GIS and Statistical Groundwater Vulnerability Modeling

Sharon L. Qi and Jason J. Gurdak

ABSTRACT

ArcGIS Desktop* and ArcGIS Workstation* were used to extract statistically significant information from various geospatial data sets for input into a statistical model of groundwater vulnerability. The product of this effort was a probability map that identified areas of vulnerability to groundwater-quality degradation. This information is of interest to a variety of water professionals because it provides a tool to help make educated decisions regarding the management of groundwater resources in the High Plains aquifer.

Because the study area includes 174,000 mi^2 in parts of eight States, a Geographic Information System (GIS) was required to efficiently extract information for 31 variables from 14 spatial-data layers at 6,416 well locations throughout the study area. The layers were both vector and raster and included information about depth to water, aquifer saturated thickness, aquifer hydraulic conductivity, aquifer specific yield, average annual precipitation rates, percentage of irrigated land and agricultural land (irrigated and nonirrigated), chemical application rates (nitrogen/phosphorus/pesticide), manure application rates, soil characteristics, and water-use estimates. For categorical data sets and certain continuous data sets (precipitation, depth to water, saturated thickness, hydraulic conductivity, specific yield, chemical applications, manure applications, and water use) the data were extracted directly from the layer at the location of each well by using a series of identity overlays. For other layers where information needed to be related to the area around a well (soil characteristics, percentage of irrigated land around a well, percentage of agricultural land around a well), buffers of varying sizes were created around each well and the information was inventoried for the buffer areas using both vector union techniques and raster map-algebra techniques.

The extracted data were used as variable input for an iterative series of statistical calculations using logistic regression. These calculations determined which of the variables (layers) or combination of variables were significantly correlated with observed nitrate concentrations in the groundwater. The variables became part of an equation defining the probability of a dissolved constituent in the ground water to be above a specified concentration. Once the probability equation was defined, the appropriate GIS layers representing the significant variables were converted to raster data sets (if they were vector data sets; raster data sets were used as is) to utilize the map algebra capabilities of ArcGIS. The significant raster data layers, the equation, and its various coefficients for each layer were put back into the GIS and, using map algebra, the probability surface was calculated and then easily visualized using the GIS.

*The use of trade names are used for descriptive purposes only and should not be considered an endorsement of a specific product by the U.S. Government.

INTRODUCTION

Nitrate contamination and the associated health concerns are one of the most common problems adversely affecting groundwater quality worldwide *(Canter, 1997)* . Nitrate is a highly mobile contaminant that originates from a variety of point and nonpoint sources. Nitrate contamination of groundwater resources is increasing because of the increase in sources from human activity *(Korom, 1992)* . The High Plains aquifer is no exception; nitrate concentrations in groundwater samples exceed the National drinking-water standard for public supply *(Litke, 2001)* , indicating the aquifer is vulnerable to nitrate contamination. Detailed monitoring of nitrate concentrations to estimate groundwater vulnerability across the aquifer is cost prohibitive because of the vast areal extent of the High Plains *(fig. 1)* . Therefore, modeling the vulnerability of the High Plains aquifer to contamination using robust statistical methods is a viable alternative to widespread groundwater monitoring. The development of a groundwater vulnerability model for nitrate in the High Plains aquifer will provide a valuable tool for water professionals responsible for making decisions about the continuing development of water resources in the High Plains aquifer. Because of the large area involved, one of the initial activities of the High Plains effort was a pilot study of groundwater vulnerability conducted in the central High Plains region to assist in developing the techniques used to extract the spatial data, to model the data using the statistical analysis technique of logistic regression, and to visualize the model results in the GIS.

DATA EXTRACTION

Spatial data were extracted in two ways. For categorical data and certain continuous data layers, the information was extracted from the layer at the well-point location by using multiple point-on-polygon identity overlays (vector layers) and the point-on-raster overlay command LATTICESPOT (raster layers) *(fig. 2)* . These techniques were used on the following data: depth to water; aquifer saturated thickness; average annual precipitation rate; nitrogen, phosphorus, and pesticide application rates; manure application rates; aquifer hydraulic conductivity; aquifer specific yield; and water-use estimates. For other continuous data sets (soil data, percentage of irrigated land, and percentage of agricultural land) where the information needed to be summarized for the area around each well, various buffers with radii of 500 m and 5,000 m were created and the data were summarized using raster map-algebra techniques *(fig. 3)* . The 5,000-m buffers of more than 6,000 wells generally were overlapping; therefore, an Arc Macro Language (AML) program was used in the GRID environment to process each buffer area individually and summarize the data in a table for each well point.

The data for each well location were exported from ArcGIS as a series of tables. These tables were then imported into the SAS® statistical package *(SAS Institute, Inc., 2001)* where the various logistic-regression models were created.

THE MODEL

Logistic regression is a statistical technique that uses one or more independent explanatory variable to predict the probability of a binary, or categorical response (the reader is referred to Hosmer and Lemeshow [2000] for detailed discussion of logistic regression). Using GIS, the spatial data were extracted for each well location and then used as input for univariate logistic-regression analyses to determine which variables were statistically related ($p < 0.1$) to observed nitrate concentrations in the groundwater. All statistically significant variables were then used as input for multivariate logistic-regression analyses to determine which combination of variables was most significantly correlated ($p < 0.05$) to the observed nitrate concentrations. For the purposes of this study, the model was designed to determine the probability of detecting nitrate concentrations greater than 4 mg/L in groundwater (the threshold at which nitrate contamination is likely the result of human activity) *(Mueller and Helsel, 1996)*.

In the pilot study, the most significant multivariate variables were the percentage of irrigated land in a 5,000-m buffer around the well *(fig. 4)* , amount of nitrogen applications from fertilizer *(fig. 5)* , and the thickness of the unsaturated zone *(fig. 6)* . The following equation was used to determine the probability of nitrate concentrations greater than 4 mg/L in the aquifer. Where *a* is the constant coefficient for the model; *b*, *c*, and *d* are the coefficients (weights) for each variable; and *e* is the natural log.

$$\text{Probability} = \frac{e^{(a + b^{*}\text{percent irrigated land} + c^{*}\text{nitrogen application} + d^{*}\text{unsaturated thickness})}}{1 + e^{(a + b^{*}\text{percent irrigated land} + c^{*}\text{nitrogen application} + d^{*}\text{unsaturated thickness})}}$$

For the model used in this study, $a = –0.2006$, $b = 0.0304$, $c = 0.0317$, and $d = –0.00802$.

THE VISUALIZATION

Once the significant model variables have been identified, they are used as input into the probability equation. This equation calculates the probability of a nitrate concentration being above a selected threshold concentration. Because the GIS supports map algebra syntax, this equation can be solved and visualized within the GIS *(fig. 7)* . Note that a variable needed for the equation may be different from the information provided by the original layer. For example, the percentage of irrigated land around each sampling well was extracted using a raster data set of the location of irrigated land in the High Plains. However, the variable required to calculate the probability surface for the entire study area was the summary variable "percentage of irrigated land in a 5,000-m buffer." Therefore, this value had to be calculated for each point (grid cell) over the entire study area, not just each well location.

The resulting product is a map illustrating the probability of detecting nitrate concentration in groundwater above a selected threshold. This product can be used by water professionals to make informed decisions about the management of groundwater resources in the High Plains.

CONTINUING EFFORTS

The initial pilot study in the central High Plains was completed in early 2003. Work has continued to include the entire High Plains aquifer area (approximately 174,000 mi^2). Spatial data for 48 variables have been extracted from 17 individual layers at 6,985 well point locations where water-quality data have been collected. Information from data sets include a subset of those used in the pilot study: depth to water, aquifer saturated thickness, average annual precipitation rates, percentage of irrigated land/agricultural land (irrigated and nonirrigated), chemical application rates (nitrogen/phosphorus/pesticide), manure application rates, soil characteristics, and water-use estimates, and four new layers—hydrogeologic region, irrigation-well density, the location of playa lakes relative to the well (thought to be locations of focused aquifer recharge), and unsaturated-zone lithology (percentages of clay and sand in the unsaturated zone). For data extractions that require a well buffer, the buffer shape was modified from a circle to a 90-degree sector, or wedge, that is oriented in the upgradient direction of groundwater flow and whose radius is determined by the hydraulic conductivity of the aquifer near the well *(fig. 8)* . This approach will summarize the spatial information only for the area thought to more directly influence the water quality in the well. The multivariate analyses are being completed to determine which variables are most significantly correlated to observed nitrate concentrations in the High Plains aquifer. Once these multivariate analyses are complete, a single probability model will be selected and a surface representing the probability of nitrate concentrations greater than 4 mg/L can be created for the entire High Plains area.

CONCLUSION

Because of the large areal extent, ArcGIS was essential for the creation of a probability map of nitrate contamination within the High Plains aquifer. Ancillary spatial data were extracted from 18 layers across the entire High Plains aquifer (174,000 mi^2). The map algebra and visualization capabilities of the GIS were used so that the probability of nitrate contamination in the aquifer derived from a statistical package could be calculated as a surface over the entire aquifer. The map product is an effective tool for water professionals in the continuing development of groundwater resources throughout the High Plains aquifer.

REFERENCES

Canter, L.W., 1997, Nitrates in groundwater: Boca Raton, CRC Press, Inc., 263 p.

Hosmer, D.W., and Lemeshow, S., 2000, Applied logistic regression: New York, John Wiley, 372 p.

Korom, S.F., 1992, Natural denitrification in the saturated zone—A review: Water Resources Research, v. 28, no. 6, 1957-1668 p.

Litke, D.W., 2001, Historical water-quality data for the High Plains Regional Groundwater study area in Colorado, Kansas, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wyoming, 1930-98: U.S. Geological Survey Water-Resources Investigations Report, 00-4254, 65 p.

Mueller, D.K., and Helsel, D.R., 1996, Nutrients in the Nation’s water—Too much of a good thing?: U.S. Geological Survey Circular 1136, 24 p.

SAS Institute, Inc., 2001, SAS software, Version 8.02 of the SAS System for PC. Copyright © 2001 SAS Institute Inc.

AUTHOR INFORMATION

Sharon L. Qi, Hydrologist, U.S. Geological Survey, 3200 SW Jefferson Way, Corvallis, OR, 97331, USA; Voice: (541)758-8815, Fax: (541)758-8806, E-mail: slqi@usgs.gov

Jason J. Gurdak, Hydrologist, U.S. Geological Survey, Box 25046, MS 415, Denver Federal Center, Lakewood, CO, 80225, USA; Voice: (303)236-4882 ext 222, Fax: (303)236-4912, E-mail: jjgurdak@usgs.gov