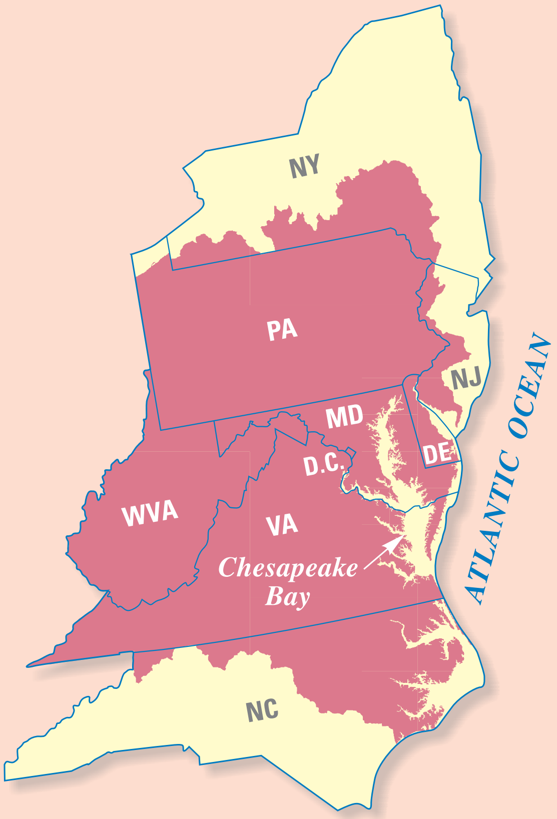
USGS
science for a changing world

125 years of science for America ★ ★ ★ ★ ★ 1879–2004

In cooperation with the
U.S. Environmental Protection Agency
Regional Vulnerability Assessment (ReVA) Program

Ground-Water Vulnerability to Nitrate Contamination at Multiple Thresholds in the Mid-Atlantic Region Using Spatial Probability Models

Scientific Investigations Report 2004-5118



U.S. Department of the Interior
U.S. Geological Survey

This Page Intentionally Left Blank

U.S. Department of the Interior
U.S. Geological Survey

Ground-Water Vulnerability to Nitrate Contamination at Multiple Thresholds in the Mid-Atlantic Region Using Spatial Probability Models

by Earl A. Greene, Andrew E. LaMotte, and Kerri-Ann Cullinan

Scientific Investigations Report 2004-5118

In cooperation with the
U.S. Environmental Protection Agency
Regional Vulnerability Assessment (ReVA) Program

125 years of science for America ★ ★ ★ ★ ★ *1879–2004*

Reston, Virginia
2005

U.S. Department of the Interior
GALE A. NORTON, Secretary

U.S. Geological Survey
Charles G. Groat, Director

The use of trade, product, or firm names in this report is for descriptive purposes only and does not imply endorsement by the U.S. Government.

The U.S. Environmental Protection Agency, through its Office of Research and Development, partially funded and collaborated in the research described in this Scientific Investigations Report under Interagency Agreement number DW 14939071 to the U.S. Geological Survey. It has been subjected to agency review and approved for publication.

For additional information contact:

District Chief
U.S. Geological Survey
8987 Yellow Brick Road
Baltimore, MD 21237

Copies of this report can be purchased from:

U.S. Geological Survey
Branch of Information Services
Box 25286
Denver, CO 80225-0286

CONTENTS

Figures

Figures—Continued

Tables

Conversion Factors, Vertical Datum, and Abbreviated Water-Quality Units

Multiply	By	To obtain
square meter (m^2)	0.0002471	acre
cubic meter (m^3)	0.0008107	acre-foot (acre-ft)
cubic meter per second (m^3/s)	35.31	cubic foot per second (ft^3/s)
meter (m)	3.281	foot (ft)
meter per kilometer (m/km)	5.27983	foot per mile (ft/mi)
liter per second (L/s)	15.85	gallon per minute (gal/min)
kilometer (km)	0.6214	mile (mi)
square kilometer (km^2)	0.3861	square mile (mi^2)

Vertical coordinate information is referenced to the North American Vertical Datum of 1988 (NAVD 88); horizontal coordinate information is referenced to the North American Datum of 1983 (NAD 83).

Concentrations of chemical constituents in water are given in either milligrams per liter (mg/L) or micrograms per liter (µg/L).

Application of fertilizer and atmospheric deposition are given in kilogram per square kilometer (kg/km^2).

This Page Intentionally Left Blank

Ground-Water Vulnerability to Nitrate Contamination at Multiple Thresholds in the Mid-Atlantic Region Using Spatial Probability Models

By Earl A. Greene, Andrew E. LaMotte, and Kerri-Ann Cullinan

Abstract

The U.S. Geological Survey, in cooperation with the U.S. Environmental Protection Agency's Regional Vulnerability Assessment Program, has developed a set of statistical tools to support regional-scale, ground-water quality and vulnerability assessments. The Regional Vulnerability Assessment Program's goals are to develop and demonstrate approaches to comprehensive, regional-scale assessments that effectively inform managers and decision-makers as to the magnitude, extent, distribution, and uncertainty of current and anticipated environmental risks. The U.S. Geological Survey is developing and exploring the use of statistical probability models to characterize the relation between ground-water quality and geographic factors in the Mid-Atlantic Region.

Available water-quality data obtained from U.S. Geological Survey National Water-Quality Assessment Program studies conducted in the Mid-Atlantic Region were used in association with geographic data (land cover, geology, soils, and others) to develop logistic-regression equations that use explanatory variables to predict the presence of a selected water-quality parameter exceeding a specified management concentration threshold. The resulting logistic-regression equations were transformed to determine the probability, *P(X)*, of a water-quality parameter exceeding a specified management threshold. Additional statistical procedures modified by the U.S. Geological Survey were used to compare the observed values to model-predicted values at each sample point. In addition, procedures to evaluate the confidence of the model predictions and estimate the uncertainty of the probability value were developed and applied. The resulting logistic-regression models were applied to the Mid-Atlantic Region to predict the spatial probability of nitrate concentrations exceeding specified management thresholds. These thresholds are usually set or established by regulators or managers at National or local levels.

At management thresholds of 1 milligram per liter and 3 milligrams per liter as nitrogen, the probability of nitrate concentrations exceeding these levels is greater than 50 percent (0.50) throughout much of the Mid-Atlantic Region. This includes extensive areas throughout central Maryland, southeastern Pennsylvania, northwestern Pennsylvania, and the Delmarva Peninsula. In addition, extensive areas in North Carolina and Virginia also have high probabilities of nitrate concentrations in ground water exceeding management thresholds of 1 milligram per liter and 3 milligrams per liter. The mapped areas showing a high predicted probability of nitrate concentrations in ground water exceeding 1 milligram per liter and 3 milligrams per liter correspond to areas that are mapped as cultivated land cover and/or overlying carbonate rocks. At a management threshold of 10 milligrams per liter (corresponding to the U.S. Environmental Protection Agency standard for nitrate in drinking water of 10 milligrams per liter), the predicted probability of nitrate concentrations in ground water exceeding this level is low for most of the Mid-Atlantic Region, except for the Delmarva Peninsula, southeastern Pennsylvania, and areas mapped as carbonate rocks in Virginia, Maryland, and Pennsylvania.

Introduction

The U.S. Environmental Protection Agency (USEPA) has recognized that regional, State, and local water managers in the Mid-Atlantic Region need to know the condition of ground-water quality at a regional scale for environmental health and human health purposes (such as public drinking-water consumption). Over the last several years, the USEPA and the U.S. Geological Survey (USGS) have been conducting National and regional assessments of water quality in ground water in the Mid-Atlantic region (Ator and Ferrari, 1997; Ferrari and others, 1997). Although these studies have addressed questions of overall water-quality conditions in the region, they were not designed to predict water quality in areas with little or no data. In order to determine ground-water vulnerability in these unknown areas, the USGS, in cooperation with the USEPA, began a study in 2001 to address this issue.

Purpose and Scope

The USGS, in cooperation with the USEPA's Regional Vulnerability Assessment (ReVA) Program, is developing a set of statistical tools to support regional-scale integrated ecological risk assessment studies. ReVA's goal is to develop and demonstrate approaches to comprehensive, regional-scale assessments that effectively inform decision-makers as to the magnitude, extent, distribution, and uncertainty of current and anticipated environmental risks. The USGS is developing and exploring the use of statistical probability models to characterize the relation between ground-water quality and geographic factors in the Mid-Atlantic Region.

This report describes a spatial statistical methodology that can be used to assess the risk of nonpoint-source contamination of ground water exceeding a management threshold in areas of the Mid-Atlantic Region where little data exits. Thresholds can be formulated and specified as management values such as 3 mg/L (milligrams per liter) of nitrate as nitrogen for environmental concerns or 10 mg/L of nitrate as nitrogen for regulating drinking water. In addition, the PRESS statistic (predictive ability) and the uncertainty (confidence interval) of the models were developed and applied to all spatial probability maps. Spatial probability maps showing the likelihood of elevated concentrations of nitrate above a management threshold were developed to identify areas that currently are vulnerable to nitrate contamination. These maps help identify areas where ground water has been affected by human activities, and can help regional and local water managers protect water supplies by targeting land-use planning solutions and implementing monitoring programs where ground water may be vulnerable.

Previous Investigations

Various investigations have shown there is a relation between ground-water quality and other variables, such as land cover and geology (Ator and Ferrari, 1997; Tesoriero and Voss, 1997). Cain and others (1989) evaluated regional ground-water quality using changes in land use. Barringer and others (1990) also related land use to ground-water quality, but discussed problems (misclassification, data closure, and spatial autocorrelations) with using land use as an explanatory variable. Some studies also have shown that when relating ground-water quality to land use, the size of the buffer around the well is an important factor. The maximum correlation between land use and nitrate concentration was reached when the land use around a well was quantified within a radius of 800 to 1,200 m (meters) (Hay and Battaglin, 1990). Tesoriero and Voss (1997) determined the best logistic-regression model relating land cover to nitrate concentration exceeding a 3-mg/L threshold was at a radius of 3,200 m.

Recent work relating regional ground-water quality to explanatory variables (land cover, geology, soils, and others) has used statistical models to explain these relations. These investigations usually use nitrate as a surrogate for ground-water contamination or for an estimate of vulnerability. A variety of statistical methods can be used to relate nitrate contamination to explanatory variables. The use of logistic-regression techniques and their application to hydrology are well documented in the literature (Walker and Duncan, 1967; Harrell and others, 1980; Helsel and Hirsch, 1992; Eckhardt and Stackelberg, 1995; Tesoriero and Voss, 1997; Nolan, 2001).

Acknowledgments

The authors thank Dr. Nagaraj Neerchal, Professor of Statistics at the University of Maryland, Baltimore County, for his statistical advice and Joseph Vrabel of the USGS for help in developing the data base for the project. The authors would also like to thank Bernard T. Nolan and A. Jim Tesoriero of the USGS for their technical reviews. Lastly, we thank Valerie Gaine, Jean Hyatt, Donna Knight, and Timothy Auer of the USGS Maryland-Delaware-D.C. District for editorial review, report preparation, and report graphics.

Data Description

Data needed for model development were obtained from previously published sources and available USGS data stored in files for ground-water and water-quality data bases. These data bases contain information obtained from National Water-Quality Assessment (NAWQA) studies, statewide ground-water-quality networks, and individual projects. Geographic data sets, such as land cover, surficial geology, soil permeability, soil organic matter, depth of soil layer, depth to water table, soil texture, hydrologic groups, manure, fertilizer, atmospheric deposition, and population density, were obtained for the Mid-Atlantic Region from published sources.

Water-quality data collected from October 1985 through September 1996 as part of the NAWQA Program (Gilliom and others, 1995) and various other USGS projects were compiled in 1997 for a regional analysis of pesticides and nitrate in surface and ground water of the Mid-Atlantic Region (Ator, 1998; Ferrari and others, 1997; Ator and Ferrari, 1997). Ground-water-quality data consisted of 1,551 samples taken from 937 different shallow observation wells (95 percent less than 33 m), with multiple samples taken at various sites (Ator, 1998; Ator and Ferrari, 1997). When multiple samples were collected at a site, the most recent sampling date was used. Ten sites were eliminated because of either missing or incomplete data, including well depth, lack of nitrate testing, and site information, thereby reducing the data set to 927 sites (fig. 1).

Dependent Variable

Many contaminants in ground water occur naturally; however, elevated nitrate concentrations in ground water typically are caused by anthropogenic (human-related) activities that include crop fertilization and domestic septic systems. Spalding and Exner (1993) suggested that nitrate may be the most widespread contaminant in ground water. Because of the extensive presence of this contaminant, nitrate concentrations in ground water may help identify environments that are susceptible to contamination (U.S. Environmental Protection Agency, 1996a).

Nitrate as nitrogen was chosen as a surrogate variable to develop spatial maps of ground-water vulnerability because it has been suggested that it can be used as an indicator of overall ground-water quality (U.S. Environmental Protection Agency, 1996a). In addition, nitrate is a USEPA-regulated contaminant and its presence in high concentrations is a potential health risk. The USEPA has set a standard for nitrate in drinking water of 10 mg/L (U.S. Environmental Protection Agency, 1996b).

Explanatory (Independent) Variables

Land cover and geology type have been related previously to the occurrence of elevated nitrate concentrations in ground water throughout the Mid-Atlantic Region (Ator and Ferrari, 1997; Nolan 2001). Land-cover data for the Mid-Atlantic Region were derived from the National Land Cover Dataset (NLCD). The NLCD was produced as a cooperative effort between the USGS and the USEPA to produce a consistent, land-cover dataset for the conterminous United States using early 1990s Landsat Thematic Mapper (TM) data purchased by the Multi-Resolution Land Characterization (MRLC) Consortium (U.S. Geological Survey, 2000). There are 15 classes of land cover that represent the Mid-Atlantic area. These 15 classes were first aggregated into 6 similar classes (water, barren, wetland, developed, cultivated, and forested), then to 3 groups (developed, cultivated, and forested/wetland). Comparison of logistic-regression analyses confirmed no loss in significance between models developed using six land-cover classes and models that were further aggregated into just three land-cover groups representing developed, cultivated, and forested/wetland. Therefore, the logistic-regression models developed for this study included land-cover groups representing developed, cultivated, and forested/wetland. These land-cover groups were developed as continuous variables with units equal to percentage of land cover (table 1, fig. 2).

Geologic type was compiled from the digital version of the geology of the conterminous United States (King and Biekman, 1974; Schruben and others, 1994) and the Mid-Atlantic Coastal Plain geologic framework (Ator and others, in press). The geologic framework of the Coastal Plain (generalized into coarse unconsolidated and fine unconsolidated material), as developed by Ator and others (in press), replaced the Coastal Plain area described by King and Biekman (1974). Geologic units were included in the statistical model as nominal categorical variables identified as Siliciclastic, Carbonate, Crystalline, coarse Coastal Plain, and fine Coastal Plain (table 1, fig. 3).

The other variables considered by the statistical model were nutrient (nitrate) source inputs consisting of application of manure and fertilizer (kilogram per square kilometer) and atmospheric deposition (kilogram per square kilometer). Population density and soils data consisting of hydrologic groups, organic matter, depth of soil layer, depth to water table, percent sand, percent silt, and percent clay also were explored by the statistical model to determine if they were significant explanatory variables (table 1).

Manure input was derived from county-based estimates of nitrogen content of animal manure in the United States for 1992 (Puckett and others, 1998). Fertilizer input was compiled from nitrogen-fertilizer sales for the conterminous United States in 1991 by county (Battaglin and Goolsby, 1994). Atmospheric deposition of nitrate was derived from 1992 data collected by the National Atmospheric Deposition Program (NADP) (National Atmospheric Deposition Program, 2000). Manure and fertilizer-input values were distributed evenly across agricultural areas of each county (cultivated land-cover variable of the model) (table 1).

Population density was derived from 1990 block group point data compiled by the USGS (Hitt, 1992). This variable was explored as a density value and as a natural log transformation. Hydrologically relevant soil characteristics were compiled by the USGS from the State Soil Geographic (STATSGO) Data Base produced by the U.S. Department of Agriculture's Natural Resources Conservation Service (Schwartz and Alexander, 1995).

Well depth has been shown to be an important variable in explaining elevated nitrate levels in ground water (Hallberg and Keeney, 1993; Nolan, 2001; Tesoriero and Voss, 1997). For this study, however, models were developed to predict the probability of nitrate concentrations exceeding a specified management threshold in areas where there are little or no data on nitrate concentrations. Because well depth does not exist in these areas, this would not be a helpful explanatory variable and was therefore not included in the statistical models.

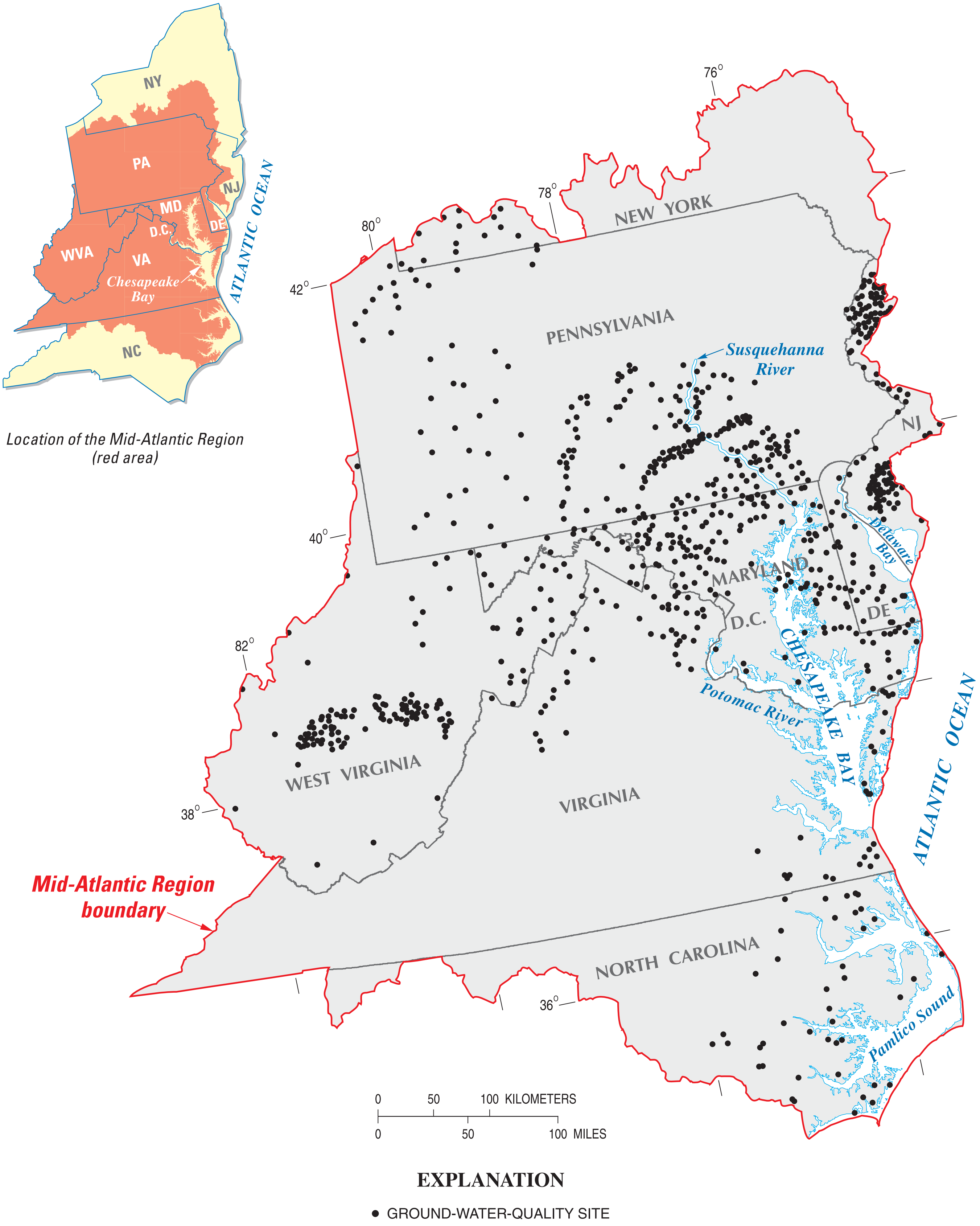


Figure 1. Location of ground-water wells sampled for nitrate in the Mid-Atlantic Region.

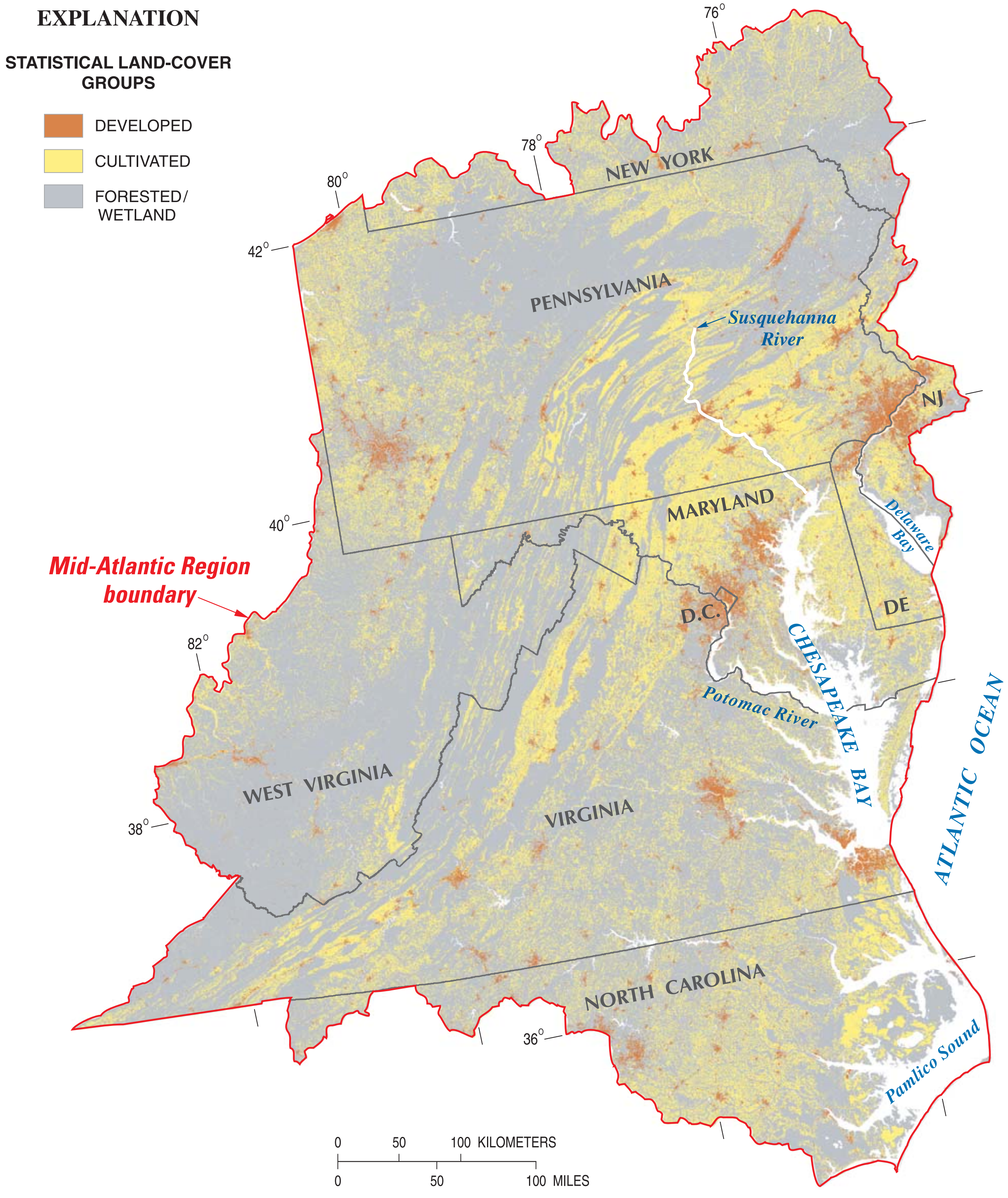


Figure 2. Land cover statistically combined into three groups in the Mid-Atlantic Region.

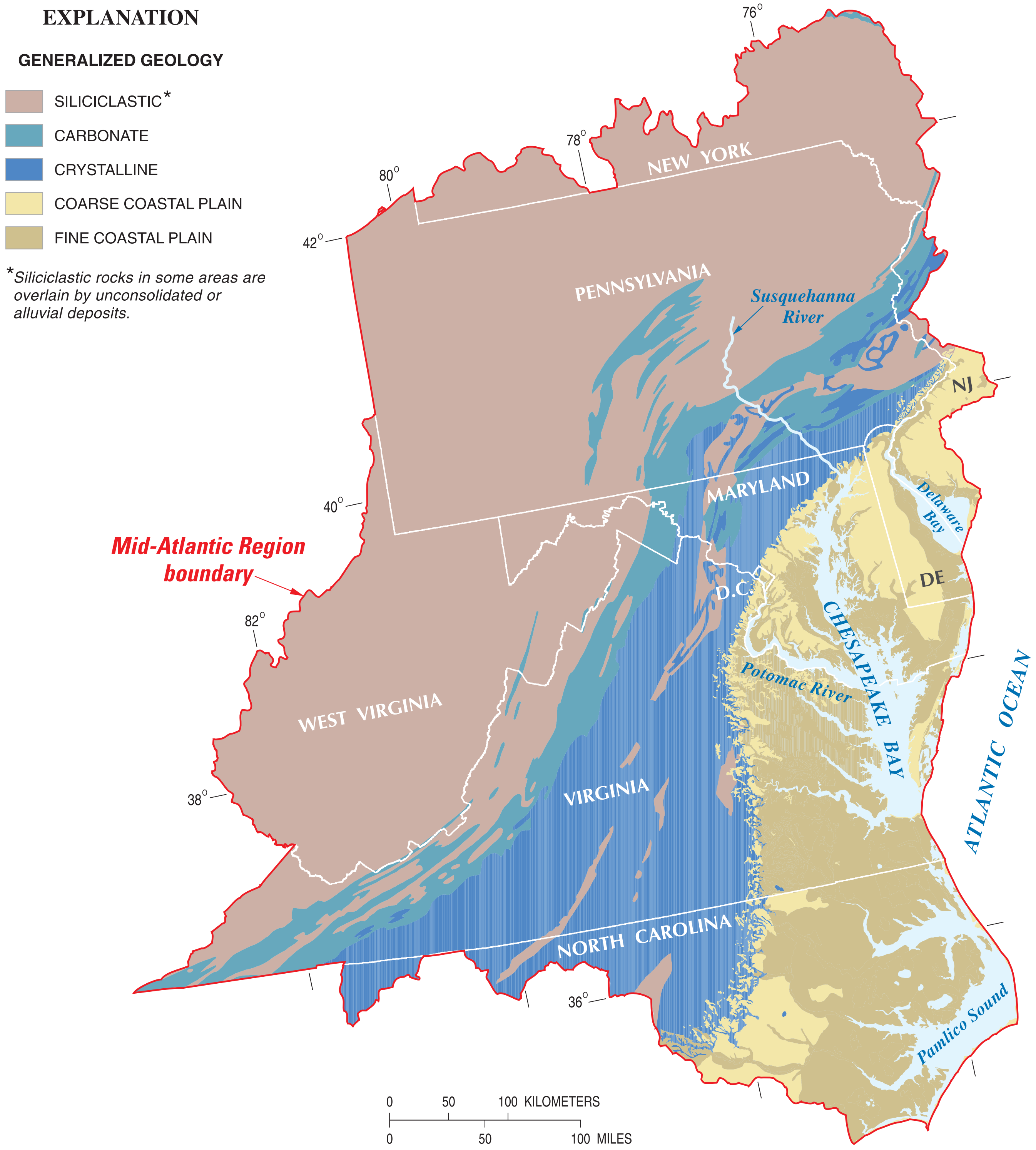


Figure 3. Generalized geology classified into five categories in the Mid-Atlantic Region (modified from King and Biekman, 1974; and Ator and others, in press).

Table 1. *Explanatory variables and their descriptive statistics explored in the logistic-regression models for the Mid-Atlantic Region*

[NA, not applicable; %, percent; kg, kilogram; km^2, square kilometer]

Variable name	Data type	Units or Categories	Minimum	Median	Maximum	Quartiles	
						25%	75%
Cultivated land cover	Continuous	Percentage	0	46	100	18	71
Developed land cover	Continuous	Percentage	0	9	84	0.1	5.4
Forested/Wetland land cover	Continuous	Percentage	0	33	100	16	59
Geology	Categorical	Siliciclastic	NA	NA	NA	NA	NA
		Carbonate	NA	NA	NA	NA	NA
		Crystalline	NA	NA	NA	NA	NA
		Coarse Coastal Plain	NA	NA	NA	NA	NA
		Fine Coastal Plain	NA	NA	NA	NA	NA
Soil permeability	Categorical	Moderate	NA	NA	NA	NA	NA
		Moderately rapid	NA	NA	NA	NA	NA
		Rapid	NA	NA	NA	NA	NA
Soil organic matter	Continuous	Percentage	0.1	0.6	52	.5	0.9
Depth of soil layer	Continuous	Meters	.7	1.5	1.8	1.2	1.6
Depth to water table	Continuous	Meters	.1	1.5	1.8	1.1	1.6
Soil texture	Continuous	Ratio (% silt and clay to % sand)	.1	.3	0.9	.2	.5
Hydrologic Group A	Continuous	Percentage	0	5	65	0	11
Population density	Continuous	People/km^2	1	40	3,022	17	98
Inorganic fertilizer application	Continuous	kg/km^2	0	1,314	1,765	590	3,556
Manure application	Continuous	kg/km^2	0	968	15,864	282	2,367
Atmospheric deposition	Continuous	kg/km^2	213	505	887	429	554

Logit Model

General linear models (GLM) of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_i x_i + \varepsilon_i \qquad (1)$$

have their basis in classical statistics, and are used to predict the response of a dependent variable (y) from a single or group of independent explanatory (x_i) variables. The intercept is β_0 and the β_i's are the slope coefficients. The unexplained variance is termed the error and is symbolized as ε_i (Ott, 1977). When the dependent variable is dichotomous (above/below, yes/no), linear regression is no longer an option and other multivariate statistical methods need to be considered. Rather than modeling the actual response with the explanatory variables, as in linear regression, the Logit model (logistic-regression model) is based on the assumption that the log-odds (logit) of the model are related to the independent variables (Brown, 1998).

Unlike linear regression, the logistic-regression procedure models a change in the logit or log-odds of the dependent variable rather than a change in the variable itself. Whereas ordinary least squares (OLS) regression attempts to minimize the distance squared between the regression line and the data points, logistic regression applies the “maximum likelihood estimation.” The maximum likelihood estimation attempts to maximize the log likelihood (LL), which reflects how likely it is (the odds) that the observed values of the dependent variable (y) may be predicted from the observed values of the independent variables (x_i).

The response probability, p_i, of the Logit model can be defined as

$$p_i = P(Y=1 \mid X_i) = \frac{e^{(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}}{1 + e^{(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}} \quad (2)$$

where, for each of the n observations, there is a response Y_i that can take on the values of 1 or 0, from corresponding independent explanatory variables X_i, $i = 1, 2, \ldots, n,$ where $X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$. β_k represents the regression coefficient corresponding to the kth explanatory variable, with β_0 representing the intercept. The vector $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ is called the vector of slope parameters. Equation (2) can be simplified further by dividing both the numerator and denominator by the numerator itself to obtain a form of the logit equation (equation 3) with a desired property of anything that is substituted for X_i, and β_i, and the probability p_i, will always be between 0 and 1.

$$p_i = P(Y=1 \mid X_i) = \frac{1}{1 + e^{(-\beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})}} \quad (3)$$

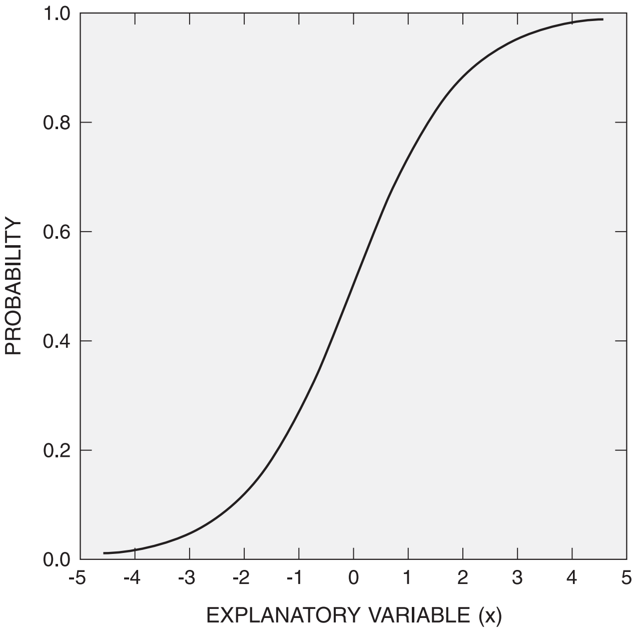


Figure 4. Logit model for a single explanatory variable (from Allison, 1999).

If there is a single explanatory variable (x) with $\beta_0 = 0$, and $\beta_1 = 1$, the simplest logit model, the equation can be graphed to produce the classic S-shaped curve shown in figure 4. As x increases or decreases, the probability p_i, always will be between 1 and 0. In addition, the graph of the S-shaped curve of the logit model shows that the probability response of a unit change in the x variable varies depending on the starting probability value. When the probability is near 0.50, a unit change in the x variable produces a large probability response; when the probability is near 0 or 1, a unit change in the x variable produces a small probability response (fig. 4).

Logistic regression was used in this investigation to identify a dichotomous response between dependent variables (for example, when the dependent variable (y) is in a binary format representing above or below) and independent variables (x) that either are continuous or categorical. Specifically, logistic-regression equations where explanatory variables (x) are used to predict the presence of nitrate concentrations (y) above a specified management threshold value (such as Maximum Contaminant Levels, or MCLs) were developed. The resulting equations then were transformed to determine the probability, $P(Y)$, of estimates of nitrate concentrations above a selected threshold value. The presence of elevated concentrations of nitrate in ground water was used as a surrogate for anthropogenic contamination of shallow ground water in the Mid-Atlantic Region.

Logistic regression allows nitrate concentrations to be evaluated at any threshold level, thereby transforming nitrate from a censored continuous variable measured in milligrams per liter to a categorical variable with classes. Understanding thresholds is important when applying logistic-regression equations to water-quality data. After a threshold has been established, sampled values are classified as either “1’s” when they are above the threshold, or “0’s” when they are below the threshold (fig. 5). These threshold levels can be set up to parallel decision-making criteria, making this modeling approach a useful decision tool for resource managers. Therefore, it is possible to predict the probability of nitrate concentrations in ground water exceeding a certain pre-determined threshold level.

Significance Testing of β Coefficients

The use of the maximum likelihood estimation (MLE) is the preferred way to test the significance of the logit coefficients (β) of the logistic-regression model (Allison, 1999; Hosmer and Lemeshow, 1989). Essentially, this test considers the null hypothesis H_o: $\beta_i = 0$, $i = 1,2,\ldots,k$, indicating that the regression coefficients β_i are not significantly different from zero, against the alternative H_a: $\beta_i \neq 0$, that at least one coefficient is different from zero. The MLE compares the ratio of the maximum of the likelihood under H_o to the maximum of the likelihood under H_a (likelihood ratio test) and seeks to maximize the log likelihood (LL) , which reflects how likely (the odds) that the observed values of the dependent variable (y) are predicted from the observed values of the independent variables (x’s).

The likelihood ratio test (G) statistic for inclusion of an independent variable can be defined as

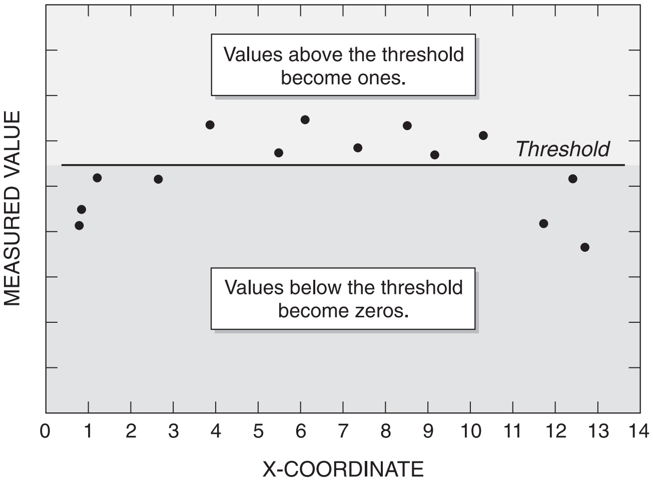


Figure 5. Transformation of measured values to ones and zeros based on a threshold value. *(Measured values above the threshold become a value of one; measured values below the threshold become zero.)*

$$G = -2 \ln \left[\frac{\textit{likelihood without the variables}}{\textit{likelihood with the variables}} \right] \qquad (4)$$

where the G statistic has a chi-square distribution with degrees of freedom depending on the number of parameters included under the null hypothesis that all β coefficients are zero. A significant G statistic means that adding the independent variables to the model, as compared to a model with only an intercept (β_0) with its corresponding β_i coefficient, improves the model (for example, at least one of the β_i coefficients is nonzero) (Hosmer and Lemeshow, 1989). To test the significance of each β_i coefficient, the model is refitted multiple times, comparing each explanatory variable in turn, to the intercept-only model.

Additional testing to evaluate the significance of each of the explanatory variables for each logistic-regression model was based on the Wald statistic. The Wald statistic was calculated for each of the β_i coefficients of the model. This test statistic (equation 5) is computed as the coefficient divided by its standard error estimate of the coefficient, with the result squared.

$$\textit{Wald Statistic} = \left[\frac{\hat{\beta}_i}{\textit{standard error } \hat{\beta}_i} \right]^2, \qquad (5)$$

$$\textit{for } i = 0, 1, 2, \ldots, k$$

This statistic, when compared to a chi-square random variable with 1 degree of freedom, was used to determine if each coefficient was significant in the model. Even though both tests were used to determine the significance of the β_i coefficients, some researchers have suggested that the likelihood ratio tests are superior to the Wald test (Hauck and Donner, 1977; Jennings, 1986; Collett, 2002).

Model Goodness-of-Fit

For this investigation, two statistical tests were used to assess the model's goodness-of-fit. The first test used was the G statistic or model chi-square test and the second test was the "Hosmer and Lemeshow's Goodness of Fit Test" (Hosmer and Lemeshow, 1989). The model chi-square test is a likelihood ratio test (equation 4), which reflects the difference between the error not including the explanatory variables (initial chi-square) and the error when the explanatory variables are included (deviance). The model chi-square follows a chi-square distribution with degrees of freedom equal to the difference in the number of explanatory variables included compared to the intercept-only model.

"Hosmer and Lemeshow's Goodness of Fit Test" (H-L) (equation 6) tests the hypothesis that the data fit the logistic-regression model. Based on the logistic-regression model, predicted probabilities associated with each observation are sorted and grouped into 10 intervals (deciles). Within each interval, the expected frequency is determined by adding the predicted probabilities. The expected frequencies are compared with the observed frequencies by the Pearson chi-square with 8 degrees of freedom (number of intervals minus 2) at the $\alpha = .05$ level of significance. If the H-L test statistic's p-value is greater than 0.05, it is implied that the model's estimates fit the data at an acceptable level. As with most model goodness-of-fit tests, as the sample size increases, the power to detect differences from the null hypothesis (no difference) based on the statistic improves (Hosmer and Lemeshow, 1989; Hosmer and others, 1997).

The H-L test statistic is:

$$\textit{Hosmer-Lemeshow test statistic} = \qquad (6)$$

$$\hat{C} = \sum_{i=1}^{10} \frac{(O_i - N_i \pi_i)^2}{N_i \pi_i (1 - \pi_i)}$$

where

N_i = the number of observations in the i^{th} decile,

O_i = the number of successes $(Y=1)$ in the i^{th} decile,

π_i = the average of the estimated probabilities, and

$\hat{C}$ = test statistic, approximated by the chi-square distribution with g-2 degrees of freedom.

Statistical Model Development

The statistical model was developed by first testing each single explanatory variable in a logistic-regression model to determine significant variables. Models that consisted of a set of single variable logistic-regression equations were analyzed. The explanatory variables that were significant individually then were tested in a multivariate logistic regression to determine how each contributed to the overall model when considered with other variables.

Area of Well Influence

Several studies have shown there is a significant relation between shallow ground-water quality (nitrate concentrations) and the types of land use near a sampled well (Cain and others, 1989; Hay and Battaglin, 1990; Barringer and others, 1990; Tesoriero and Voss, 1997). Using different types of statistical procedures (linear regression, rank correlations, logistic regression), these investigations all showed that the size of the land use/cover buffer around a sampled well is significant when evaluating ground-water quality in relation to land use/cover.

In developing the logistic-regression models for explanatory variables, the land cover within an optimal radius was used to establish an area of influence around each well (fig. 6). The same area of well influence that was determined for land cover was applied to all explanatory variables that were a function of area. Explanatory variables that were dependent on the spatial area around the well (values varied in space) included land cover, soil permeability, soil organic matter, depth of soil layer, depth to water table, clay content of the soil, silt content of the soil, and hydrologic groups.

The optimal land-cover radius was determined by examining how well the logistic-regression models at different radii fit the nitrate data sampled at each well. Using land-cover data, logistic-regression models for 8 radii that ranged from 500 to 4,000 m in 500-m steps were tested. The best-fit model was determined by finding the radii that maximized the likelihood ratio test (G) and H-L test statistics. The best-fit model that maximized the test statistics for nitrate concentrations exceeding a threshold of 3 mg/L was at 1,500 m (fig. 7). Repeating this analysis for thresholds of 1 mg/L through 10 mg/L indicated that the test statistics also were maximized at a radius of 1,500 m. Results of this analysis indicate that the radius that contributes the most information to the model using land cover is at an optimal radius of 1,500 m regardless of the threshold level that is chosen.

Single Variable Models

To develop logistic-regression models that relate an event of an occurrence of nitrate concentrations above a predetermined threshold to an explanatory variable, various variables were tested individually to determine if they were significant predictors. Single-variable models that were considered and tested included land cover, surficial geology, soil permeability, soil organic matter, depth of soil layer, depth to water table, soil texture, hydrologic groups, manure, fertilizer, atmospheric deposition, and population density.

Regression coefficients and statistics for models developed using these variables are presented in table 2. All explanatory variables were found to be significant ($\alpha = 0.05$), except for Hydrologic Group A at the lower thresholds and atmospheric deposition for all thresholds. Even though Hydrologic Group A was not significant at the lower thresholds, this variable was included in the multivariate model because at higher thresholds, it was significant. Atmospheric deposition was dropped as an explanatory variable for the multivariate models at all thresholds.

Multivariate Model

The multivariate logistic-regression equations combined the results in direct relations between all tested explanatory variables and the probability of exceeding nitrate concentrations at a selected threshold value. The logistic-regression models were tested using goodness-of-fit statistical procedures (likelihood ratio test (G) and H-L test statistics) to define the final logistic-regression models at multiple thresholds.

Stepwise Logistic Regression

When building the logistic-regression model, the use of a stepwise selection procedure facilitated the development of models at multiple thresholds. The development of the models in a sequential fashion allowed for the examination of a collection of models that otherwise would not have been examined. A statistical algorithm that adds or removes variables from the model based on a pre-determined decision rule was used to develop the model. The decision rule usually is based on the statistical significance of the variable’s model coefficient. At any step in the procedure, the most important variable (in statistical terms) was the one that produced the greatest change in the log-likelihood relative to the model not containing the variable.

The stepwise selection procedure actually is a modification of the simpler forward selection procedure. With forward selection, the initial model begins with only the intercept and variables are added to the model one at a time. If the associated variable is considered significant at the $\alpha = 0.2$ level of significance, it is entered into the model. One or more backward elimination steps follow the forward selection by examining the maximum likelihood estimates for each variable in the model. Those variables that are not significant at the $\alpha = 0.05$ level of significance are removed from the model. The process continues until no more variables can offer a change in the log-odds, and therefore can not be added to or removed from the model.

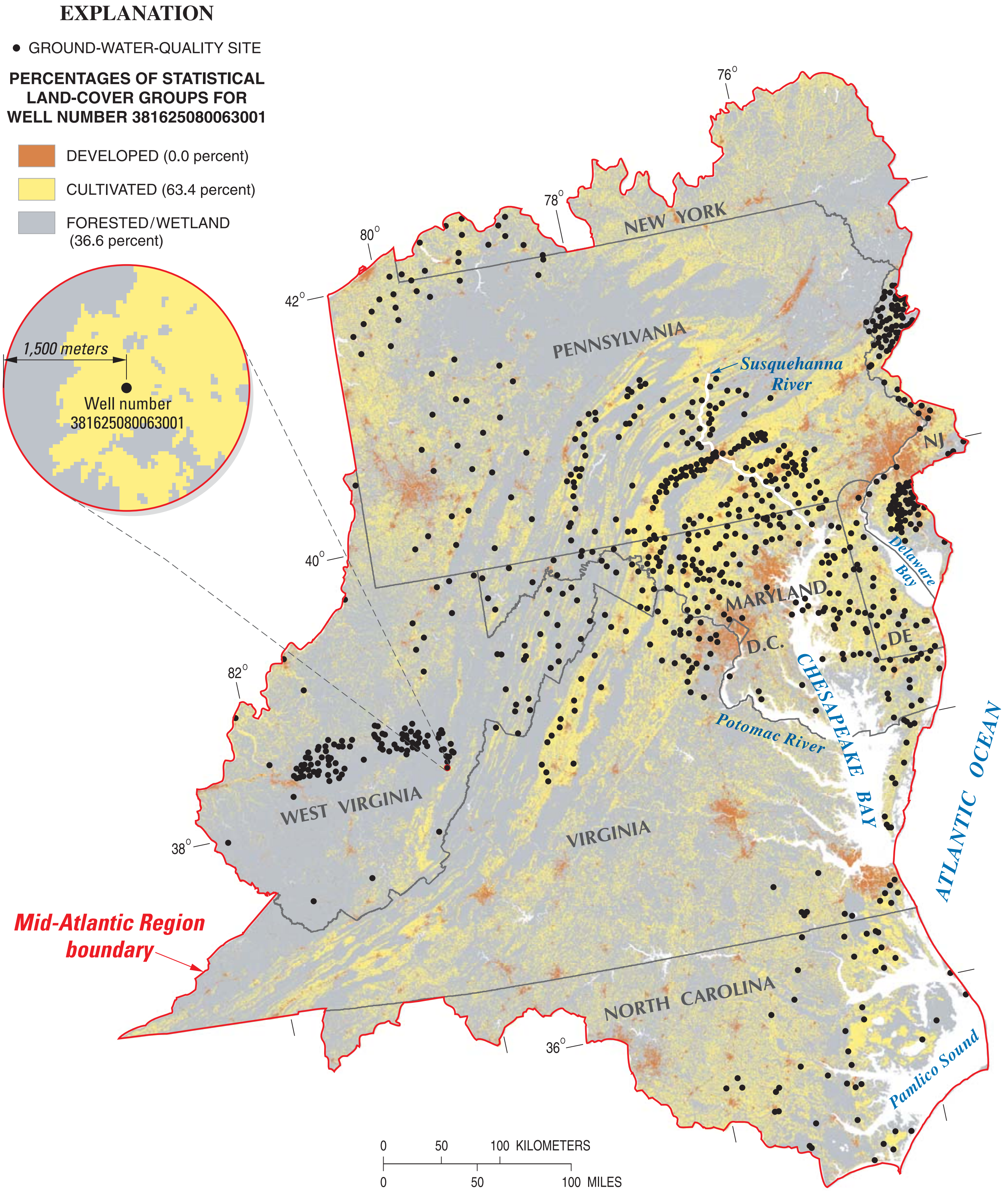


Figure 6. The extraction of land-cover variables within a statistical area of influence around each ground-water well sampled in the Mid-Atlantic Region.

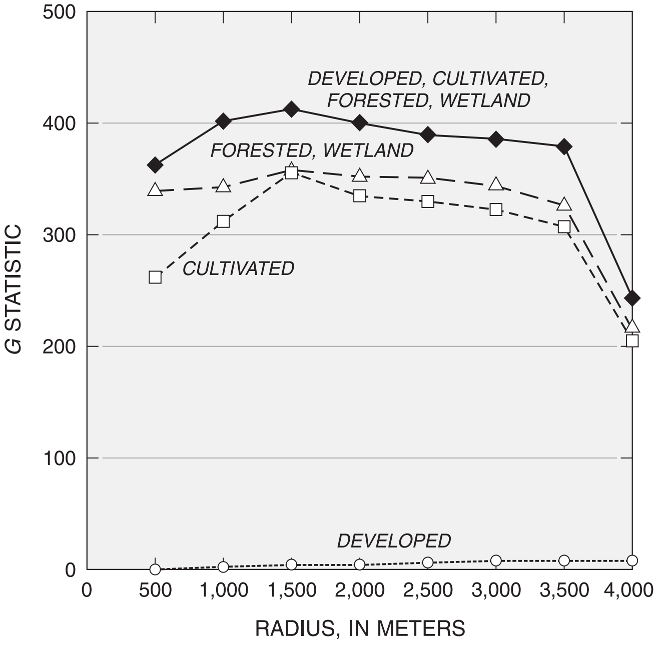


Figure 7. Optimal radius for the best-fit model using land-cover variables for nitrate concentrations in ground water exceeding a threshold of 3 milligrams per liter as nitrogen for the Mid-Atlantic Region.

Multicollinearity and Interaction Terms

In developing general linear and logistic-regression models, multicollinearity and interaction among the explanatory variables must be examined. Multicollinearity occurs when there are strong dependencies among the explanatory variables. This interaction among variables will mean that the odds ratio for explanatory variables may vary with the value of another explanatory variable. It is possible to have an interaction in which no pairs of variables are highly correlated, but various variables together may be highly interdependent (Allison, 1999).

Interaction among explanatory variables was addressed early in the model-forming stage. All interaction terms were evaluated before any individual variable was eliminated. All explanatory variables that had significant interaction terms were included in the model. The final multivariate model variables, coefficients, and interaction terms for significant explanatory variables ($p \leq 0.05$) that were included in the model are presented in table 3. Cultivated land cover and surficial geology were the only significant variables for all thresholds. Other variables were significant at certain thresholds and were only included in the model when appropriate (table 3).

Table 2. *Likelihood ratio tests for single variable logistic-regression models for the predicted probability of nitrate concentrations in ground water from the Mid-Atlantic Region exceeding management thresholds of 1 milligram per liter, 3 milligrams per liter, and 10 milligrams per liter as nitrogen*

[mg/L, milligram per liter; <, less than]

Variable	Threshold 1 mg/L		Threshold 3 mg/L		Threshold 10 mg/L	
	Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value
Land cover (including developed, cultivated, and forested/wetland)	415.6207	<0.0001	412.2763	<0.0001	202.5157	<0.0001
Geology	262.3879	< .0001	210.7784	< .0001	94.5039	< .0001
Soil permeability	39.3112	< .0001	43.7395	< .0001	37.1519	< .0001
Soil organic matter	11.5722	< .0001	39.7203	< .0001	20.7725	< .0001
Depth of soil layer	180.5545	< .0001	193.2178	< .0001	89.6187	< .0001
Depth to water table	22.5801	< .0001	30.7165	< .0001	13.5250	< .0001
Soil texture, including percent silt and clay	18.4296	< .0001	24.2340	< .0001	23.0355	< .0001
Hydrologic Group A	0.7826	.3764	2.8388	.0920	24.3645	< .0001
Population density	24.2443	< .0001	8.1261	.0044	10.0032	.0016
Inorganic fertilizer application	132.3953	< .0001	137.1383	< .0001	43.6472	< .0001
Manure application	76.1945	< .0001	96.3165	< .0001	100.2533	< .0001
Atmospheric deposition	.0038	.9508	0.4737	0.4913	.0675	.7949

Table 3. *Final multivariate model significant variables, coefficients, and interaction terms for the predicted probability of nitrate concentrations in ground water from the Mid-Atlantic Region exceeding thresholds of 1 milligram per liter through 10 milligrams per liter as nitrogen*

[4.13408 Coefficients significant when *p*-values are less than or equal to 0.05; ▨ Coefficients not significant when *p*-values are greater than 0.05]

Threshold, in milligrams per liter	VARIABLES												
	Intercept	**Cultivated land cover**	**Developed land cover**	**Forested / wetland land cover**	**Geology 1 (Siliciclastic / Fine Coastal Plain)**	**Geology 2 (Carbonate / Fine Coastal Plain)**	**Geology 3 (Crystalline / Fine Coastal Plain)**	**Geology 4 (Coarse Coastal Plain / Fine Coastal Plain)**	**Depth of soil layer**	**Soil organic matter**	**Soil texture**	**Hydrologic Group A**	**Manure application**
1.0	-11.93477	4.13408	4.46423		3.76149	3.32044	3.59361	-46.97709	0.13493		3.78615	-1.49253	
2.0	-7.73812	4.17508	3.64069		-0.96083	0.31189	-0.05322	-0.56985	0.08869		4.06430	-4.24487	
3.0	-10.11441	4.61047	3.90884		-1.16910	-1.00821	-0.25770	1.25243	0.08782		4.32576	-2.58499	
4.0	-5.88696	5.43187	2.26244		-0.41940	0.75934	0.34462	0.04731	0.07164	-0.35688	-6.63612		
5.0	-4.23939	3.72503		-2.17038	-0.37520	0.84054	0.26953	0.05820	0.06848	-0.40537	-6.83051		
6.0	-6.33372	5.33130			-0.62890	0.55199	-0.18910	0.48216	0.06316	-0.41058	-4.88176		
7.0	-7.42080	5.98300			-0.69071	0.62048	-0.30996	0.49358	0.06956	-0.37862	-5.65256		
8.0	-5.59087	5.88522			-1.07735	0.06230	-0.11394	1.07914					0.00010
9.0	-5.90109	5.65248			-1.17667	0.28074	-0.05287	1.21497					0.00011
10.0	-5.82783	5.10470	-9.98260		-1.44675	0.30028	0.23513	1.25856					0.00015

Threshold, in milligrams per liter	INTERACTION TERMS												
	Cultivated land cover * Geology 1	**Cultivated land cover * Geology 2**	**Cultivated land cover * Geology 3**	**Cultivated land cover * Geology 4**	**Cultivated land cover * Hydrologic Group A**	**Developed land cover * Geology 1**	**Developed land cover * Geology 2**	**Developed land cover * Geology 3**	**Developed land cover * Geology 4**	**Geology 1 * Depth of soil layer**	**Geology 2 * Depth of soil layer**	**Geology 3 * Depth of soil layer**	**Geology 4 * Depth of soil layer**
1.0	0.93658	1.53188	0.19431	1.02639	10.07229	4.57696	0.34336	-0.58182	3.19501	-0.08571	-0.06456	-0.05284	0.78621
2.0	1.55053	0.59546	0.97702	1.24037	13.23667	3.38802	-1.35698	0.16497	2.30817				
3.0	1.51478	1.42262	0.61235	-0.37405	10.52195								
4.0													
5.0													
6.0													
7.0													
8.0													
9.0													
10.0													

Evaluation of Model Performance

Predictive statistical models need to be evaluated in order to establish that the model works satisfactorily for data that were not used in model development (Altman and Royston, 2000). In developing logistic-regression models, various researchers have suggested that calibration, discrimination, and validation should be used to evaluate the predictive model (Altman and Royston, 2000; Hosmer and others, 1997; Justice and others, 1999; Mittlbock and Schemper, 1996; Schemper and Stare, 1996).

Calibration

During the calibration step, the models are assessed to determine whether the predicted probabilities agree with the observed probabilities. Two model calibrations were developed—the first looked at whether the predictions agreed on average with the observed probabilities, and the second investigated the calibration in more detail and the agreement between the predictions and observations over the entire range of predictions.

An example of the final calibration based on the method developed by Hosmer and Lemeshow (1980), Lemeshow and Hosmer (1982), and Hosmer, Lemeshow, and Klar (1988) is shown in figure 8. The actual observations (observed) compared to the percent predicted (expected) probabilities for nitrate concentrations exceeding management thresholds of 3 mg/L, for a model with only land cover, and for a final model with all significant variables, are presented in this figure. The 1:1 line in figure 8 illustrates the ideal calibration. A simple measure of fit (calibration) used in this investigation was the slope of the regression line for the observed and predicted values within a decile of risk. When the predictions were greater than the observations, then the slope was less than one. If the predictions were less than the observations, then the slope was greater than one. This form of the "mis-calibration" was compared statistically to a model with an intercept equal to 0 and a slope of 1 (Harrell and others, 1980) and the best-fit model was developed for each threshold.

Discrimination

The ability of the predictive model to distinguish between sites with nitrate concentrations greater than the threshold and sites with nitrate concentrations less than the threshold was quantified using the measure of concordance (c statistic). The c statistic ranges between 0.5 and 1 for good models (Allison, 1999). The closer the c statistic is to 1, the better the model is at distinguishing the correct binary outcome. The c statistic for the final predictive model at each of the thresholds is presented in table 4.

Model Validation

When developing general linear models, it is important to consider a class of statistical tools that are measures of how well the model can predict the response of a dependent variable (y) from a single or group of independent explanatory (x_i) variables. This type of assessment is often called model validation. This assessment is a different statistical measure than the model goodness-of-fit statistic. It is possible to have a model that predicts the dependent variable very well, but has a poor goodness-of-fit statistic (H-L test statistic). It is also possible that a model will have a very good goodness-of-fit statistic but a very poor prediction statistic. In developing general linear models, statistics such as R^2 are used to estimate the proportion of the variance "explained" by the independent (explanatory) variables. In logistic regression, there is not a simple statistic that is equivalent to the R^2 in linear regression. Although an R^2 can be calculated, and is an estimate of how much variance is explained, it is specific to the particular data set and cannot be used to compare between different models. In logistic regression, three measures can be used in place of the R^2, such as the D statistic, the Brier score, or the PRESS statistic, which was used for this investigation (Hosmer and Lemeshow, 1989; SAS Institute, Inc., 1990).

The PRESS statistic can be interpreted as the power of model predictability and can be used for model validation (SAS Institute, Inc., 1990). The PRESS statistic is defined as the prediction sum of squares and its purpose is to evaluate the regression model. A PRESS statistic for ordinary regression is available with SAS Institute, Inc. (1990) software, but there is not a similar algorithm for logistic regression. Therefore, as part of this investigation, a PRESS statistic was developed for use in logistic regression for model validation.

Theoretically, the computation of the PRESS statistic begins by taking the i^{th} observation out of the sample, building the model from the remaining n–1 observations, and then evaluating the model using the explanatory variables from the i^{th} observation. The i^{th} PRESS residual is calculated by subtracting this new predicted value from the actual value for the dependent variable. Calculating and summing these residuals over all n sample data points gives the PRESS statistic for the model. The best model can be considered the one with the smallest PRESS statistic value.

A logistic-regression model predicts the probability of an event occurring rather than the value of the dependent variable directly. To make the PRESS statistic procedure valid for logistic regression, each predicted probability was transformed into a 1 or 0 whether it was likely to have exceeded the threshold or not. When the predicted probability was greater than or equal to 0.5, the value was interpreted as a 1 for that threshold, otherwise it was taken to be 0. These transformed values are used to estimate the probability of misclassification. After comparing these values to the actual values for that threshold, the PRESS statistic results can be calculated as a count of the number of "correctly predicted" observations. To measure the prediction power of the model, the "correctly predicted" observations are divided by the total number of observations to give a percentage of correctly predicted terms.

The PRESS statistic was used to determine the ability of each model to correctly predict the occurrence of nitrate concentrations above a predetermined threshold value. The PRESS statistic results for each threshold are listed in

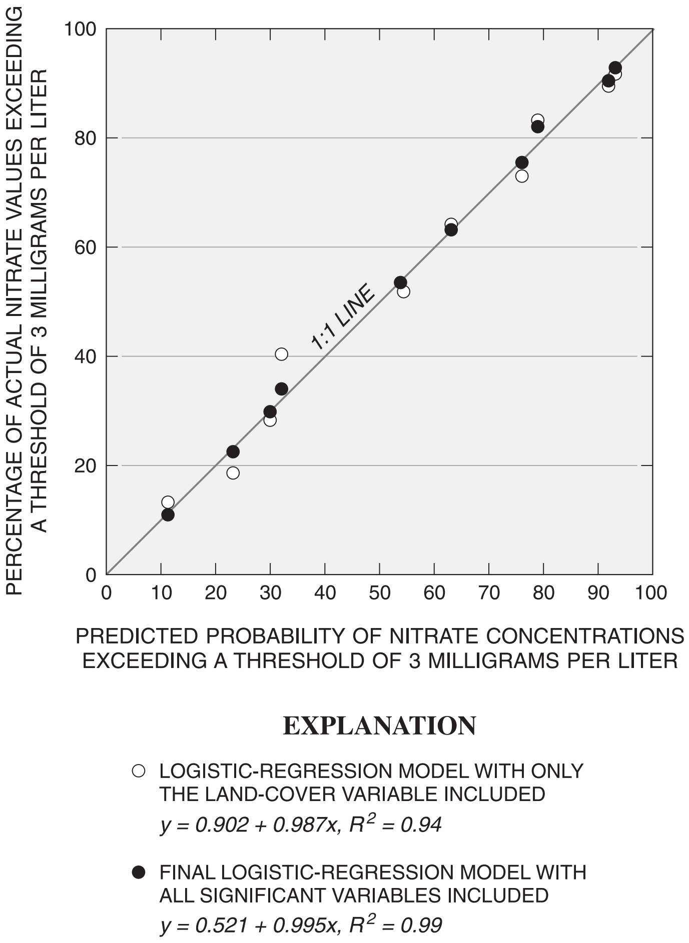


Figure 8. Example calibration of the best-fit model comparing the predicted (expected) probability of nitrate concentrations in ground water to the percentage of the actual (observed) nitrate concentrations exceeding a threshold of 3 milligrams per liter as nitrogen for the Mid-Atlantic Region.

table 5. This statistic illustrates that even at the lower thresholds, the model is correctly predicting nitrate concentrations above the threshold about 81 percent of the time, and generally increases to about 89 percent at the higher thresholds (table 5).

Additional statistics based on the ordinal measures of association also can be used to measure the predictive power of the model. The ordinal measures (Somer's D and Gamma) were considered and maximized when building the statistical model. Both measures range between 0 and 1, and the values closer to 1 correspond to stronger associations between the predicted and observed values (Allison, 1999; SAS Institute, Inc., 1990). The ordinal measures for the final model at each of the thresholds are listed in table 6.

Table 4. *The c statistic results for each multivariate logistic-regression model to predict the probability of nitrate concentrations in ground water from the Mid-Atlantic Region exceeding thresholds of 1 milligram per liter through 10 milligrams per liter as nitrogen*

[mg/L, milligram per liter]

Thresholds	c statistic
1 mg/L	0.902
2 mg/L	.900
3 mg/L	.895
4 mg/L	.893
5 mg/L	.893
6 mg/L	.893
7 mg/L	.903
8 mg/L	.899
9 mg/L	.893
10 mg/L	.908

Table 5. *PRESS statistic results for each multivariate logistic-regression model to predict the probability of nitrate concentrations in ground water from the Mid-Atlantic Region exceeding thresholds of 1 milligram per liter through 10 milligrams per liter as nitrogen*

[mg/L, milligram per liter]

Thresholds	PRESS statistic (decimal percent of correctly predicted observations)
1 mg/L	0.813
2 mg/L	.814
3 mg/L	.823
4 mg/L	.822
5 mg/L	.815
6 mg/L	.829
7 mg/L	.856
8 mg/L	.861
9 mg/L	.875
10 mg/L	.891

Table 6. *Ordinal measure results for each multivariate logistic-regression model to predict the probability of nitrate concentrations in ground water from the Mid-Atlantic Region exceeding thresholds of 1 milligram per liter through 10 milligrams per liter as nitrogen*

[mg/L, milligram per liter]

Thresholds	Somer's D	Gamma
1 mg/L	0.804	0.805
2 mg/L	.800	.801
3 mg/L	.789	.790
4 mg/L	.786	.787
5 mg/L	.786	.787
6 mg/L	.786	.787
7 mg/L	.806	.807
8 mg/L	.777	.778
9 mg/L	.786	.788
10 mg/L	.817	.818

Confidence Interval

Confidence intervals for the best-fit logistic-regression model for each threshold were evaluated in order to obtain the range in values with a specified probability (95 percent) containing the regression coefficients. Two types of confidence interval maps were developed as part of this investigation—one mapping the length of the 95-percent confidence interval, and one displaying the upper bound of the interval, or the worst-case probability within a 95-percent confidence interval. The frequency for the predicted probabilities compared to the length of the confidence intervals for these predictions for nitrate in ground water exceeding the 1 mg/L, 3 mg/L, and 10 mg/L management thresholds is shown in figure 9. The lengths of the confidence intervals gradually became smaller (increase in confidence) as the thresholds increase. Thus, a smaller confidence-interval length indicates a greater confidence in the prediction, whereas a larger length depicts less confidence in the prediction.

Analysis of Variable Effects

A Wald Type III analysis was performed on each model to determine the relative importance of each significant variable to the overall logistic-regression model at each threshold (Hosmer and Lemeshow, 1989). The histogram of the statistics associated with each significant variable at each threshold from 1 mg/L through 10 mg/L is shown in figure 10. The Wald Type III "Analysis of Effects" includes variables that are categorical, continuous, and have interaction terms to determine the relative importance of each variable to the model. Each threshold analysis corresponds to 100 percent of the Type III Wald Statistic with each variable representing the weighted proportion of the statistic.

The variables that have the most effect or that are most significant to the model are cultivated land cover and geology. For the lower thresholds, developed land cover and depth of soil layer are significant explanatory variables. At the upper thresholds, the variable manure application becomes a significant explanatory variable for the model.

Ground-Water Vulnerability to Nitrate Contamination in the Mid-Atlantic Region

Geographic information system (GIS) methods were used to generate an array of regularly spaced points on a 1,500-m grid representing the Mid-Atlantic Region. These points represent the center of each grid cell and correspond to the center of the 1,500-m radius that was determined to be the best-fit model from the land-cover analysis. Logistic-regression model coefficients that were developed from the explanatory variables were applied to each point on the grid, and maps were developed that illustrate the probability of nitrate concentrations exceeding thresholds of 1 mg/L through 10 mg/L. These maps show the vulnerability of ground water to nonpoint nitrate contamination at these thresholds (fig. 11).

Along with the probability maps of ground-water vulnerability in the Mid-Atlantic Region, the uncertainty of the probability estimates in the form of confidence maps was generated (fig. 11). The length of the confidence interval helps discern the error in the probability estimate. The upper bound of the confidence interval for each estimate can be thought of as the worst-case scenario, or highest possible probability, of nitrate concentrations in ground water exceeding a certain threshold within a 95-percent confidence interval.

The predicted probability of nitrate concentrations exceeding management thresholds of 1 mg/L, 3 mg/L, and 10 mg/L, along with the associated upper bound (worst-case scenario) of the 95-percent confidence interval of these predictions are shown in figures 11a–f. At management thresholds of 1 mg/L and 3 mg/L, the probability of nitrate concentrations exceeding these levels is low throughout much of the Mid-Atlantic Region. Near the northern part of the Chesapeake Bay, however, the probability of nitrate concentrations exceeding management thresholds of 1 mg/L and 3 mg/L is high (greater than 0.5). This area includes much of central Maryland, southeastern and northwestern Pennsylvania, and the Delmarva Peninsula. In addition, extensive areas in North Carolina and Virginia also show high probabilities of nitrate concentrations in ground water exceeding management thresholds of 1 mg/L and 3 mg/L (figs. 11a–d). The mapped areas showing a high predicted probability of

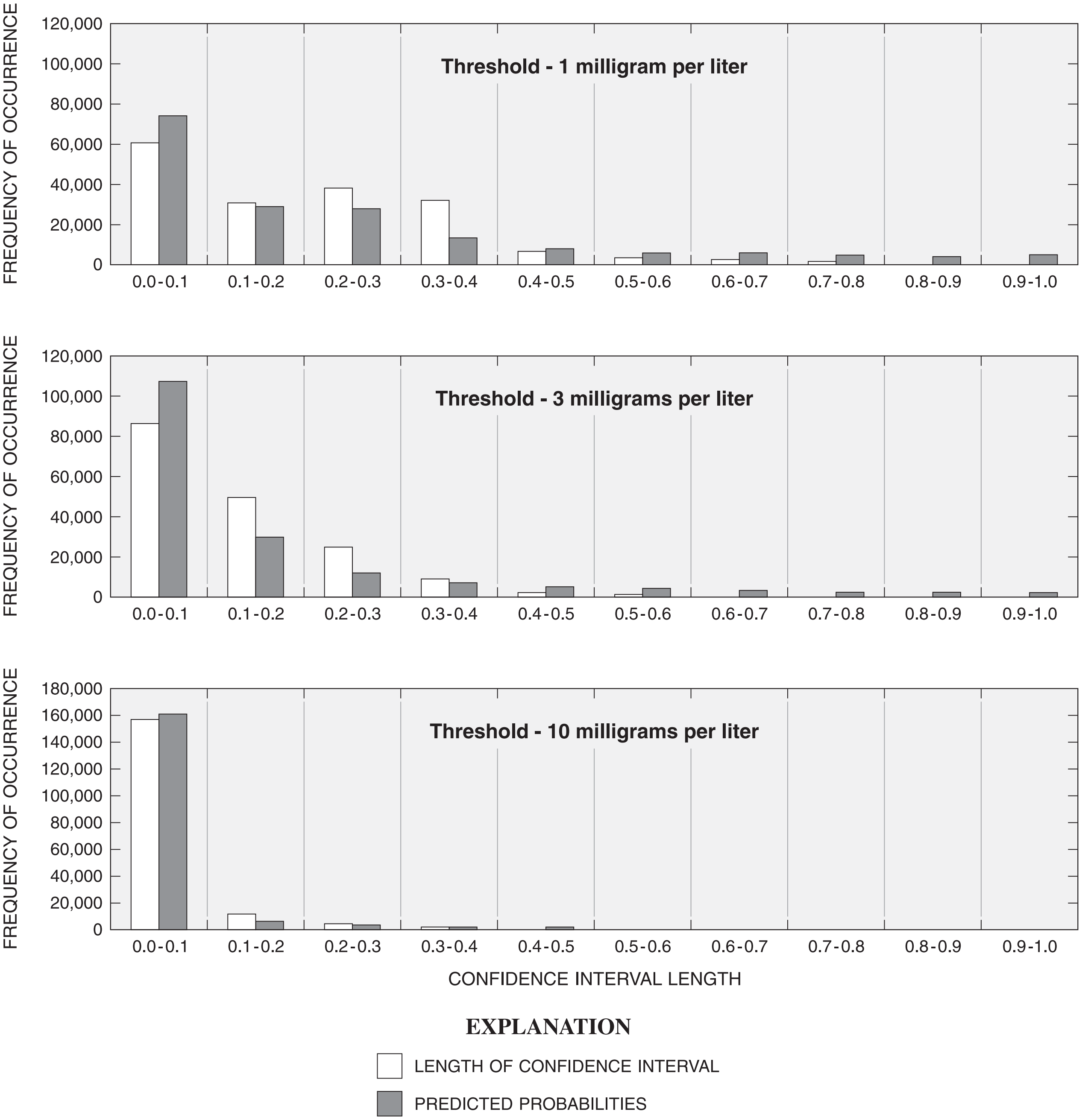


Figure 9. Frequencies of the confidence interval length and corresponding predicted probabilities of nitrate concentrations in ground water from the Mid-Atlantic Region exceeding management thresholds of 1 milligram per liter, 3 milligrams per liter, and 10 milligrams per liter as nitrogen.

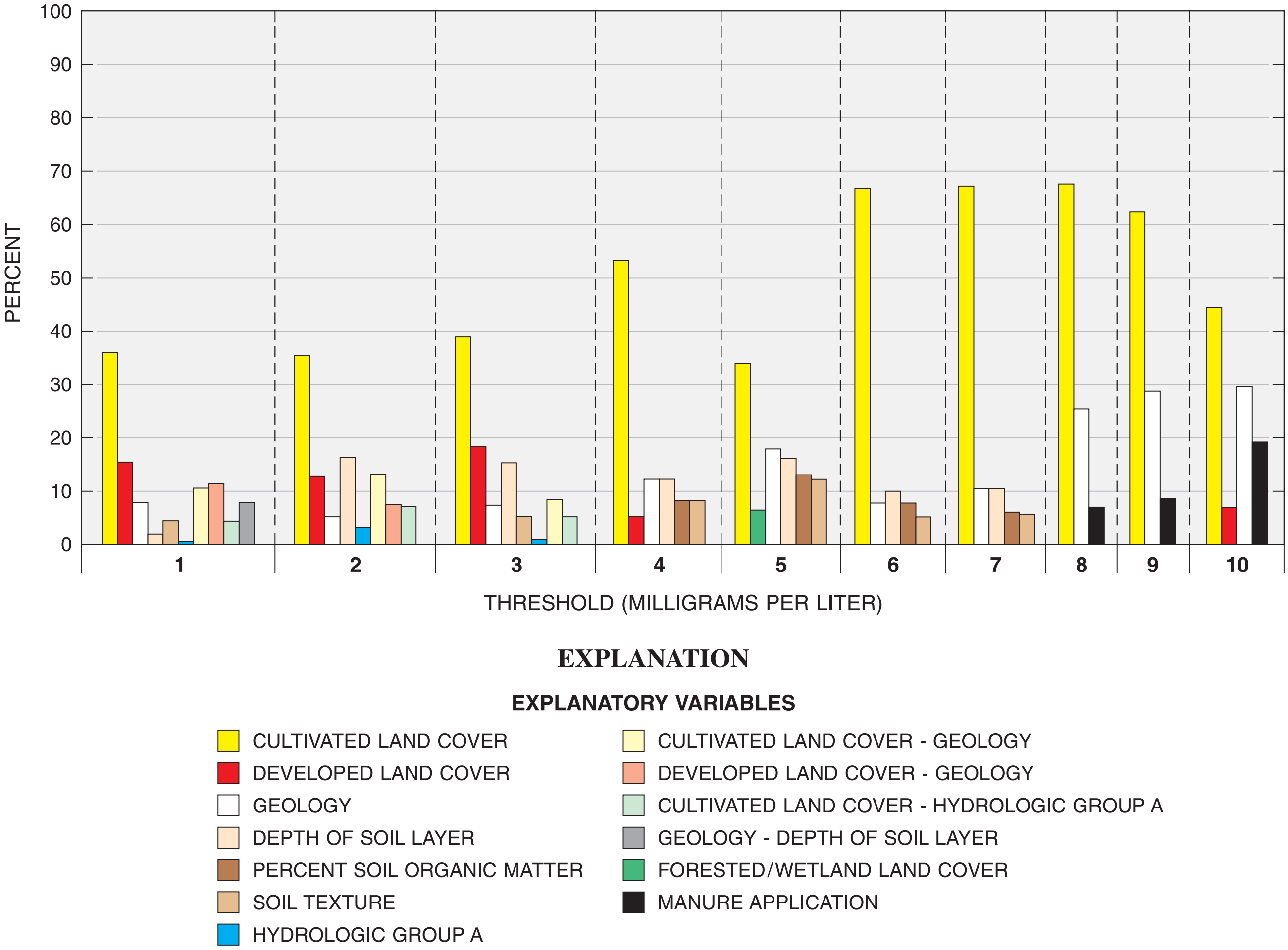


Figure 10. Wald Type III analysis for relative importance of explanatory variables' contribution to the model at each threshold for ground water in the Mid-Atlantic Region.

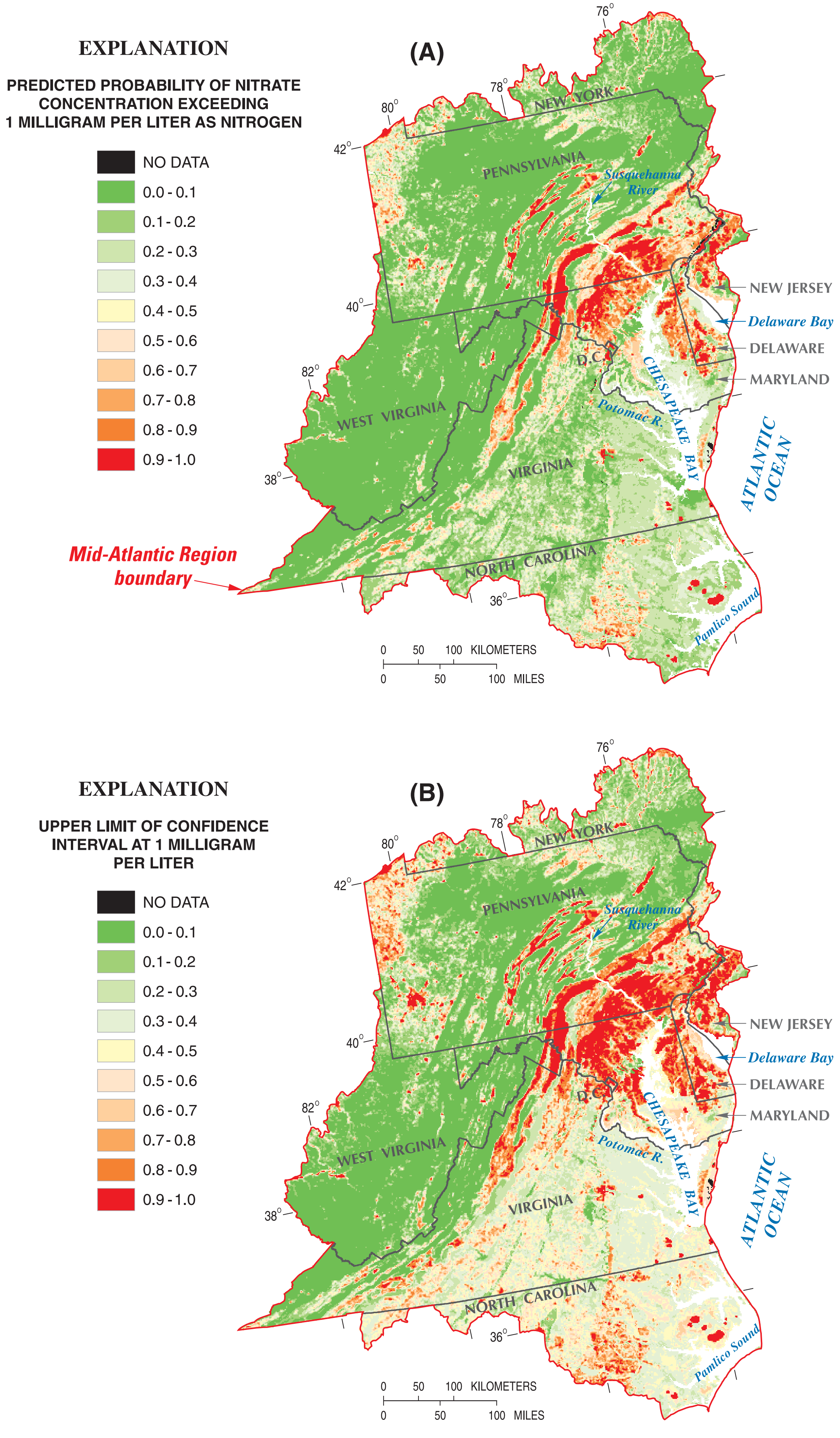


Figure 11. **(A)** Probability of nitrate concentrations in ground water exceeding 1 milligram per liter as nitrogen, and **(B)** upper limit of the confidence interval at 1 milligram per liter in the Mid-Atlantic Region.

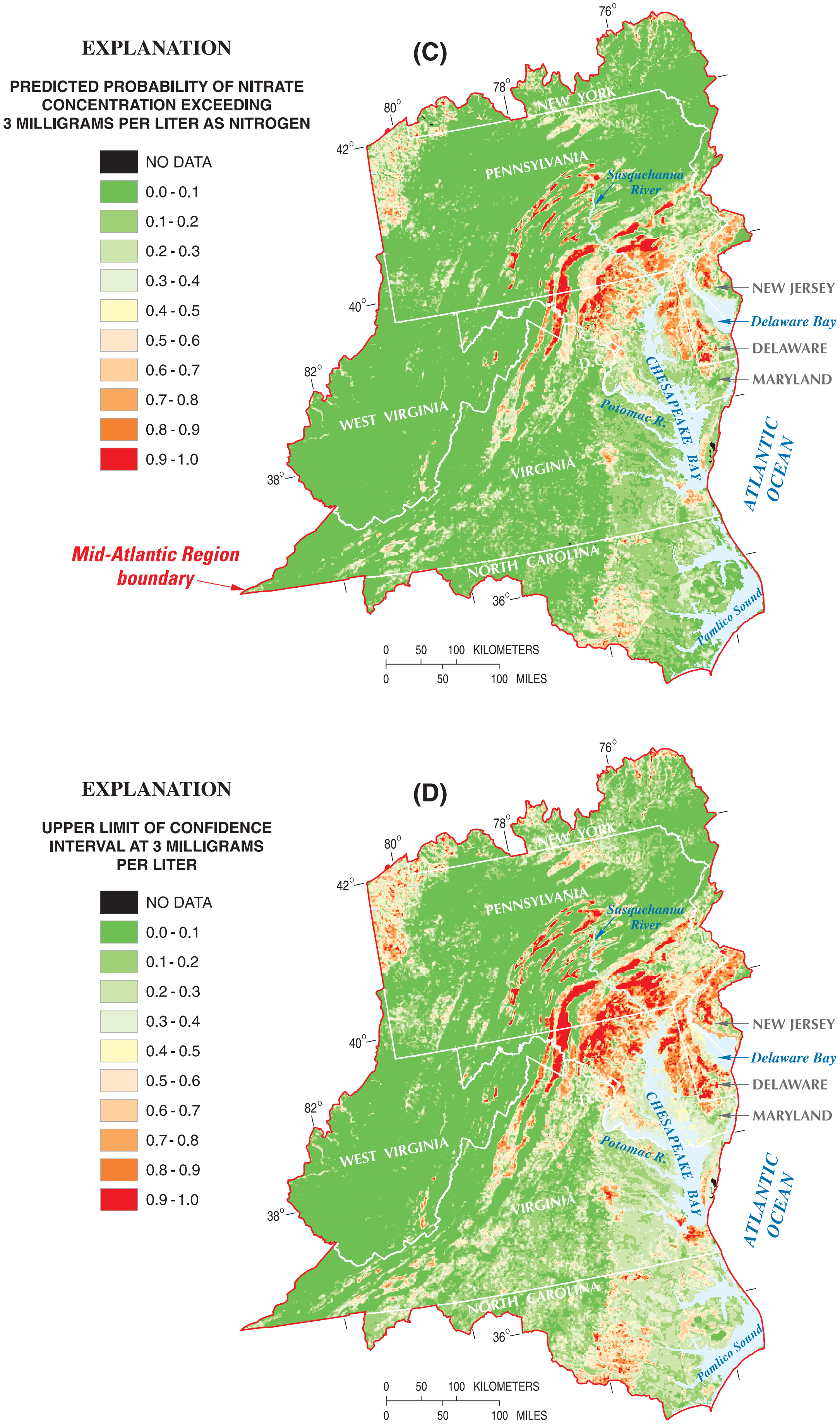


Figure 11. **(C)** Probability of nitrate concentrations in ground water exceeding 3 milligrams per liter as nitrogen, and **(D)** upper limit of the confidence interval at 3 milligrams per liter in the Mid-Atlantic Region. -- Continued

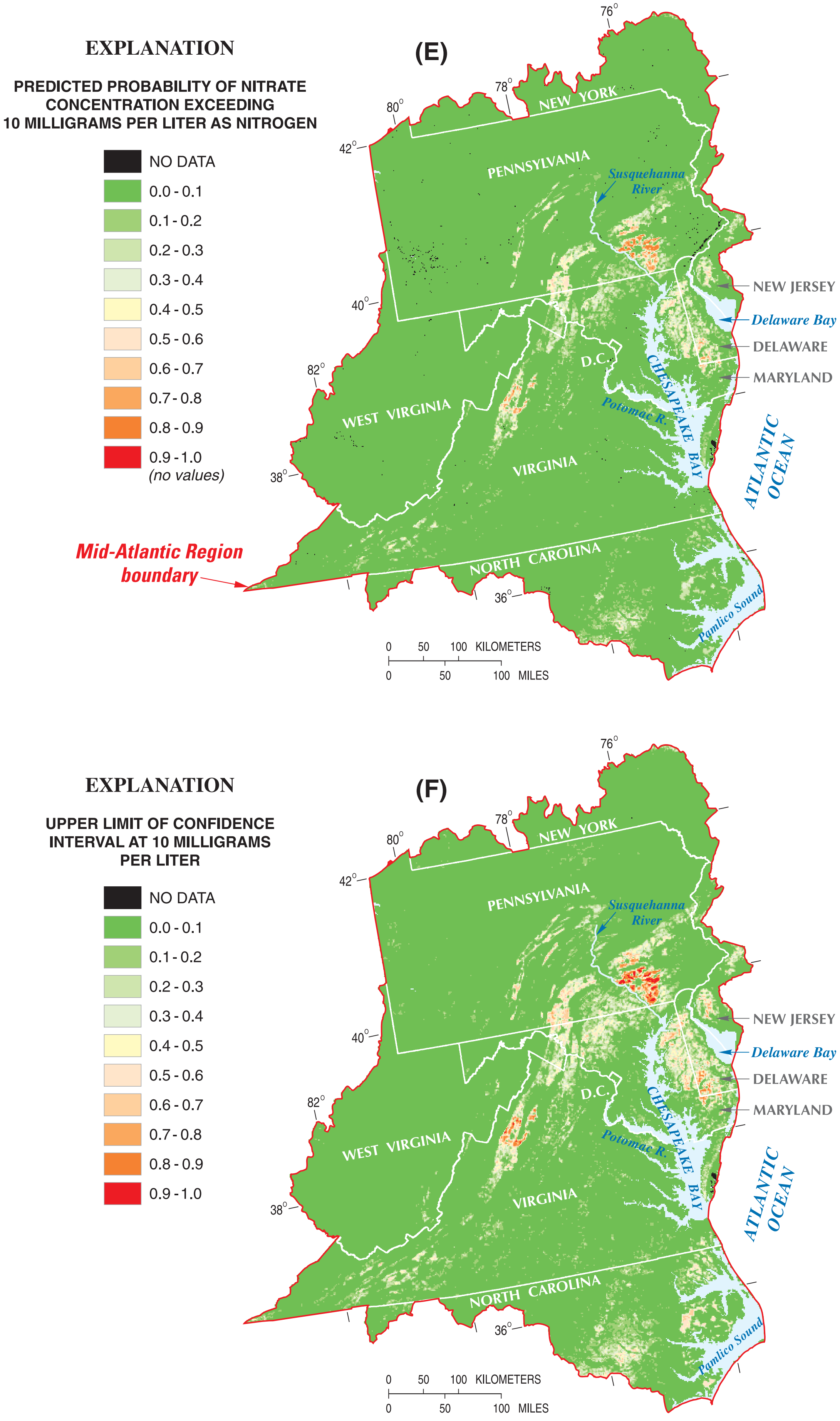


Figure 11. **(E)** Probability of nitrate concentrations in ground water exceeding 10 milligrams per liter as nitrogen, and **(F)** upper limit of the confidence interval at 10 milligrams per liter in the Mid-Atlantic Region. -- Continued

nitrate concentrations in ground water exceeding 1 mg/L and 3 mg/L correspond to areas that are mapped as cultivated land cover (fig. 2) and/or overlying areas mapped as carbonate rocks (fig. 3).

Using a management threshold of 10 mg/L (corresponding to the USEPA standard for nitrate in drinking water of 10 mg/L), the predicted probability of nitrate concentrations in ground water exceeding this level is low (less than 0.5) for most of the Mid-Atlantic Region. Areas within the Delmarva Peninsula and southeastern Pennsylvania, in addition to the mapped carbonate areas in Virginia, Maryland, and Pennsylvania, have high probabilities (greater than 0.5) of nitrate in ground water exceeding 10 mg/L (figs. 11e–f).

Summary

The U.S. Environmental Protection Agency has recognized the need for regional, State, and local water managers in the Mid-Atlantic Region to assess ground-water-quality conditions at a regional scale for environmental and human health purposes. The U.S. Geological Survey, in cooperation with the U.S. Environmental Protection Agency's Regional Vulnerability Assessment Program, has developed a set of statistical probability models to characterize the relation between ground-water quality and geographic factors in the Mid-Atlantic Region to support regional-scale, integrated ecological risk-assessment studies. These models can be used to assess the risk of nonpoint-source contamination of ground water in areas of the Mid-Atlantic Region where little data are available at multiple management thresholds.

Logistic-regression models provide a statistical link between a dependent variable (nitrate concentration in ground water) and explanatory variables. Ground-water data for 927 sites sampled by the U.S. Geological Survey National Water-Quality Assessment Program and other projects were used to develop logistic-regression models to quantify the relation between geographic factors (land cover, geology, soils, and others) and the probability of nitrate in ground water exceeding a specified management concentration threshold. Thresholds can be formulated and specified as management values, such as 3 milligrams per liter for environmental concerns or 10 milligrams per liter for regulating drinking water.

A geographic information system in combination with logistic regression was used to establish the area of land-cover influence around each ground-water sample site. The influence of land cover was analyzed by developing logistic-regression equations at various radii (areas) to find the model that best described this influence. The radius (area) that best described the influence of land cover on nitrate concentrations was 1,500 meters.

Each of the other explanatory variables were all developed within this 1,500-meter radius and tested to determine the significance of including these variables in the logistic-regression model. In some cases, the explanatory variable within this 1,500-meter radius was a continuous value (inorganic fertilizer application); in other instances, it was a weighted percentage (land cover) or a categorical variable (geology).

Individual explanatory variables that were significant were combined into an overall logistic-regression model using a stepwise selection process that developed an appropriate combination of variables to maximize the model's performance, and by using a goodness-of-fit test incorporating Hosmer-Lemeshow, Wald, and PRESS statistics. In addition, multicollinearity and interaction terms were analyzed.

Geographic-information-system methods were used to generate an array of regularly spaced points to map the variables on a 1,500-meter grid representing the Mid-Atlantic Region. These points represent the center of each grid cell, and correspond to the center of the 1,500-meter radius that was determined to be the best-fit model from the land-cover analysis. Logistic-regression model coefficients that were developed from the explanatory variables were applied to each point, and maps that determined the probability of nitrate concentrations exceeding management thresholds of 1 milligram per liter through 10 milligrams per liter as nitrogen were developed.

In addition to the probability maps showing ground-water vulnerability to nitrate contamination in the Mid-Atlantic Region, uncertainty in the form of the confidence maps was generated and displayed. The upper bound of the confidence interval for each prediction of the probabilities of nitrate in ground water exceeding a threshold of 1 milligram per liter through 10 milligrams per liter was developed. Using these bounds, the worst-case scenario, or highest possible probability, of nitrate concentrations in ground water exceeding a certain management threshold was displayed.

The probability of nitrate concentrations exceeding management thresholds of 1 milligram per liter and 3 milligrams per liter is extensive (greater than 0.5) throughout much of the Mid-Atlantic Region, especially near the northern part of the Chesapeake Bay. This area includes much of Maryland, southeastern and northwestern Pennsylvania, and the Delmarva Peninsula. In addition, extensive areas in North Carolina and Virginia also show high probabilities of nitrate concentrations in ground water exceeding these management thresholds. At a management threshold of 10 milligrams per liter (corresponding to the U.S. Environmental Protection Agency standard for nitrate in drinking water of 10 milligrams per liter), the predicted probability of nitrate concentrations in ground water exceeding this level is low (less than 0.5) for most of the Mid-Atlantic Region. Areas in the Delmarva Peninsula and southeastern Pennsylvania, along with the mapped carbonate areas in Virginia, Maryland, and Pennsylvania, show a higher predicted probability (greater than 0.5) of nitrate concentrations exceeding 10 milligrams per liter.

References Cited

Allison, P.D., 1999, Logistic regression using the SAS system—Theory and application: Cary, N.C., SAS Institute, Inc., 288 p.

Altman, D.G., and Royston, P., 2000, What do we mean by validating a prognostic model?: Statistics in Medicine, v. 19, issue 4, p. 453–473.

Ator, S.W., 1998, Nitrate and pesticide data for waters of the Mid-Atlantic Region: U.S. Geological Survey Open-File Report 98–158, 5 p.

Ator, S.W., Denver, J.M., Krantz, D.E., Newell, W.L., and Martucci, S.K., in press, A surficial hydrogeologic framework for the Mid-Atlantic Coastal Plain: U.S. Geological Survey Professional Paper 1680.

Ator, S.W., and Ferrari, M.J., 1997, Nitrate and selected pesticides in ground water of the Mid-Atlantic Region: U.S. Geological Survey Water-Resources Investigations Report 97–4139, 8 p.

Barringer, Thomas, Dunn, D., Battaglin, W., and Vowinkel, E., 1990, Problems and methods involved in relating land use to ground-water quality: Water Resources Bulletin, v. 26, no. 1, p. 1–9.

Battaglin, W.A., and Goolsby, D.A., 1994, Spatial data in geographic information system format on agricultural chemical use, land use, and cropping practices in the United States: U.S. Geological Survey Water-Resources Investigations Report 94–4176, 87 p.

Brown, C.E., 1998, Applied multivariate statistics in geohydrology and related sciences: New York, Springer, 248 p.

Cain, Douglas, Helsel, D.R., and Ragone, S.E., 1989, Preliminary evaluations of regional ground-water quality in relation to land use: Ground Water, v. 27, no. 2, p. 230–244.

Collet, David, 2002, Modelling binary data, 2^{d} ed.: London, England, Chapman and Hall/CRC, 408 p.

Eckhardt, D.A., and Stackelberg, P.E., 1995, Relation of ground-water quality to land use on Long Island, New York: Ground Water, v. 33, no. 6, p. 1,019–1,033.

Ferrari, M.J., Ator, S.W., Blomquist, J.D., and Dysart, J.E., 1997, Pesticides in surface water of the Mid-Atlantic Region: U.S. Geological Survey Water-Resources Investigations Report 97–4280, 12 p.

Gilliom, R.J., Alley, W. M., and Gurtz, M.E., 1995, Design of the National Water-Quality Assessment Program—Occurrence and distribution of water-quality conditions: U.S. Geological Survey Circular 1112, 33 p.

Hallberg, G.R., and Keeney, D.R., 1993, Nitrate, *in*: Alley, W.M., (ed.), Regional ground-water quality: New York, Van Nostrand Reinhold, p. 297–322.

Harrell, F.E., Lee, K.L., and McKinnis, R.A., 1980, Proceedings for large regression problems requiring maximum likelihood estimation, *in* Proceedings of the Fifth Annual SAS Users Group International Conference, Cary, N.C., SAS Institute, Inc., p. 1,031–1,036.

Hauck, W.W., and Donner, A., 1977, Walds test as applied to hypotheses in logit analysis: Journal of the American Statistical Association, v. 72, p. 851–853.

Hay, L.E., and Battaglin, W.A., 1990, Effects of land-use buffer size on Spearman's rank partial correlations of land use and shallow ground-water quality: U.S. Geological Survey Water-Resources Investigations Report 89–4163, 28 p.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources: New York, Elsevier, 522 p.

Hitt, K.J., 1992, Digital map file of 1990 population and housing data for the United States: U.S. Geological Survey, digital data accessed on December 27, 2002, at URL *http://water.usgs.gov/lookup/getspatial?uspop90*

Hosmer, D.W., Hosmer, T., Le Cressie, S., and Lemeshow, S., 1997, A comparison of goodness-of-fit tests for the logistic regression model: Statistics in Medicine, v., 16, p. 965–980.

Hosmer, D.W. and Lemeshow, S. 1980, A goodness-of-fit test for the multiple regression model: Communications in Statistics, A10, p. 1,043–1,069.

_____ 1989, Applied logistic regression: New York, Wiley and Sons, 307 p.

Hosmer, D.W., Lemeshow, S., and Klar, J. 1988, Goodness-of-fit testing for multiple regression analysis when the estimates of probabilities are small: Biometrical Journal, v. 30, p. 911–924.

Jennings, D.E., 1986, Judging inference adequacy in logistic regression: Journal of the American Statistical Association, v. 81, p. 471–476.

Justice, A.C., Covinsky, K.E., and Berlin, J.A., 1999, Assessing the generalizability of prognostic information: Annual Internal Medicine, v. 130, p. 515–524.

King, P.B., and Biekman, H.M., 1974, Geologic map of the United States: U.S. Geological Survey, 3 sheets, scale 1:2,500,000.

Lemeshow, S., and Hosmer, D.W., 1982, The use of goodness-of-fit statistics in the development of logistic regression models: American Journal of Epidemiology, v. 115, p. 92–106.

Mittlbock, M., and Schemper, M., 1996, Explained variation for logistic regression: Statistics in Medicine, v. 15, p. 1,987–1,997.

National Atmospheric Deposition Program (NRSP-3)/ National Trends Network, 2000, NADP Program Office: Champaign, Ill., Illinois State Water Survey, 16 p.

Nolan, B.T., 2001, Relating nitrogen sources and aquifer susceptibility to nitrate in shallow ground waters of the United States: Ground Water, v. 39, no. 2, p. 290–299.

Ott, Lyman, 1977, Introduction to statistical methods and data analysis: North Scituate, Mass., Duxbury Press, 730 p.

Puckett, L.J., Hitt, K.J., and Alexander, R.B., 1998, County-based estimates of nitrogen and phosphorus content of animal manure in the United States for 1982, 1987, and 1992: U.S. Geological Survey, digital data, accessed December 20, 2002 at URL *http://water.usgs.gov/lookup/getspatial?manure*

SAS Institute, Inc., 1990, SAS User's Guide: Statistics, 6th ed., Cary, N.C., SAS Institute, Inc. 1,686 p.

Schemper, M., and Stare, J., 1996, Explained variation in survival analysis: Statistics in Medicine, v. 15, p. 1,999–2,012.

Schruben, P.G., Arnt, R.E., Bawiec, W.J., King, P.B., and Biekman, H.M., 1994, Geology of the conterminous United States at 1:2,500,000 Scale—A digital representation of the 1974 P.B. King and H.M. Biekman Map: U.S. Geological Survey Digital Data Series DDS–11, U.S. Geological Survey, Reston, Va.

Schwartz, G.E., and Alexander, R.B., 1995, State Soil Geographic (STATSGO) Data Base for the conterminous United States: U.S. Geological Survey Open-File Report 95–449, digital data accessed on December 27, 2002, at URL *http://water.usgs.gov/lookup/getspatial?ussoils*

Spalding, R.F., and Exner, M.E., 1993, Occurrence of nitrate in groundwater—A review: Journal of Environmental Quality, v. 22, p. 392–402.

Tesoriero, A.J., and Voss, F.D., 1997, Predicting the probability of elevated nitrate concentrations in the Puget Sound Basin: Implications for aquifer susceptibility and vulnerability: Ground Water, v. 39, no. 2, p. 1,029–1,039.

U.S. Environmental Protection Agency, 1996a, Environmental indicators of water quality in the United States: Washington, D.C., U.S. Environmental Protection Agency, Office of Water, EPA 841-R-96-002, 25 p.

_____ 1996b, Drinking water regulations and health advisories: Washington, D.C., U.S. Environmental Protection Agency, Office of Water, EPA 822-B-96-002, 11 p.

U.S. Geological Survey, 2000, National Land Cover Dataset: U.S. Geological Survey Fact Sheet 108–00, 1 p.

Walker, S.H., and Duncan, D.B., 1967, Estimation of the probability of an event as a function of several independent variables: Biometrika, v. 54, p. 167–179.