



**PREDICTION OF GROUND WATER VULNERABILITY
USING AN INTEGRATED GIS-BASED NEURO-FUZZY TECHNIQUES**

B. DIXON¹

¹ Environmental Science, Policy and Geography, 140 seventh ave south, St. Petersburg, Fl 33701, 727 553 4025,
bdixon@stpt.usf.edu

ABSTRACT

There is a need to develop new modeling techniques that assess ground water vulnerability with less expensive data and which are robust when data are uncertain and incomplete. Incorporation of Geographic Information Systems (GIS) with a modeling approach that is robust has the potential for creating a successful modeling tool. The specific objective of this study was to develop a model using Neuro-fuzzy techniques in a GIS to predict ground water vulnerability. The Neuro-fuzzy model was developed in JAVA using four plausible parameters deemed critical in transporting contaminants in and through the soil profile. These parameters include soil hydrologic group, depth of the soil profile, soil structure (pedality points) of the soil A horizon and landuse. The model was validated using nitrate-N concentration data. The majority of the highly vulnerable areas predicted by the model coincided with agricultural landuse, moderately deep to deep soils, soil hydrologic group C (moderately low Ksat) and high pedality points (high water transmitting properties of the soil structure). The proposed methodology has potential for facilitating ground water vulnerability modeling at a regional scale and can be used for other regions, but would require incorporation of appropriate input parameters suitable for the region. This study is the first step toward incorporation of Neuro-fuzzy techniques, GIS, GPS and remote sensing in the assessment of ground water vulnerability from non-point source contaminants.

Key Words: GIS, Spatial Modeling, Remote Sensing, Fuzzy Logic, Neural Networks

INTRODUCTION

Contamination of ground water has become a major concern of local, state and federal agencies involved with the management of water quality and quantity and their relationship to human health. Delineation of vulnerable areas and selective applications of agricultural chemicals in those areas can minimize contamination of ground water. However, assessment of ground water vulnerability or delineation of monitoring zones is not easy because contamination depends upon numerous, complexly interacting parameters. Uncertainty is inherent in all methods of assessing ground water vulnerability and arises from errors in obtaining data, the natural spatial and temporal variability of the hydrogeologic parameters in the field, and in the numerical approximation and computerization (National Research Council, 1993).

Existing ground water vulnerability assessment methods may be grouped into three categories: overlay and index, statistical, and process-based simulation models (National Research Council, 1993). Overlay and index methods have been developed because of limitations in process-based models and lack of monitoring data required for statistical methods (National Research Council, 1993). Advent of Geographic Information System (GIS) facilitated adoption of this modeling approach to watershed and regional scales. Despite its common use at the regional scale, overlay and index methods do not have inherent mechanisms to deal with uncertainties, nor do these models consider landuse and landcover (LULC) and management. Most models based on overlay and index methods use physiographic parameters and do not consider anthropogenic aspects of vulnerability. Therefore, there is a need to develop a methodology that can extract information from imprecise data.

Corwin, et.al., (1996) suggested that an integrated system of advanced information technologies such as Global Positioning Systems (GPS), GIS, geostatistics, remote sensing,

solute transport modeling, neural networks (NN), Fuzzy Logic, and uncertainty analysis could provide a framework from which real-time or simulated assessment of non-point source (NPS) pollution can be made. Burrough (1996) suggested that there are potential benefits in GIS-based modeling of solute transport at the regional scale He also stated that when appropriate interfaces are available, GIS-based approaches can help model two-, three-, and four dimensional situations, sensitivity analysis and error propagation studies. The results seen in terms of spatial context will enhance greater understanding of the modeling problem (Burrough, 1996). A Neuro-fuzzy system is a fuzzy system that is trained by a learning algorithm from NN theory. Neuro-fuzzy modeling is an approach where the fusion of NN and Fuzzy Logic find their greatest strengths. These two techniques complement each other. This approach employs heuristic learning strategies derived from the domain of NN theory to support the development of a fuzzy system. It is possible to completely map NN knowledge to Fuzzy Logic (Khan, 1999).

GOALS AND OBJECTIVES

The specific objective of this research was to develop a modeling approach that loosely couples Neuro-fuzzy techniques and GIS to predict ground water vulnerability in a relatively large watershed in northwest Arkansas having mixed LULC, and variably permeable soils over the karstified Boone Formation.

SIGNIFICANCE

This research used GIS, GPS, remote sensing, and a fusion of NN and fuzzy logic techniques along with relevant interactions of soil properties and LULC on ground water quality of watersheds in a karst region. The Neuro-fuzzy models developed in a GIS have the inherent

capability to deal with uncertainties in the data, tolerate imprecision, and can extract information from incomplete datasets. Expert knowledge, which is a valuable source of information on the physical, chemical and biological parameters that are hard to measure, as well as experimental information were also incorporated into modeling.

Simple and readily available, but meaningful, parameters were used in the modeling to ensure global application of the models especially for environmental policy development. The model was developed using relevant soil properties and LULC as input data. In the selection of parameters it was assumed that since the underlying geology is the Boone Formation, which is highly fractured, the variability of water and contaminant transmitting properties of soils as well as attenuation processes of the overlying soils govern the vulnerability of the ground water. Once the contaminant moves below the soil zone, it will eventually reach the ground water due to the ubiquitous presence of fractures, low consumption capabilities and the considerable amounts of water flowing vertically in this humid region (Al-Rashidy, 1999). Therefore, only soil and LULC related parameters were used in this research.

Application of Neuro-fuzzy techniques to the prediction of ground water vulnerability does not provide exact solutions. The output from the Neuro-fuzzy model was displayed in the form of a map that shows regions of ground water in the watershed having more or less potential vulnerability to $\text{NO}_3\text{-N}$ contamination. In addition, a table was developed to present the areal extent of the vulnerability categories. This article discusses a methodology to integrate a Neuro-fuzzy technique in a GIS to predict ground water vulnerability.

The proposed methodology has the potential to facilitate the modeling of ground water vulnerability at a regional scale. Methodologies employed in this project are applicable, and readily transferable, to other watersheds with different physiographic settings to delineate ground

water vulnerability. However, this approach would require incorporation of appropriate input parameters suitable for the region. For example, if the geology of an area is different from the study area, geological factors should be incorporated to account for potential resistance to water and contaminants transport processes. This study is a first step toward incorporation of Neuro-fuzzy techniques in a GIS and would require modifications for wider ranges of applications. For a more detailed discussion of the results and sensitivity of the models to training parameters and scale issues, the reader is referred to the work reported by Dixon (2001).

STUDY AREA

Location

The Neuro-fuzzy models were developed based on the characteristics of the four sub-basins of the Illinois River Watershed including the Savoy Experimental Watershed (SEW). The study area, which is located east of the Arkansas-Oklahoma border (Figure 1), is an intensely monitored watershed in northwest Arkansas. The SEW is a University of Arkansas (U of A)

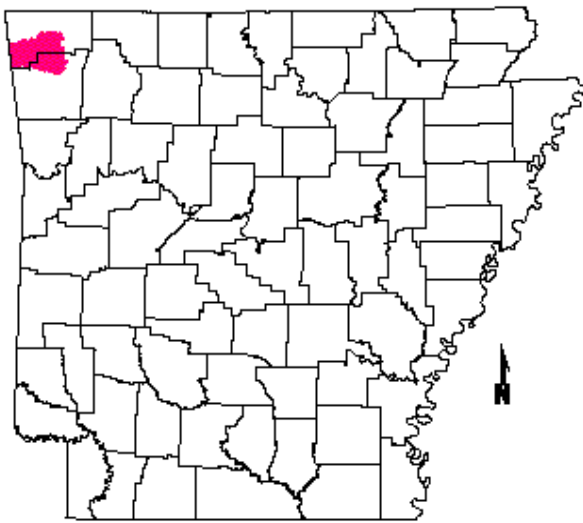


Figure 1. Location of the study area in northwest Arkansas, USA.

property of approximately 1250 ha located in the Illinois River watershed, 24 km west of the U of A campus in Fayetteville, AR. The study area has an area of about 109,000 ha (270,000 acres) and is characterized by 10 dominant soil series with Clarksville and Nixa soils occupying about 16% and 21% of the study area, respectively. The major LULC of the study area is forests (64%) and

agriculture, primarily tall fescue and bermuda pasture (23%).

The Boone aquifer, which underlies the study area in northwest Arkansas, has been shown to have higher nitrate-N ($\text{NO}_3\text{-N}$) concentrations than the national median (Peterson, et. al., 1998). The dominant landuse (LULC) of this area is agriculture (primarily pasture/cattle and woodlands) and an encroaching urbanization. The major sources of nitrogen in the study area are poultry/cattle wastes, inorganic fertilizers (Petersen et. al., 1998) and septic filter fields. Many of the soils in the Ozark Region are highly permeable and well drained and the geology is limestone interbedded with chert.

METHODOLOGY

Development of the Model Inputs

This study used several primary and secondary digital databases. The primary digital databases were obtained from numerous sources and in various formats (Table 1). Watershed boundaries were used to delineate the study area, however, they were not used as model inputs. The location of wells and springs and accompanying water quality data were used to assess the performance of the models. Geology and slope parameters were not used as model inputs, they were used to cross check the vulnerability zones and fine-tune the predictions. Elevation data were used to generate slopes for the watershed. The GIS software used in this study were GRASS 4.2 and ArcView 3.2.

The primary digital data layers were (i) watershed boundaries, (ii) location of wells and springs, (iii) water quality, (v) geology, (vi) soils, (vi) LULC, and (vii) digital elevation models (DEMs).

The primary data layers were manipulated in a GIS to generate secondary data layers. The secondary data layers were used in the models either as inputs to the models or were used to fine-tune the rule bases (Table 2). The inputs for the Neuro-fuzzy models were (i) soil hydrologic group, (ii) depth of the soil horizons, (iii) soil structure of the A horizon and (iv) LULC.

Table 1. Description of primary data layers.

Primary Data layers	Source	Scale/resolution	Comments
Watershed boundaries	NRCS	1:100,000	Digital
Location of Springs/wells	Field determined	N. A.	GPS & AWRC Publication
Water Quality data	Collected at SEW and surroundings	N. A.	ADEQ Lab. and AWRC publication
Geology	Arkansas Geological Commission	1:24,000	Digital
Soils	NRCS and Iowa State	1:24,000	Mylar for primary and Tabular for secondary attributes
LULC	Oklahoma State University (1985) and NRCS (1996)	1:24,000 1 m	Mylar Digital
DEMs	USGS	30 m	Digital

- *NRCS: Natural Resources Conservation Services*
- *AWRC: Arkansas Water Resources Center*
- *ADEQ: Arkansas Dept. of Environmental Quality*
- *USGS: U.S. Geological Survey*

Table 2. Primary and secondary data layers and their use in the research.

Primary data	Secondary data	Model Inputs	Validation	Fine-Tune
Soils	Hydrologic groups	4		
	Structure	4		
	Depth of the profile	4		
LULC	LULC	4		
DEMs	Slope			4
Location of wells/springs			4	
Water quality			4	

1. Watershed Boundaries

The watershed boundaries were used to delineate the study area. This data layer, which was not used in the modeling processes, was provided by Natural Resources Conservation Service (NRCS) in a digital format. The digital data were available in a Digital Line Graph – 3 (DLG-3) format, which is readily compatible with the GIS software used for the research.

2. Spring and Well Data

Locations and names of springs and wells were obtained through field inventory with a GPS and from the Arkansas Water Resources Center (AWRC) report (Smith and Steele, 1990). The GPS data were in latitude and longitude format. The location data were converted into UTM format using the GRASS 4.2 command `m.l12u`. The UTM coordinate file was brought into Microsoft Excel 97 and saved as .csv files. Then the ArcView command ‘`gps2shape`’ was used on the file names ‘`mapdata.csv`’. The AWRC report (Smith and Steele, 1990) provided well location in decimal degrees that were converted into degree:minutes:seconds (d:m:s) format using MSEXcel 1997. Once the data were converted into d:m:s format, the steps described for GPS data were used with the AWRC data set.

3. Ground Water Quality

The ground water quality data were obtained from two sources: Arkansas Department of Environmental Quality (ADEQ) Laboratory (Tim Kresse, ADEQ, written Commun. 2000) and AWRC publications. The water quality data provided by ADEQ were collected with respect to storm events during 1998 and 1999 for 24 different wells and springs. These data were analyzed by ADEQ Laboratory personnel for about 40 different ions and compounds. In this study, $\text{NO}_3\text{-N}$ was the water quality parameter used to compare the validity of the models because application

of animal wastes to pasture is a routine management practice and this compound readily leaches to ground water. The discharge records and concentration level of $\text{NO}_3\text{-N}$ data for springs and only concentration data for wells were stored in a relational table. In addition, AWRC provided historical data consisting of 20 wells (Smith and Steele, 1990) for the study area. These wells were sampled during the wet season of 1990 and analyzed for $\text{NO}_3\text{-N}$ and other ions and compounds in the AWRC water quality laboratory. The inclusion of historical ground water quality data added temporal variability as well as uncertainty in the data. However, inclusion of historical data (Smith and Steele, 1990) added spatial variability to the data set as they reduced the clustering of the wells of the ADEQ data set.

4. Soils

The Soil Survey Geographic Database (SSURGO) soils maps (1:24,000) were obtained from NRCS and digitized at the Soil Physics Laboratory of U of A (Mitra, et al., 1997). Tabular data for the hydrologic groups were obtained from SSURGO database for soil map units. Soil map units were reclassified into appropriate categories to generate maps of hydrologic groups. The data layers for soil map units were then reclassified into a soil series level map. This step was necessary because SSURGO level data do not contain information on depth of the profile and soil structure required by the Neuro-fuzzy models. Soils data for depth of the profile and soil structure were obtained from the Official Soil Series Description database of Iowa State (<http://www.statlab.iastate.edu/soils/nsdaf/>; viewed 6/16/00). The soil series map was reclassified to generate maps for soil structure and depth of the profile. Soil structure, specifically pedality, were classified according to the scheme developed by (Lin et, al. 1999) to indicate water transmitting properties of the soils. In soils with compound pedality, weighted averages were used for the horizon. The depth of the profile was estimated by excluding Cr and

R horizons from the published soil descriptions. The GRASS (4.2) command 'r.reclass' were used for all reclassification routines.

5. LULC

In this study, two sets of LULC data were used (i) LULC data for 1985 provided by Oklahoma State University. These maps were originally developed by the Lockheed Corporation from 1:24,000 scale aerial photographs. Interpretation of LULC in line formats data were copied from aerial photographs to acetate maps and scanned in the Soil Physics Laboratory at U of A. (ii) The other set of LULC data were obtained from NRCS-Washington County Office. This map was for SEW and its surroundings and shows types of pasture, i.e. whether it is bermuda or tall fescue pastures. The field boundaries were drawn on a Digital Orthophoto Quarter Quads (DOQQ) and ground truthing was done in 1996 to complete the map. In this map, the fields were numbered and tabular data was provided to associate attribute data with the field numbers. Tabular data provided records on acreage of the field and type of pasture.

Development of Neuro-Fuzzy Models and Integrating Models in a GIS Platform.

The Neuro-fuzzy software NEFCLASS-J (NEuro Fuzzy CLASSfier) for the JAVA platform was used (Nauck and Kruse, 1999). A NEFCLASS-J is a three layer fuzzy perceptron. NEFCLASS-J uses pattern vectors $x = (x_1, \dots, x_n) \in R^n$ and class C is a subset of R^n (Figure 2). It assumes that intersections between two different classes are empty. In this study, a supervised learning algorithm based on fuzzy error backpropagation was used. The fuzzy sets and the linguistic rules, which perform this approximation and define the resulting NEFCLASS-J systems, were obtained from a set of examples provided in the training data sets. Since the NEFCLASS-J is written in JAVA, the output function was customized in JAVA to tie the model output in a GIS. The NEFCLASS-J offers learning algorithms to create structure (rule base) and

parameters (fuzzy sets) of a fuzzy classifier from data. Therefore, the learning algorithm uses constraints which can be selected according to the requirements of the application data.

The Neuro-fuzzy techniques were used in the research as a two step process. First, the training data set was used with the Neuro-fuzzy software to generate a classifier, fuzzy sets and rule bases. Second, once the software was trained, the application data set was used with the Neuro-fuzzy models. The outputs from the models were used to generate ground water vulnerability maps. The steps involved to generate a classifier were: 1) train data with 51% of all cases, 2) determine optimal consequents, 3) selection of consequents is complete, 4) writing the rules to

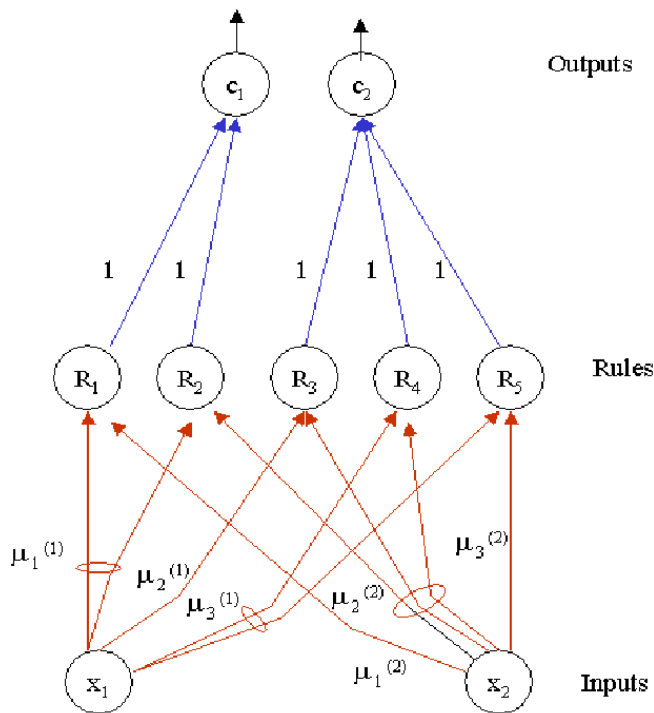


Figure 2. Schematics of the Neuro-fuzzy model used in this study,

the log file, 5) obey the learning rule criteria, 6) trimming the rule base to generate final rule base to use with the application data, 7) validate the data and 8) report training session.

NEFCLASS-J was used with trapezoidal membership functions. This parameterized membership function, which was chosen because of its simplicity, not only reduces system design time, but can facilitate the automated tuning of the system with

desired changes (Yen and Langari, 1998). The changes of the membership function for the trapezoidal shape can be obtained from the widening or narrowing of the membership function and the corresponding changes of the related parameters (Yen and Langari, 1998). This simple

membership function facilitated an important principle underlying the theories of Fuzzy Logic: exploring cost effective approximate solutions. The trapezoid membership function required four parameters a, b, c and d and the peak of the membership function is 1. Mathematically the function can be defined as:

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

The training data sets were obtained for the entire watershed from the GIS software GRASS using the command r.stats. Four input parameters to the command r.stats included the data sets used in the Neuro-fuzzy model, viz. soil hydrologic group, LULC, depth of the profile and structure (pedality points) of the soils as it indicates water transmitting properties. This GRASS command generated a table representing all of the possible combinations of input parameters found in the watershed. 202 patterns were identified. The output table from GRASS was imported in dBASE (IV) and the data were classified based on expert's opinion. The format of the input dBASE table is presented in Table 3. The first 4 columns represented input parameters such as hydrologic groups, LULC, depth of the soil horizon and soil structure. The remaining 4 columns indicated the classifier. All of the training data sets were reclassified according to the high, moderately high, moderate and low potential for ground water vulnerability.

Table 3. Example of the classifier for potential vulnerability categories used in the training data sets.

Hydrologic Group	LULC	Depth*	Soil Structure**	High	Moderately High	Moderate	Low
10	10	15	38	1	0	0	0
10	10	18	38	1	0	0	0
10	10	36	53	1	0	0	0
10	10	60	38	0	1	0	0
10	10	66	34	0	1	0	0
10	10	72	20	0	0	1	0
10	10	72	38	0	1	0	0
10	10	78	34	0	1	0	0
10	10	97	34	0	0	1	0
10	50	72	20	0	0	0	1
10	50	78	34	0	0	0	1

*depth in inches

** pedality points

Integration of Databases, and Neuro-Fuzzy Models in a GIS Platform

All of the relevant (soils, LULC and water quality) primary and reclassified secondary data were stored in dBASE IV. These data had a relational join with spatial data that facilitates graphical display on the ArcView GIS platform. The JAVA programming language was used to integrate databases with Neuro-Fuzzy models on the GIS platform. A custom program was written to export output from NEFCCLASS-J, the Neuro-fuzzy software, to the GIS software, GRASS, to generate maps.

Comparison of Neuro-fuzzy Models with Field Data

This project emphasized the likelihood of a location being classified as contaminated rather than focusing on accurately estimating NO₃-N concentrations in the ground water. Therefore, comparison of predictions between Neuro-fuzzy models with field water quality data was performed with respect to concentration classes. The NO₃-N concentration levels (mg/l) for the wells and springs were classified into four categories: low (< 0.5 mg/l), moderate (0.5 – 3 mg/l), moderately high (3 – 10mg/l) and high (> 10 mg/l). These categories were chosen to

indicate no anthropogenic input (<0.5 mg/l), low anthropogenic input (0.5 – 3 mg/l), significant anthropogenic input (3 – 10 mg/l) and above Maximum Contamination Level or MCL (10 mg/l).

The Neuro-fuzzy model predictions were compared with the water quality data sets (field data) to obtain information on relative suitability of the modeling techniques for predicting ground water vulnerability in the watershed. It should be noted that due to the point nature of the water quality data and inherent spatial and temporal variability associated with the water quality data, a comparison of well and spring data (point) and vulnerability maps (spatial) is not suitable for determining the best modeling approach in an absolute sense.

Three sets of coincidence analyses were performed between (i) Neuro-fuzzy model and input data, (ii) Neuro-fuzzy model and slope and geology of the area, and (iii) Neuro-fuzzy model prediction and well/spring concentration data.

RESULTS AND DISCUSSION

Spatial Characteristics of Primary Data layers

Soils

The study area contains 44 soil series. The two most dominant soil series are Nixa and Captina. These soils occupy about 21% and 18% of the study area, respectively (Table 4). Nixa soils are found mainly in the north and northeast part of the watershed while Captina soils are found predominantly in the eastern part (Figure 3). Small patches of Captina soils are also found in the west. Clarksville soils comprise about 16% of the area and are found in the central part of the watershed. Peridge soils comprise about 2% of the study area and are found mainly along the streams valleys. A few patches of Peridge soil also occur in the eastern part of the watershed.

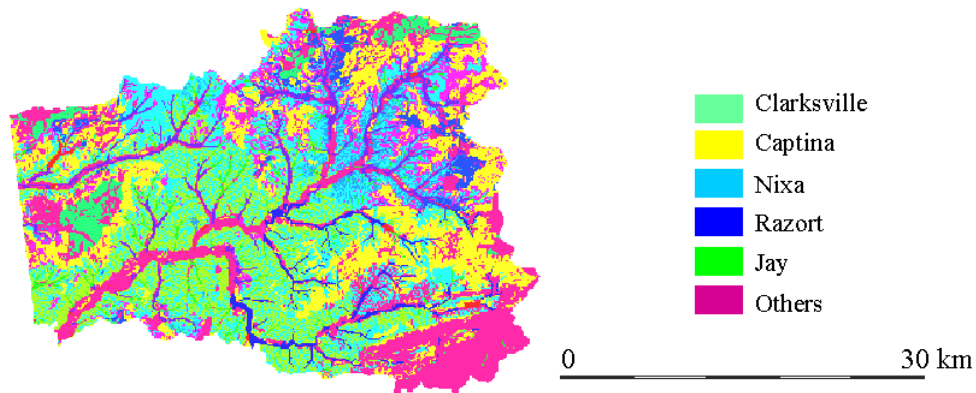


Figure 3. Spatial distribution of major soil series in the watershed

Table 4. Areal distribution of the major soils in the watershed.

Major Series	ha	%	Hydrologic Group	Pedality classes
Captina	20,107	18.4	C	Moderately high
Clarksville	17,363	15.9	B	Very high
Elsah	2,502	2.3	B	Moderately high
Jay	3,321	3.1	C	High
Nixa	22,856	20.9	C	High
Noark	2,859	2.6	B	Moderately high
Peridge	2,599	2.4	B	Low
Razort	2,440	2.2	B	Moderately high
Secesh	4,580	4.2	B	High
Tonti	7,596	6.9	C	Moderately High
Other soil series	22,565	20.7	B,C,D	N.A.
Water	480	0.4	N.A.	N.A.
Total	109,268	100		

LULC

The watershed is characterized by mixed LULC. Agriculture, particularly tall fescue and bermuda grass pasture, covers about 64% of the study area (Table 5). About 23% of the study area is covered by forests that are found in the central part of the watershed. Urban LULC, which covers about 10% of the study area, is found mainly in eastern part of the study area (Figure 4).

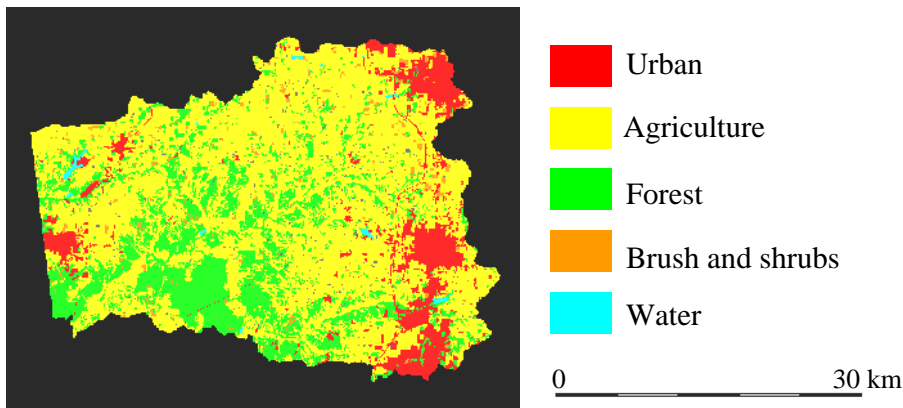


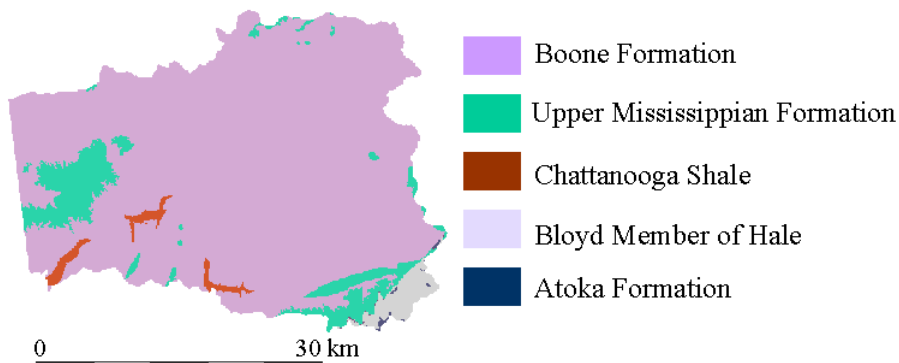
Figure 4. Spatial distribution of LULC in the watershed

Table 6. Areal distribution of geology in the watershed.

Geology	ha	%
Atoka Formation	161	0.1
Bloyd Member of the Hale	2,350	2.2
Cane Hill Member of Hale	46	0
Upper Mississippian Formation	9,224	8.4
Boone Formation	95,976	87.8
Chattanooga Shale	1,511	1.5
Total	109,268	100

Geology

The major rock unit, which occupies 88% of the study area, is the Mississippian age



Boone Formation (Figure 5). The Boone Formation is a limestone with varying amounts of densely interbedded chert ranging between 30

Figure 5. Spatial distribution of geology in the watershed

and 60% by volume. In NW Arkansas, the Boone Formation typically occurs between 300 – 350 feet (91 – 106 m) (Croneis, 1930). The Upper Mississippian Formation, which is a combination of Pitkin limestone, Fayetteville shale and the Batesville sandstone, occupies about 8% of the study area (Table 6). This geological formation is found in patches all over the study area (Figure 5). The primary porosity of the Boone Formation is low but the secondary porosity of this formation is high due to the presence of numerous fractures (Curtis, 2000; Chitsazan, 1980; Razaie, 1979).

Table 6. Areal distribution of geology in the watershed.

Geology	ha	%
Atoka Formation	161	0.1
Bloyd Member of the Hale	2,350	2.2
Cane Hill Member of Hale	46	0
Upper Mississippian Formation	9,224	8.4
Boone Formation	95,976	87.8
Chattanooga Shale	1,511	1.5
Total	109,268	100

Slopes

The slopes in the watershed vary from 0 to > 31 degrees. The slopes were classified according to the scheme of Hays (1995). About 39% of the study area has slopes that are classed as nearly level (0 – 2 degrees) and are found further away from the stream beds (Figure 6).

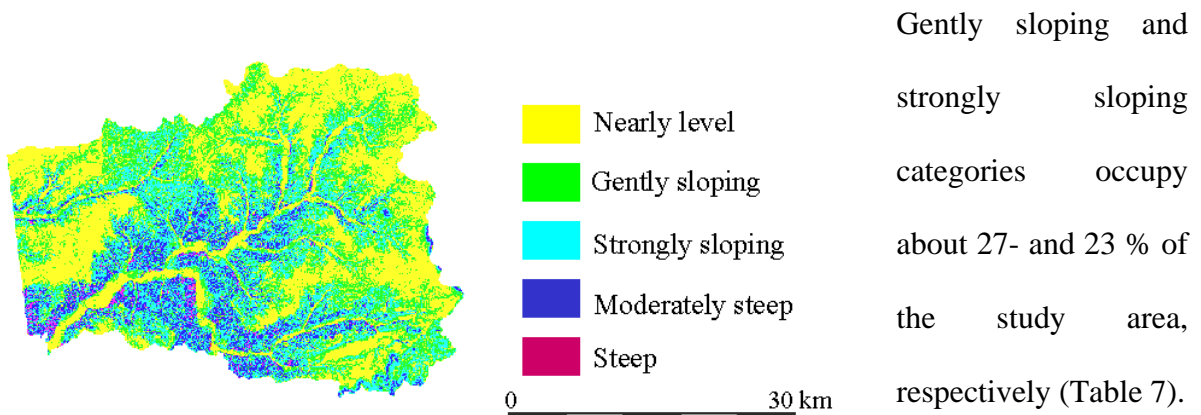


Figure 6. Spatial distribution of slope in the watershed

Table 7. Areal distribution of slopes in the watershed.

Slopes (degrees)	ha	%
Nearly level (0 - 2)	42,146	38.6
Gently sloping (3 - 4)	30,006	27.5
Strongly sloping (5 - 9)	25,226	23.1
Moderately steep (10 - 16)	10,208	9.3
Steep (17 - 30)	1,659	1.5
Very steep (> 31)	23	0
Total	109,268	100

Spatial Characteristics of Model Inputs

Soil Hydrologic Groups

The soils in the watershed were classified into three soil hydrologic groups B, C and D (Figure 7). About 54% of the land area in the watershed was in soil hydrologic group C (Table 8). Hydrologic group C indicates that Ksat is moderately low and internal free water occurrence is deeper than shallow. Hydrologic group B covers about 40% of the watershed and occurs along the stream valleys. Soil hydrologic group B indicates Ksat is moderately high and free water occurrence is deep or very deep. Hydrologic group D, which occurred in small patches across the watershed and indicates low Ksat values, occupies slightly more than 5% of the land area (Soil Division Staff, 1993).

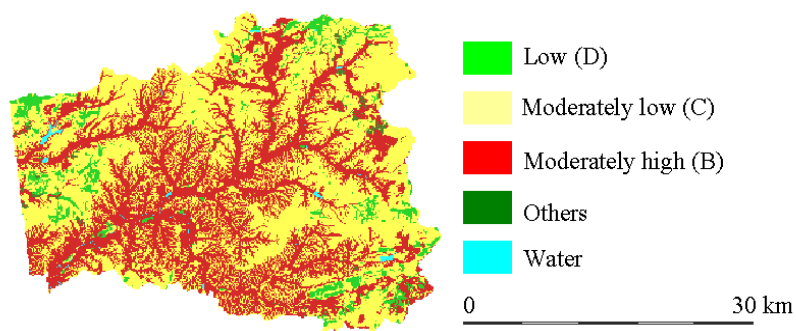


Figure 7. Spatial distribution of soil hydrologic groups in the watershed

Table 8. Areal distribution of soil hydrologic groups in the watershed.

Soil Hydrologic Groups	ha	%
Low (D)	5,713	5.2
Moderately low (C)	59,065	54
Moderately high (B)	43,249	39.6
Others	761	0.7
Water	480	0.5
Total	109,268	100

Depth of the Soil Profile

The depth of the soil profiles was estimated from the soil series description for the solum thickness excluding the Cr and R horizons. About 83% of the study area has deep or very deep soil profiles (Table 9). Deep soil profiles are found all over the watershed whereas very deep soils occur along the stream valleys (Figure 8). Moderately deep soils comprise 15% of the study area and also occur along the stream valleys. Moderately shallow soils (31 – 50 inches) are found in small patches across the watershed and occupies about 1% of the study area.

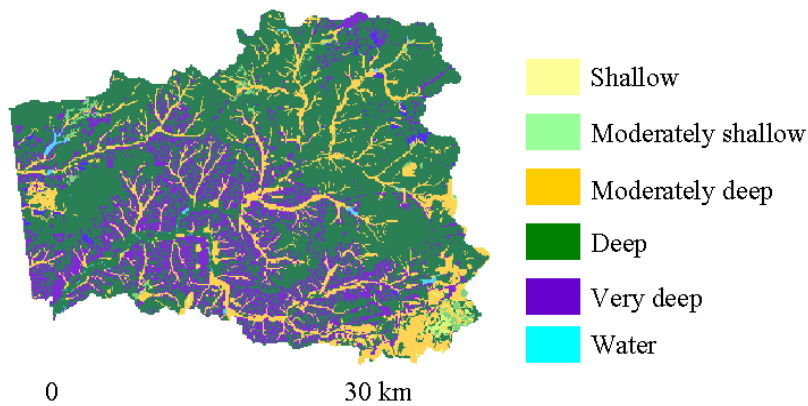


Figure 8. Spatial distribution of depth of the soil profile in the watershed

Table 9. Areal distribution of depth of the soil profile in the watershed.

Depth (inches)	ha	%
Shallow (9 - 30)	494	0.5
Moderately shallow (31 - 50)	1,180	1
Moderately deep (55 - 69)	15,927	15
Deep (70 – 85)	68,025	62
Very deep (> 85)	22401	21
Others	761	0.1
Water	480	0.4
Total	109,268	100

Soil Structure (pedality)

The structure (pedality) of the soil varies within the soil profile. For the Neuro-fuzzy model of the watershed, only the structural properties or pedality of the surface horizon (A) was used. The soil structure of the A horizon was reclassified according to the pedality points outlined by Lin et. al, (1999). For each soils the final pedality points were obtained by adding all of the points for ped size, ped shape and ped grade of the central concept of each soil series.

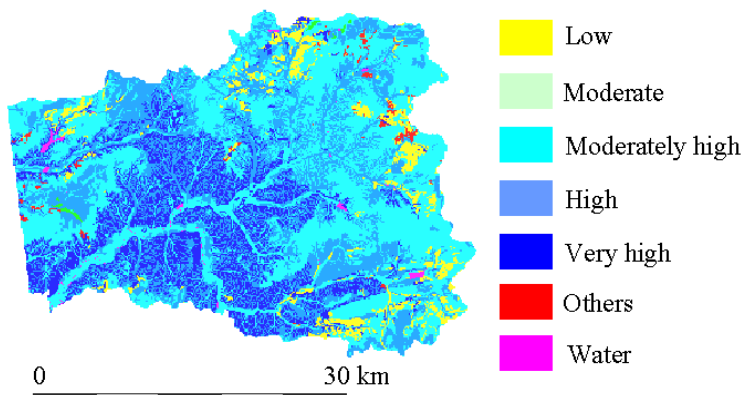


Figure 9. Spatial distribution of soil structure (pedality) in the watershed

Pedality points were regrouped to indicate water transmission potential in the profile. Five pedality groups were generated based on the total pedality points. For example, pedality point ranging from 10 – 17 was considered low (Table 10). Low pedality points are found in

small patches across the watershed (Figure 9). High (40 – 50) and very high (>51) pedality points together occupy about 49% of the watershed and are found mainly in the central part of the watershed. Most of the high pedality points are associated with coarse textured soils found closer to the main stream of the watershed. Table 10 shows pedality points for the major soils in the study area.

Table 10. Areal distribution of pedality points for soils in the watershed.

Soil Structure (Pedality Points)	ha	%
Low (10 - 17)	5,747	5.2
Moderate (18 - 30)	205	0.2
Moderately high (31 - 40)	47,656	43.6
High (40 - 50)	36,419	33.3
Very high (> 51)	18,001	16.5
Water	480	0.5
Others	760	0.7
Total	109,268	100

GIS-Based Neuro-fuzzy Model

The Neuro-fuzzy model was developed using trapezoidal membership functions. A dataset consisting of 202 combinations of patterns from input data was used to train the net. These patterns are also referred to as ‘cases’. The application data set for the watershed consisted of 2,662,528 rows and four columns of input data consisting of hydrologic groups, LULC, depth of the profile and soil structure for the study area. Four fuzzy sets were developed for each input parameter.

Characteristics of the Neuro-fuzzy Model

The parameter settings used for the model are presented in the Table 11. The software NEFCLASS-J developed by Nauck and Kruse (1999) was used for this study. The training data set was composed of 202 rows. Out of 202 rows 46 cases were classified as class 1 (high), 72

cases as class 2 (moderately high), 65 cases as class 3 (moderate) and 19 cases as class 4 (low). The validation technique ‘single test’ which randomly divides the data into two sets according to a given percentage value, was used to develop the model. The training process used 49% and the validation process used 51% of all cases presented in the data sets. A total of 41 possible rules were found. The optimal consequents were determined. ‘Best per class’ rule learning strategy was used for training and formulation of the rule base. The maximum numbers of rules were determined automatically. This option selects under the constraints of the size of the rule base the best per class. A final rule base with 29 rules was created. This rule base covered all patterns.

Table 11. The parameter settings for the Neuro-fuzzy model.

Parameters	Settings
Training data file	BasinsCLASSIF.dat
Number of fuzzy sets	4
Type of fuzzy sets	Trapezoidal
Aggregation function	Maximum
Interpretation of classification results	Winner takes all (WTA)
Size of rule base	Automatically determined
Learning rule procedure	Best per class
Fuzzy sets constraints	(i) Keep relative order (ii) Always overlap
Rule weights	Not used
Learning rate	0.1
Validation	Single test (50%). 50% of the data withheld from the training
Stop Control	Maximum number of epochs = 100 Minimum number of epochs = 0 Number of epochs after optimum = 10 Admissible classification errors = 0

Performance on training and validation data are presented in Tables 12 - 14. Table 12 indicates that about 15% of the training data which fell within the high category coincided with high category, about 24% of moderately high category coincided with moderately high category, 21% of moderate category coincided with moderate and 5% of the low category coincided with

low. For the training data sets, correct classification was 65 (65%) and number of misclassified entries were 35 (35%). For the validation data sets (49% of the training sets), the correct classification was 41 (40%) and number of misclassified entries were 61 (59%). Details of the rule bases and fuzzy sets used in this study can be found in Dixon (2001). Statistical characteristics and correlation analysis of the training data sets are presented in Tables 15 and 16. Tables 17 and 18 presented statistical characteristics and correlation analysis of application data, which consisted of 2,660,864 rows.

Table 12 . Performance of the training data (%) for ground water vulnerability classes.

Vulnerability classes	High	Moderately high	Moderate	Low	Not classified	Total
High	15	8	0	0	0	23
Moderately high	5	24	7	0	0	36
Moderate	0	5	21	2	4	32
Low	0	2	2	5	0	0
Total	20	39	30	7	4	100

Table 13. Performance of the validation data (%) for ground water vulnerability classes.

Vulnerability classes	High	Moderately high	Moderate	Low	Non Classified	Total
High	10 (9.80%)	8 (7.84%)	1 (0.98%)	0 (0.00%)	4 (3.92%)	23 (22.55%)
Moderately high	4 (3.92%)	19 (18.63%)	4 (3.92%)	3 (2.94%)	6 (5.88%)	36 (35.29%)
Moderate	2 (1.96%)	8 (7.84%)	11 (10.78%)	1 (0.98%)	11 (10.78%)	33 (32.35%)
Low	0 (0.00%)	4 (3.92%)	5 (4.90%)	1 (0.98%)	0 (0.00%)	10 (9.80%)
Total	16 (15.69%)	39 (38.24%)	21 (20.59%)	5 (4.90%)	21 (20.59%)	102 (100.00%)

Table 14. Characteristics of the training sets.

Learning Procedure	Patterns	Misclassification	Errors
Training	100	23	55
Validation	102	61	79

Table 15. Statistics for training data.

Input Variables	mean	std. deviation	minimum	maximum	missing
Var 1 Hydrologic Group	24.55	9.23	10	50	0
Var 2 LULC	32.77	16.98	10	60	0
Var 3 Depth	74.09	26.43	9	151	0
Var 4 Sructure	38.17	11.43	14	53	0

Table 16. Correlation for training data.

Input variables	1	2	3	4	Class
Hydrologic groups (1)	1	0.05	0.48	0.23	0.26
LULC (2)		1	0.05	-0.01	-0.21
Depth (3)			1	0.23	0.47
Structure (4)				1	-0.28

Table 17. Statistics for application data.

Input Variables	mean	std. deviation	minimum	maximum	missing
Var 1 Hydrologic Group	10.81	12.52	0	50	0
Var 2 LULC	11.01	13.85	0	60	0
Var 3 Depth	35.5	40.05	0	151	0
Var 4 Structure	19.13	21.91	0	53	0

Table 18. Correlation for application data.

Input variables	var 1	var 2	var 3	var 4	class
var 1 Hydrologic group	1	0.86	0.94	0.92	0
var 2 LULC		1	0.86	0.88	0
var 3 Depth			1	0.95	0
var 4 Structure				1	0

Spatial Distribution of Neuro-fuzzy Model

The spatial distribution of ground water vulnerability predicted from the Neuro-fuzzy model using the four input parameters, soil hydrologic groups, LULC, depth of the soil profile and soil structure (pedality points) is shown in Figure 10 and summarized in Table 19. The high ground water vulnerability category coincided with regions of hydrologic group C, agricultural land use, deep soils and high pedality points (40 – 50). Moderately high ground water vulnerability categories comprised 25% of the watershed and was distributed across watershed. The moderate vulnerability category coincided with urban LULC and soil hydrologic group C.

Fine tuning of the training data sets was required to improve the model's prediction from contradictory data. About 13% of the watershed area was not classified by the preliminary

model. This could be attributed to the validation technique used in the model, i.e. the single test. The results of the learning processes provide information on the smallest error caused during all propagations through the classifier and the error of the training data of the same cycle. This validation technique does not cross-validate error during the training processes.

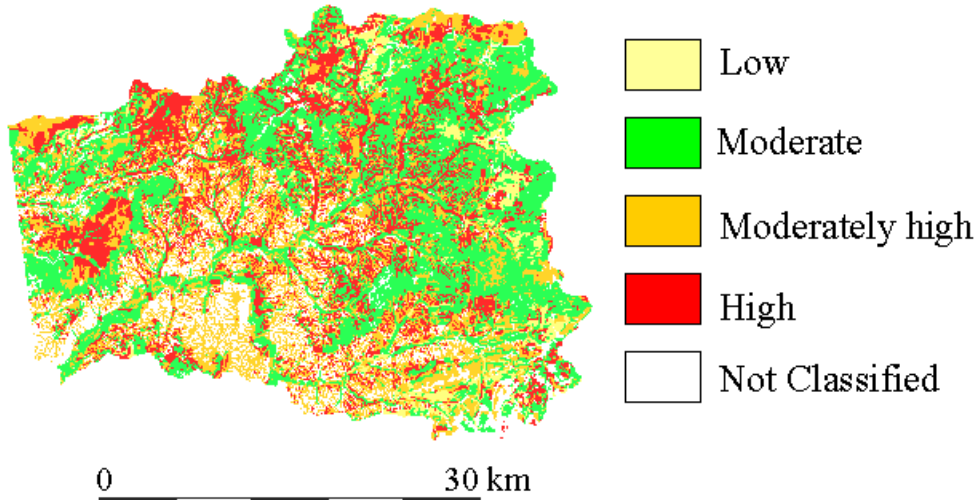


Figure 10. Spatial distribution of ground water vulnerability from the Neuro-fuzzy models in the watershed.

Table 19. Areal distribution of ground water vulnerability in the watershed.

Vulnerability	acres	ha	%
Non Classified (0)	37,375	15,125	13.8
High (1)	59,617	24,127	22
Moderately high (2)	68,975	27,914	25.8
Moderate (3)	93,277	37,748	34.5
Low (4)	10,758	4,354	3.9
Total	270,002	109,268	100

Coincidence Reports for the Neuro-fuzzy Model

Coincidence reports provided information on the mutual occurrence of the ground water vulnerability categories and the physical characteristics of the watershed. Two sets of coincidence reports were prepared and are shown in Figures 11 – 12. A higher proportion of the

highly vulnerable categories coincided with the soil hydrologic group C. Ideally, highly vulnerable areas were expected to coincide with soil hydrologic group B. However, in this watershed almost equal areas of soil hydrologic groups B and C coincided with moderately high vulnerability category (Figure 11a). As expected, a higher proportion of the highly vulnerable areas coincided with agricultural landuse. About 32,000 ha of agricultural land also coincided with moderately vulnerable categories (Figure 11b). A higher proportion of soils with deep profiles coincided with moderately high ground water vulnerable areas. About 10,000 ha of the very deep soils coincided with moderately high vulnerability categories. About 32,000 ha of the moderate vulnerability categories also coincided with deep soils (Figure 11c). Almost all of the high ground water vulnerability category coincided with soil structure or high pedality points with high water transmitting capabilities through the profiles (Figure 11d). The majority of the moderately vulnerable categories coincided with moderately high pedality points. Almost equal moderately high vulnerability category coincided with high and moderately high pedality points.

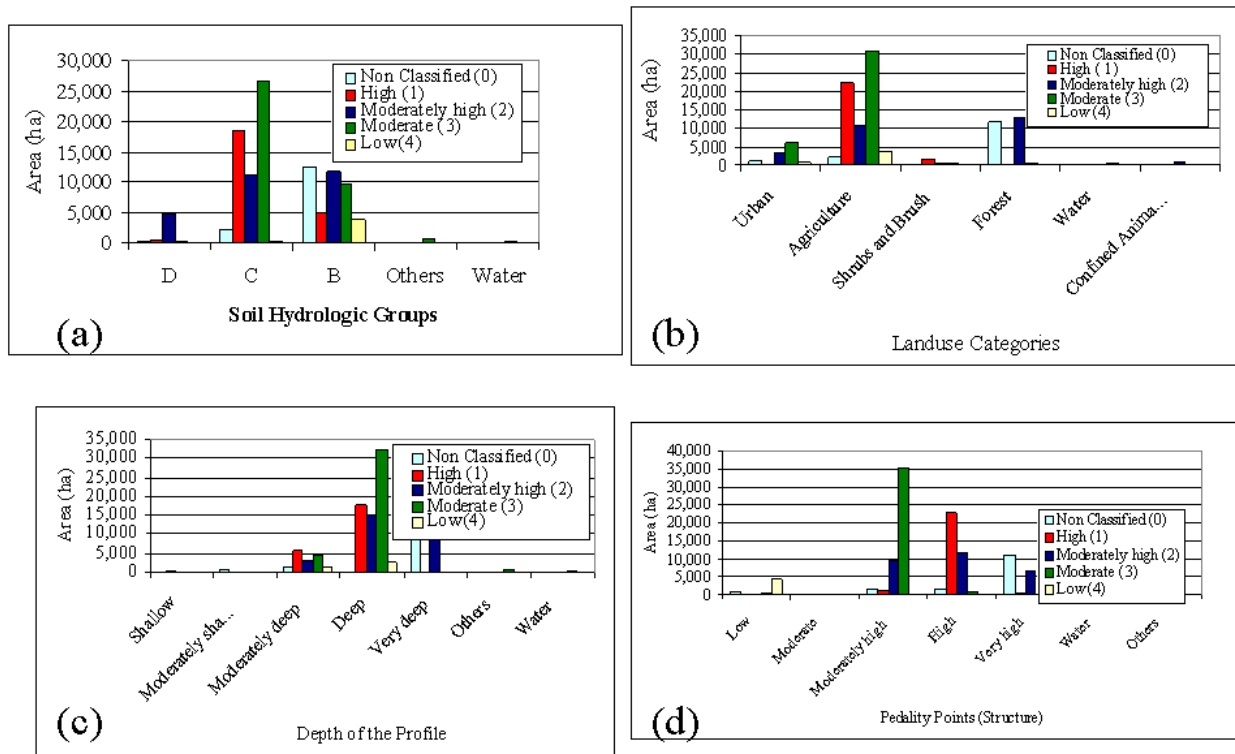


Figure 11. Mutual occurrence of Neuro-fuzzy-based ground water vulnerability categories and model inputs: (a) soil hydrologic groups, (b) landuse, (c) Depth of the soil profile and (d) Pedality Points (soil structure)

Nixa soils coincided with high or moderately high vulnerability categories. Almost all of Captina soil area was classified as moderately vulnerable. About 6,000 ha of the Clarksville soil area was classified as moderately high. Almost equal area with the highly vulnerable category coincided with nearly level, gentle and strong slopes (Figure 12a). The nearly level slope category also coincided with moderately vulnerable category. Areas with a combination of high vulnerability and nearly level slope have greater contamination potential than areas with high vulnerability but strong slopes. Almost all of the highly vulnerable area coincided with the

Boone geological formation followed by the Upper Mississippian Formation. About 32,000 ha of the watershed mapped as the Boone Formation coincided with the moderate vulnerability categories. The Boone formation covers about 87% of the study area (Figure 12 b).

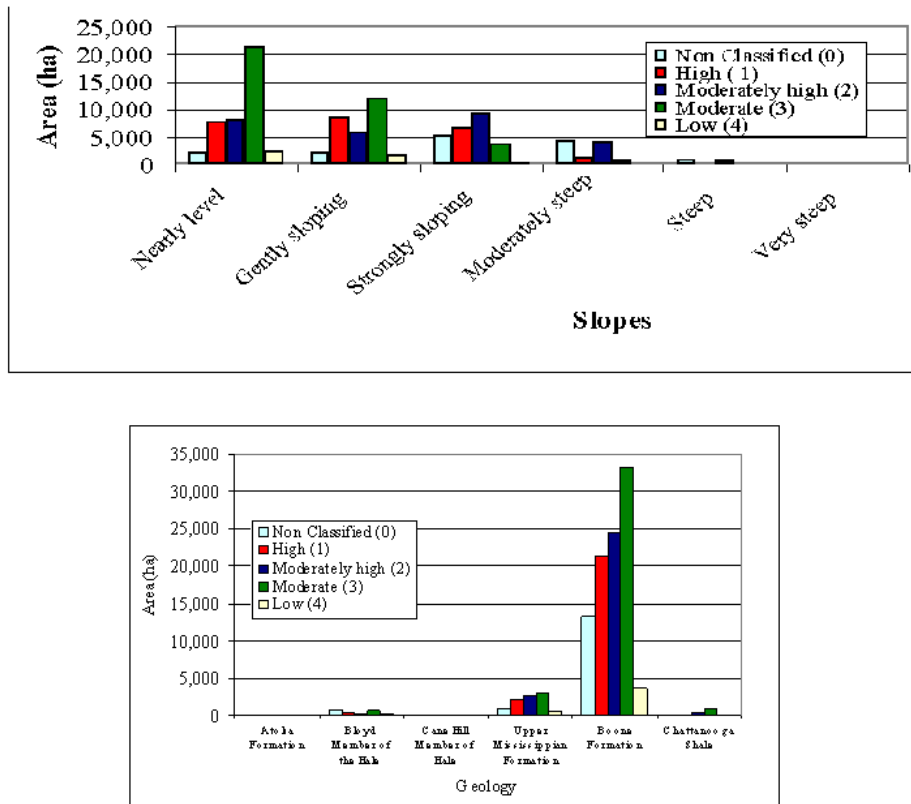


Figure 12. Mutual Occurrence of Neuro-fuzzy-based ground water vulnerability categories and validating parameters: (a) slopes and (c) geology

Coincidence with Field Data.

A set of coincidence reports was generated between vulnerability categories and the classes nitrate-N concentration data for all 44 wells and springs (Figure 13). The nitrate-N concentration data were classified into four categories: low (<0.5 mg/l), moderate (0.5 – 3 mg/l), moderately

high (3 – 10 mg/l) and high (> 10 mg/l). Two wells were classified as high concentration, one well coincided with high vulnerability and the other wells coincided with the moderate vulnerable category. Relatively higher number of wells with moderately high concentration level coincided with moderate vulnerability category followed by moderately high vulnerability category (Figure 13). Almost equal numbers of wells with moderate concentration level coincided with moderate and moderately high vulnerability categories. Two wells with moderately high concentration level coincided with low vulnerability area. Location and nitrate-N concentration levels (mg/l) of wells are shown in Figure 14.

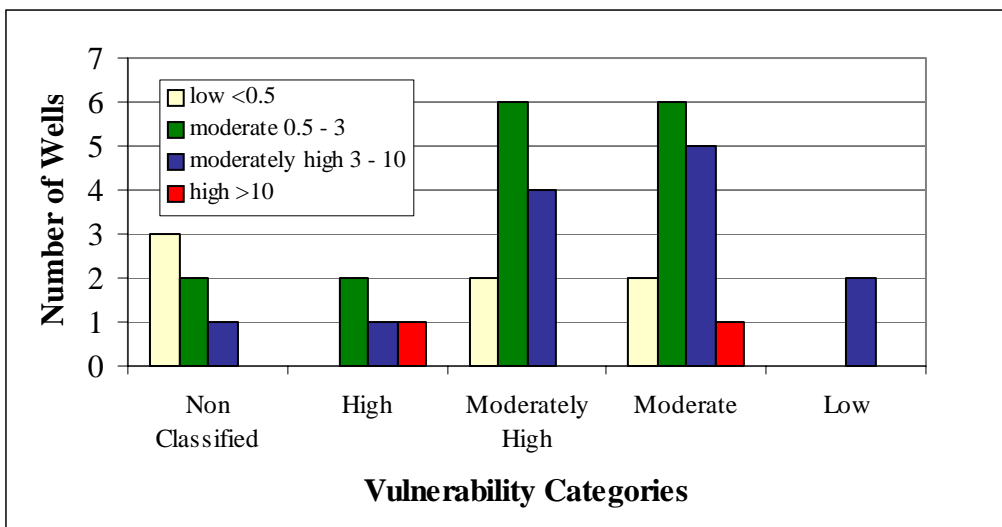


Figure 13. Coincidence results between vulnerability classes and well water quality a data.

Coincidence analyses between model inputs and well concentration data are presented in Figure 15. Only two wells sampled in the study area were categorized in the highly contaminated category. One of each associated with urban and agricultural LULC, moderately deep and deep soil profile, moderately high and high soil structure and one each with hydrologic groups B and C (Figure 15). The majority of the moderately high concentration levels are associated with

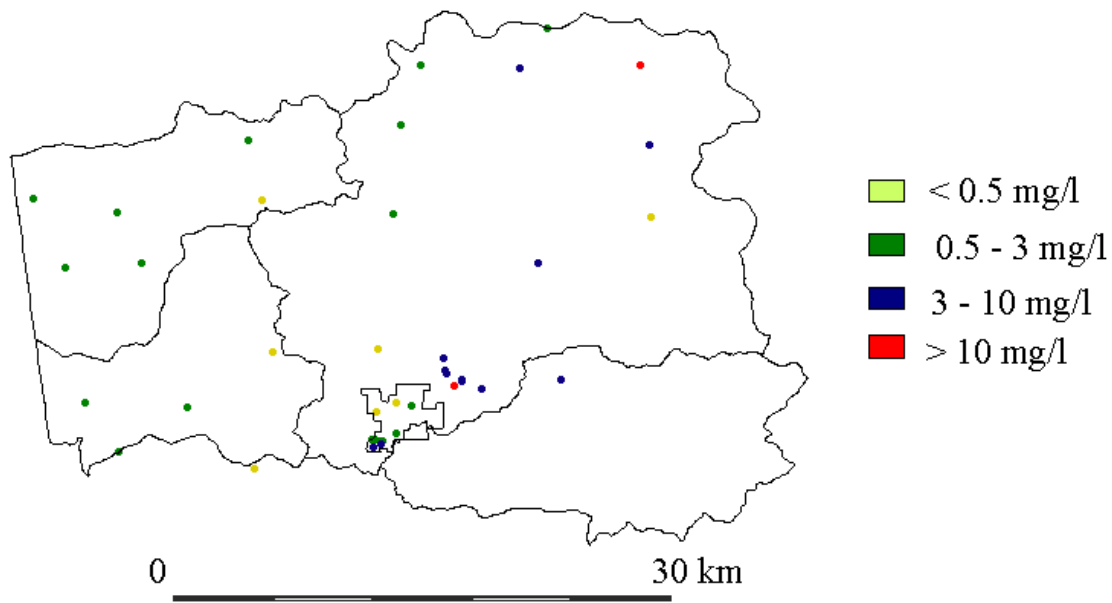


Figure 14. Location and nitrate-N contamination levels of wells in the watershed

agricultural landuse, moderately deep soil profile, soil hydrologic group B and moderately high pedality points. As mentioned earlier, the well concentration data were not collected during the same time nor by the same agencies and were compiled from different sources, therefore, this data set contained some additional uncertainty (point vs. areal) and variability. As a result, the comparison between well data and vulnerability categories should not be considered ‘absolute’ parameter in determining applicability of the model. Neuro-fuzzy approaches are capable of dealing with uncertainty of the data when they are used as input to the model, however, this approach can not help overcome the uncertainty of the data used to validate the model.

The data set for water quality was not considered to be adequate to determine the ability of the models to predict ground water vulnerability since it had inherent uncertainty. This brings out an interesting aspect of solute transport modeling on a regional scale as pointed out by Burrough (1996). He mentioned that most of the environmental data are collected on a project basis rather than in a systematic way which poses a problem in development of solute transport models in a regional scale. One of the goals of this research was to examine the usefulness of existing data in regional scale modeling of ground water vulnerability since this will reduce cost of modeling. Drilling new wells to validate the model will be cost prohibitive.

Moreover, well and spring data are point data, a validation technique that, compared

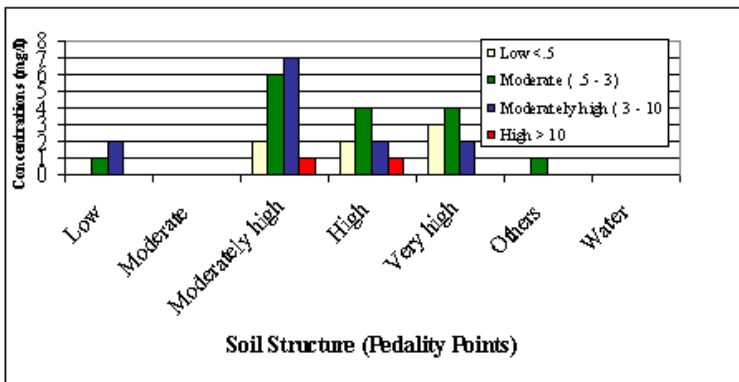
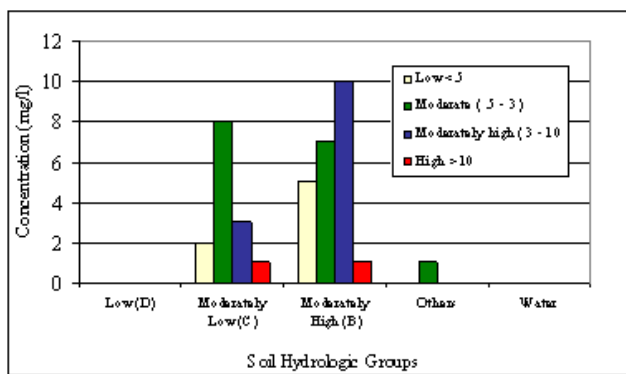
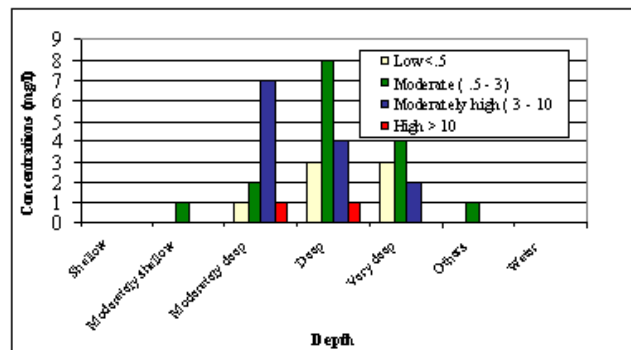
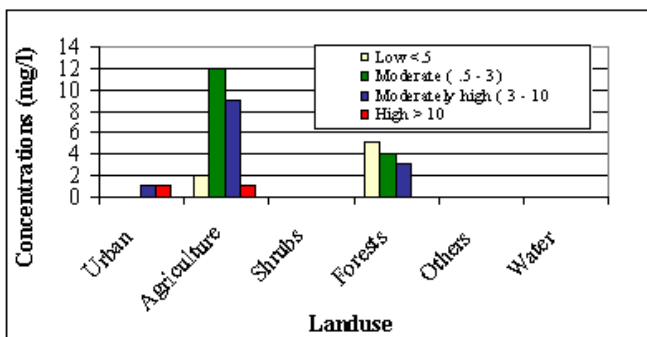


Figure 15. Coincidence between well contamination level and model inputs

point data with spatial data predicted by the model, has inherent uncertainty. Further studies are required with a larger set of water quality data collected all over the watershed. Use of geostatistical tools to generate concentration surface and comparison of that surface with the map generated by the vulnerability model could be useful in the assessment of the model performance. However, this might not be the most cost effective strategy for regional scale planning.

SUMMARY

About 22% of the watershed area was classified as highly vulnerable and almost all of the highly vulnerable areas coincided with agricultural landuse, moderately deep and deep soils, soil hydrologic group C and high pedality points. As expected, the vulnerability map showed high coincidence patterns between the highly vulnerable areas and LULC and soil structure (pedality points). However, the coincidence patterns between the vulnerability categories and depth of the soil profile and soil hydrologic groups did not show the expected patterns of coincidence. The expected patterns were: the shallower the profile, the higher the vulnerability; or the higher the Ksat of a soil hydrologic group, the higher the vulnerability (i.e. soil hydrologic group B is more vulnerable than C). These discrepancies could be attributed to the fact that the majority of the agricultural landuse coincided with soil hydrologic group C (40,178 ha) followed by soil hydrologic group B (24,807 ha). About 47,818 ha of the watershed with deep soils coincided with agricultural landuse. Only two wells sampled in the study area had NO₃-N concentration of greater than 10 mg/l. One of each coincided with highly and moderately vulnerable areas, agricultural and urban landuse, moderately deep and deep soil profile, soil hydrologic groups B and C and pedality points of moderately high and high. The water quality data were not sufficient to characterize the entire watershed. Locations of the wells and spring used in the study

to validate the model had a spatial bias, and therefore, were not especially useful in validating the Neuro-fuzzy model. As mentioned earlier due to the inherent uncertainty (temporal and point vs. areal) associated with the well concentration data, the data set should not be considered as ‘absolute’ parameter in determining performance of the model. As mentioned earlier, Neuro-fuzzy approaches are capable of dealing with uncertainty of the data when they are used as input to the model, however, this approach can not help overcome the uncertainty of the data used to validate the model.

CONCLUSION

This research used Neuro-fuzzy techniques to predict ground water vulnerability in northwest Arkansas. These techniques allowed incorporation of expert’s opinion in the models, which is a valuable source of information, particularly for the parameters that are hard to measure and vary over space and time. The models developed in this research used simple soil parameters, including depth of the soil profile, soil hydrologic groups and pedality points of the A horizon and LULC to ensure global scope of the model. Since the underlying geology of the watershed is primarily the Boone Formation, which is highly fractured, it was assumed that any contaminants that reached the Boone formation would also move to the ground water because this formation poses little hydraulic resistance to flow.

Application of Neuro-fuzzy techniques to the prediction of ground water vulnerability does not provide exact solutions. Fuzzy systems, which are used to exploit the tolerance for imprecise solutions, are useful because they are easy to use, handle, and understand. Use of the NEFCLASS-J tool provided all necessary statistics. No further statistical tools or computation with statistical tools were required. The preliminary model needed to be fine tuned through fine tuning of the rule base and classifier. From this research it is evident that the tool NEFCLASS-J

could not automatically create the classifier. It supports the user but it cannot do all the work because a precise and interpretable fuzzy classifier can hardly be found by an automatic learning process. The NEFCLASS-J needs experts' opinion and tuning.

This methodology has potential in facilitating modeling ground water vulnerability at a regional scale. This methodology can be used for other regions, however, this approach would require incorporation of appropriate input parameters suitable for the region. For example, if the geology of an area is different from the study area, geological factors should be incorporated to account for potential resistance to water and contaminants transport processes. This study is a first step toward incorporation of Neuro-fuzzy techniques in a GIS and would require modifications for wider range of application.

Project supported by: Arkansas Water Resources Research Grant

REFERENCES

- Al-Rashidy, S. 1999. Hydrogeologic controls of ground water in the shallow mantled karst aquifer, Copperhead Spring, Savoy Experimental Watershed, Northwest Arkansas. M.S. Thesis. University of Arkansas. p. 117
- Burrough, P. A. 1996. Opportunities and Limitations of GIS-based Modeling of Solute Transport at the Regional Scale. In (D. Corning and K. Loague eds.), Application of GIS to the non-point source pollutants in the vadose zone, SSSA, Special Publication # 48, Madison, WI. p.367
- Chitsazan, Manouchehr. 1980. Hydrogeologic evaluation of the Boone-St. Joe carbonate aquifer. M.S. Thesis, Department of Geology, University of Arkansas, Fayetteville. p.140.
- 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 (D. L. Corwin, K. Loague and T. R. Ellsworth, eds.) Assessment of non-point source pollution in the vadose zone. Geophysical Monograph 108. AGS, Washington D.C. p.386
- Croneis, C. 1930. Geology of the Paleozoic region with special reference to oil and gas possibilities. Arkansas Geologic Survey Bulletin # 3. p. 457.
- Curtis, D. L. 2000. An integrated rapid hydrogeologic approach to delineate areas affected by advective transport in mantled karst, with an application to Clear Creek Basin, Washington County, Arkansas. Ph.D. Dissertation, University of Arkansas, Fayetteville. p. 121.
- Dixon, B. 2001. Application of Neuro-fuzzy techniques to predict ground water vulnerability in NW Arkansas. Doctoral Dissertation. University of Arkansas. Environmental Dynamics. p. 265
- Hays, T. S. 1995. Digital Characterization of the Illinois River Watershed and simulated phosphorus loading in representative sub-basins. M.S. Thesis. Department of Agronomy, University of Arkansas, Fayetteville. p. 156.
- Hunter G.J. and M. F. Goodchild. 1996. Communicating uncertainty in spatial databases. Transactions in GIS 1, 1:13-24
- Khan, E. 1999. Neural fuzzy based intelligent systems and applications In (Jain, L.C and N. M. Martin eds.) Fusion of neural networks, fuzzy sets and genetic algorithms: industrial applications. CRC Press, Washington, D.C. p. 331.

- Lin, H. S., K. J. McInnes, L.P. Wilding, and C. T. Hallmark. 1999. Effects of soil morphology on hydraulic Properties: I. quantification of soil morphology. *Soil Science Society of America Journal*. 63:948- 953.
- Mitra, B., J. M. McKimney, H. D. Scott, and T. S. Hays. 1997. Development and use of digital databases in agricultural research. *Trends in Agronomy*. 1:1- 17.
- National Research Council (NRC). 1993. Ground water vulnerability assessment: contamination potential under conditions of uncertainty. Committee on Techniques for Assessing Ground Water Vulnerability, Water Science and technology Board, Commission on Geosciences, Environment, and Resources. National Academy Press, Washington D.C. p.179.
- Nauck, D., F. Klawonn, and R. Kruse. 1997. *Foundations on Neuro-Fuzzy Systems*. John Wiley, Chichester. p. 356
- Nauck, U. and R. Kruse. 1999. Design and Implementation of a Neuro-fuzzy data analysis tool in JAVA. Manual. Technical University of Braunschweig. Germany.
- Petersen, J.C., Adamski, J.C., Bell, Davis, J.V., Femmer, S.R., Freiwald, D.A., and R.. L. Joseph. 1998, Water Quality in the Ozark Plateaus, Arkansas, Kansas, Missouri, and Oklahoma, 1992-95.U.S. Geological Survey Circular 1158, on line at < URL: <http://water.usgs.gov/pubs/circ1158>>
- Razaie, N. 1979. The hydrogeology of the Boone-St. Joe aquifer of Benton County, Arkansas. M.S. Thesis, Department of Geology, University of Arkansas, Fayetteville. p. 147.
- Smith, C. R. and K. F. Steele. 1990. Nitrate concentrations of ground water in Benton County, Arkansas. AWRC Pub. # 73. p. 48.
- Soil Division Staff. 1993. *Soil Survey Manual*. USDA Handbook. No.18. U.S. Gov. Print. Office, Washington, DC. p. 436.
- Steyaert L.T. and M. F. Goodchild 1994. Integrating geographic information systems and environmental simulation models: a status review. In (Michener W.K., ed) *Environmental information management and analysis*. Taylor & Francis. p. 333-355.
- Yager, R. R., Ovchinnikov, S., R. M. Tong and H. T. Nguyen. 1987. Coping with the imprecision of real world: an interview with L. A. Zadeh. In (Yager, R. R., Ovchinnikov, S., R. M. Tong and H. T. Nguyen, Eds.) *Fuzzy sets and applications: selected papers by L. A. Zadeh*. John Wiley and Sons, NY. p.436
- Yen, J. and R. Langari. 1998. *Fuzzy Logic: intelligence, control and information*. Prentice Hall. NJ. p. 548.
- Zadeh, A. L. 1965. Fuzzy sets. *Information and Control*. 8:338-353