincidence reports with well water quality data did not yield conclusive results, nor should they be used as absolute indicators. Although for our project integration and use of NN software with GIS was cumbersome and time consuming – it predicted better. With the fine-tune of the contradictory data manually could yield similar vulnerability maps both from NN and Neuro-fuzzy models and improve well predictions. Further studies needed. A larger data set of well contamination may provide better results.

References

1.Aller, L. T., J.H. Bennett, R. Lehr, J. Petty, and G. Hackett. 1987. DRASTIC: A Standardized System for Evaluating Ground Water Pollution Potential Using Geo-Hydrogeologic Settings. U.S. Environmental Protection Agency Report. EPA600/2-87/036.

2.Burrough, P.A., R.A. MacMillan, and W. Deursen. 1992. Fuzzy classification methods for determining land suitability from soil profile observations. J. Soil Sci. 43:193-210.

3. Burrough, P. A. 1996. Opportunities and Limitations of GIS-based Modeling of Solute Transport at the Regional Scale. In *Application of GIS to the non-point source pollutants in the vadose zone*, SSSA, Special Publication # 48, D. Corning and K. Loague eds., Madison, WI.

4.Corwin, D. L, K. Loague and T. R. Ellsworth. 1996. Introduction to Non-Point Source Pollution in the Vadose Zone with Advanced Information Technologies. In: Assessment of Non-Point Source Pollution in the Vadose Zone, ed. D. L. Corwin, K. Loague and T. R. Eliisworth. Geophysical Monograph 108. AGS, Washington D.C

5.Dixon, B. "Application of neuro-fuzzy techniques to predict ground water vulnerability in NW Arkansas." Ph.D. diss., University of Arkansas, 2001.

6. Dixon, B., H.D. Scott, J.C. Dixon, and K.F. Steele. 2002. Prediction of Aquifer Vulnerability to Pesticides Using Fuzzy Rule-Based Models at the Regional Scale. Physical Geography 23:130 – 152.
7. De Gruijter, J. J, D. J. J. Walvoort, and P. F. M. Van Gaans. 1997. Continuos soil maps - a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma. 77:169-195.
8. Khan, E. 1999. Neural fuzzy based intelligent systems and applications. In *Fusion of neural networks, fuzzy sets, and genetic algorithms: Industrial applications*, edited by L.C. Jain and N.M. Martin. CRC Press, Washington, D.C. p. 331.
9. Nauck, U. and R. Kruse. 1999. Design and Implementation of a neuro-fuzzy data analysis tool in JAVA. Manual. Technical University of Braunschweig. Germany.
10. Smith, P. A., Scott, H.D., and Fugitt, T. (1994). Influence of geographic database scale on prediction of groundwater vulnerability to pesticides. *Journal of Soil Contamination*, Vol. 3, 285-298.
11. Wang, F., G. Hall and Subaryono. 1990. Fuzzy information representation and processing in conventional GIS software: database design and application, Int. J. GIS. 4:261-283.