EVALUATION OF STATISTICAL MODELS TO PREDICT CHEMICAL QUALITY OF SHALLOW GROUND WATER IN THE PINE BARRENS OF SUFFOLK COUNTY, LONG ISLAND, NEW YORK

by Paul E. Stackelberg and Steven F. Siwiec

U.S. GEOLOGICAL SURVEY

Water-Resources Investigations Report 92-4100

Prepared in cooperation with the

LONG ISLAND REGIONAL PLANNING BOARD

Syosset, New York

1993

U.S. DEPARTMENT OF THE INTERIOR

BRUCE BABBITT, Secretary

U.S. Geological Survey

Dallas L. Peck, Director

For additional information
write to:

U.S. Geological Survey
5 Aerial Way
Syosset, NY 11791

Copies of this report may be
purchased from:

U.S. Geological Survey
Books and Open-File Reports
P.O. Box 25425
Denver, CO 80225

CONTENTS

Page

ILLUSTRATIONS

ILLUSTRATIONS (continued)

Page

TABLES

CONVERSION FACTORS AND ABBREVIATIONS

Multiply	By	To obtain
	Length	
foot (ft)	0.3048	meter
mile (mi)	1.609	kilometer
	Area	
acre	0.4047	hectares
square mile (mi^2)	2.590	square kilometer
	Volume	
gallon (gal)	3.785	liter
	Mass	
ton	0.9072	megagram
	Hydraulic conductivity	
foot per day (ft/d)	0.3048	meter per day
	Gradient	
foot per mile (ft/mi)	0.1894	meter per kilometer

Other abbreviations used in this report:

parts per million (ppm)
micrograms per liter (µg/L)

EVALUATION OF STATISTICAL MODELS TO PREDICT CHEMICAL QUALITY OF SHALLOW GROUND WATER IN THE PINE BARRENS OF SUFFOLK COUNTY, LONG ISLAND, NEW YORK

by Paul E. Stackelberg and Steven F. Siwiec

Abstract

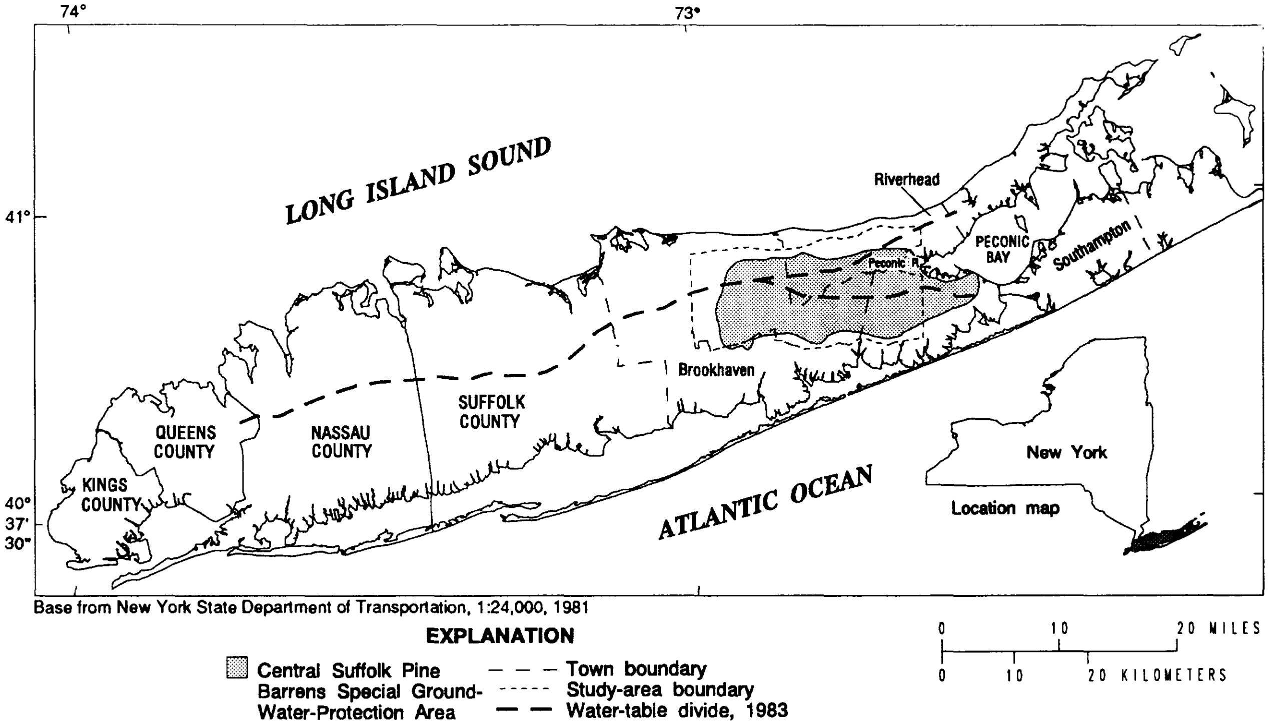
Maximum-likelihood logistic-regression models were evaluated for effectiveness in the prediction of ground-water contamination within a study area in the Pine Barrens of Suffolk County. The models relate the presence or absence of volatile organic compounds (VOC's) at concentrations equal to or exceeding 1 microgram per liter within the shallow ground-water system to population density and land use. The logistic function was used to develop models that quantify the probability of VOC detection within the shallow ground-water system. Population-density values and land-use percentages used by the models were those computed for 1/4-square-mile grid cells of an overlay of the study area by a geographical information system (GIS).

Results suggest that the predictive ability of the models is governed primarily by the degree of resolution of explanatory data. Increased data resolution gave improved model fits, as indicated by the appropriate statistical diagnostics. The predictive ability of the models also depends on the presence of appropriate ranges of explanatory data in modeled areas. Application of models to areas with differing land use and population density resulted in a decrease in the ability of the models to predict the presence of VOC's within the shallow ground-water system, as indicated by an analysis of model residuals.

The modeling technique evaluated in this study, used in conjunction with GIS technology, proved useful in the examination of correlations among shallow ground-water quality and variables related to human activities and, thus, could help decisionmakers predict the effects of proposed development on shallow ground-water quality in areas where models are properly fitted to explanatory data. In addition, the technique can provide maximum-likelihood estimates of shallow ground-water quality at less cost than a ground-water-sampling program in areas where such data are scarce or absent but where explanatory variables can be quantified.

INTRODUCTION

The aquifer system of Long Island, N.Y., is the sole source of freshwater for 2.8 million residents of Nassau and Suffolk Counties (fig. 1). The surficial water-table aquifer, referred to as the upper glacial aquifer, is highly susceptible to contamination by substances introduced at or near the land surface because it is directly exposed at land surface and is highly permeable. In recent decades, extensive development and urbanization have resulted in ground-water contamination from a variety of sources, and this contamination has, in some places, restricted the consumptive use of water from the upper glacial aquifer. The effects of human activities on the quality and quantity of Long Island's ground water have, therefore, become a major concern throughout Nassau and Suffolk Counties.



Base from New York State Department of Transportation, 1:24,000, 1981

Figure 1.--Location of study area and of Pine Barrens Special Ground Water Protection Area (SGPA) in Suffolk County, N.Y.

Special Ground-Water-Protection Areas

In an effort to protect and maintain the quality of ground water on Long Island, the State of New York, in 1987, appropriated funds for the preparation and implementation of watershed-protection plans in "special ground-water-protection areas" (SGPA's) within federally designated sole-source aquifers in counties with a population exceeding 1 million (Trunzo and others, 1987). SGPA's are defined as "significant, largely undeveloped or sparsely developed geographic areas ... which recharge portions of the deep flow aquifer system" (New York State Department of Environmental Conservation, 1986, p. IV-63). The purpose of this action by the State of New York was to--

(1) Establish procedures for designating SGPA's;

(2) delineate areas of vital importance in maintaining water quality in sole-source aquifers and characterize the hydrogeology, water quality, and land uses within these areas;

(3) ensure that these areas are protected and managed to maintain or improve ground-water quality;

(4) establish procedures for the development and implementation of an individual management team for each SGPA;

(5) implement a part of the State ground-water-management program on Long Island to serve as a model for future statewide application; and

(6) establish guidelines for Federal-State cooperation in the planning, funding, and implementation of SGPA-management plans.

An additional intent of this action was to strengthen controls on nonpoint-source contamination to protect potable water beneath aquifer recharge areas.

The concept of SGPA's was first introduced by the New York State Department of Environmental Conservation (NYSDEC) in their 1983 Draft Long Island Groundwater Management Program (NYSDEC, 1983). In this document, the NYSDEC defined criteria for the delineation of SGPA's and proposed nine such areas in Nassau and Suffolk Counties. Some modification of boundaries has resulted from detailed review of SGPA's with regard to the established criteria since the initial designation of these areas.

In 1987, as mandated in Article 55 of the New York State Environmental Conservation Law, an advisory group known as the SGPA Advisory Council (SGPAAC) was established to assist the Long Island Regional Planning Board (LIRPB) in the development, review, and implementation of an SGPA management plan. The SGPAAC includes representatives of State and local government agencies with water-resource management responsibilities, and members of conservation organizations that have property or other interests in one or more of the SGPA's on Long Island. The SGPAAC meets on a regular basis to discuss issues related to management options, proposed zoning changes, development activities, and pending legislation within Long Island's SGPA's. In addition, the Council hears progress reports on special projects dealing with specific issues relevant to their mission.

Background and Objective of this Study

In 1988, the U.S. Geological Survey (USGS), in cooperation with the LIRPB, began a study to evaluate a statistical modeling technique that describes the relations among variables representing human activities and chemical quality of shallow ground water. The techniques used in this study are largely based on work by Eckhardt and Helsel (1988), Eckhardt and others (1989b), and S.J. Cauller (U.S. Geological Survey, written commun., 1991). The study entailed evaluation of selected statistical models, developed by these investigators, for effectiveness as predictive tools within a pilot SGPA in Suffolk County. The models describe the correlation between variables representing human activities (land use and population density) and the occurrence of volatile organic compounds (VOC's) in shallow ground water.

The study was limited to the water-table aquifer in a ground-water recharge area where the primary direction of ground-water flow and associated contaminant transport is assumed to be vertically downward to underlying aquifers. Concentrations of VOC's, as a group, were considered to be representative of overall ground-water quality conditions because their presence is generally considered to be indicative of human influences.

The objective of the present study was to evaluate statistical models that relate land use and population density to the presence or absence of VOC's in shallow ground water as potential tools for decisionmaking regarding future development within SGPA's. This study consisted of four phases: (1) examination of ways to improve the models through the use of fine-scale data and a data-aggregation technique; (2) application of the models to a pilot SGPA on Long Island; (3) verification of model predictions of VOC detection within the study area; and (4) evaluation of the transferability of selected models to areas of similar climatic and geohydrologic conditions.

Purpose and Scope

This report (1) describes the population distribution, land-use patterns, hydrogeology, and ground-water quality in the study area; (2) summarizes results of the statistical model evaluations, including methods used to develop, test, and improve the models; (3) presents results of an analysis of residuals; and (4) discusses model-verification results and model transferability. It also discusses limitations of the models, data requirements, and scale considera-ions. Results are presented as a series of computer-generated maps that illustrate the probability of VOC detection within the shallow ground-water system as a function of variables that represent human activities in the overlying area.

Previous Studies

Several studies by the USGS have assessed the quality of the nation's ground-water reserves and the nature and extent of ground-water contamination. The primary objective of these studies was the statistical evaluation of relations among variables representing human activities and the chemical quality of underlying ground water (Helsel and Ragone, 1984). Results of a study completed on Long Island (Eckhardt and others, 1989a,b) indicate that the presence of VOC's in the water-table aquifer is statistically related to land-use patterns and population density in the overlying area. Similar studies that have related ground-water quality to land use and hydrogeology within the Potomac-Raritan-Magothy aquifer system in New Jersey also found statistically significant correlations between certain land-use categories and the presence of nitrate-nitrogen and VOC's (Vowinkel and Battaglin, 1989; Hay and Battaglin, 1990). Similar results from other studies to evaluate the correlation between variables representing hydrogeology or human activities and shallow-ground-water quality are described in S.J. Cauller (U.S. Geological Survey, written commun., 1991), Grady and Weaver (1988), Barton and others (1987), Rutledge (1987), Chen and Druliner (1987), and Cain and Edelmann (1986). Results of these investigations provide the conceptual basis for the study described herein.

Acknowledgments

Representatives of the LIRPB who provided data and guidance throughout this project include Lee E. Koppelman, Arthur Kunz, Edith Tanenbaum, Roy Fedelem, Gary Palumbo, Peter Lambert, and Anthony Tucci.

STUDY AREA

The area selected for study encompasses 163 mi^2 in central Suffolk County and coincides, in large part, with the Central Suffolk Pine Barrens SGPA (fig. 1). The study area contains parts of the Townships of Brookhaven, Riverhead, and Southampton. The Central Suffolk Pine Barrens SGPA is the largest of nine SGPA's on Long Island and has been given a high priority by the Special Ground Water Protection Area Advisory Council for protection of the quality and quantity of the underlying ground water.

Population Density and Land Use

Human activities within the study area can be assessed and described in terms of land-use categories and population distribution. The pattern of land use within the study area in 1973, as compiled at a scale of 1:250,000, is depicted in figure 2A and summarized in table 1A (p. 9); the pattern in 1981 is shown at a scale of 1:24,000 in figure 2B and summarized in table 2B (p. 9). The 1981 delineations are the basis for the following discussion.

The predominant land-use categories within the study area are vacant land and open space/recreational land, which together constitute about 56 percent of the area and are concentrated mainly in the Towns of Brookhaven and Southampton (fig. 1). The next-largest category is agricultural land, which is concentrated largely in the Town of Riverhead and constitutes about 15 percent of the study area. The transportation/utilities category constitutes about 10 percent of the study area and is represented mainly by two large facilities--an airport in the southeastern corner of the study area in the Town of Southampton, and an industrial plant near the center of the study area in the Towns of Brookhaven and Riverhead. The residential land-use category represents 9 percent of the study area and is concentrated primarily in the Towns of Brookhaven and Riverhead. Institutional land use constitutes about 6.6 percent of the study area and is represented mostly by a laboratory facility in the Town of Brookhaven. Commercial/industrial land makes up about 2 percent of the study area and is concentrated mostly along major transportation corridors in the Towns of Brookhaven and Riverhead. Water bodies account for less than 1 percent of the study area.

The general distribution and density of population within the study area, as indicated by the distribution and density of dwelling units, is shown in figure 3. This figure indicates that residential land with 4.99 or fewer dwelling units per acre constitutes 7.6 percent of the study area, and residential land with five or more dwelling units per acre constitutes less than 2 percent of the study area. Most of the residential land is within the Towns of Brookhaven and Riverhead, and the heaviest concentration is in the westernmost part of the study area. Current and projected population figures for the townships within the study area are listed in figure 3; these indicate that most of the population within the study area resides in the Town of Brookhaven.

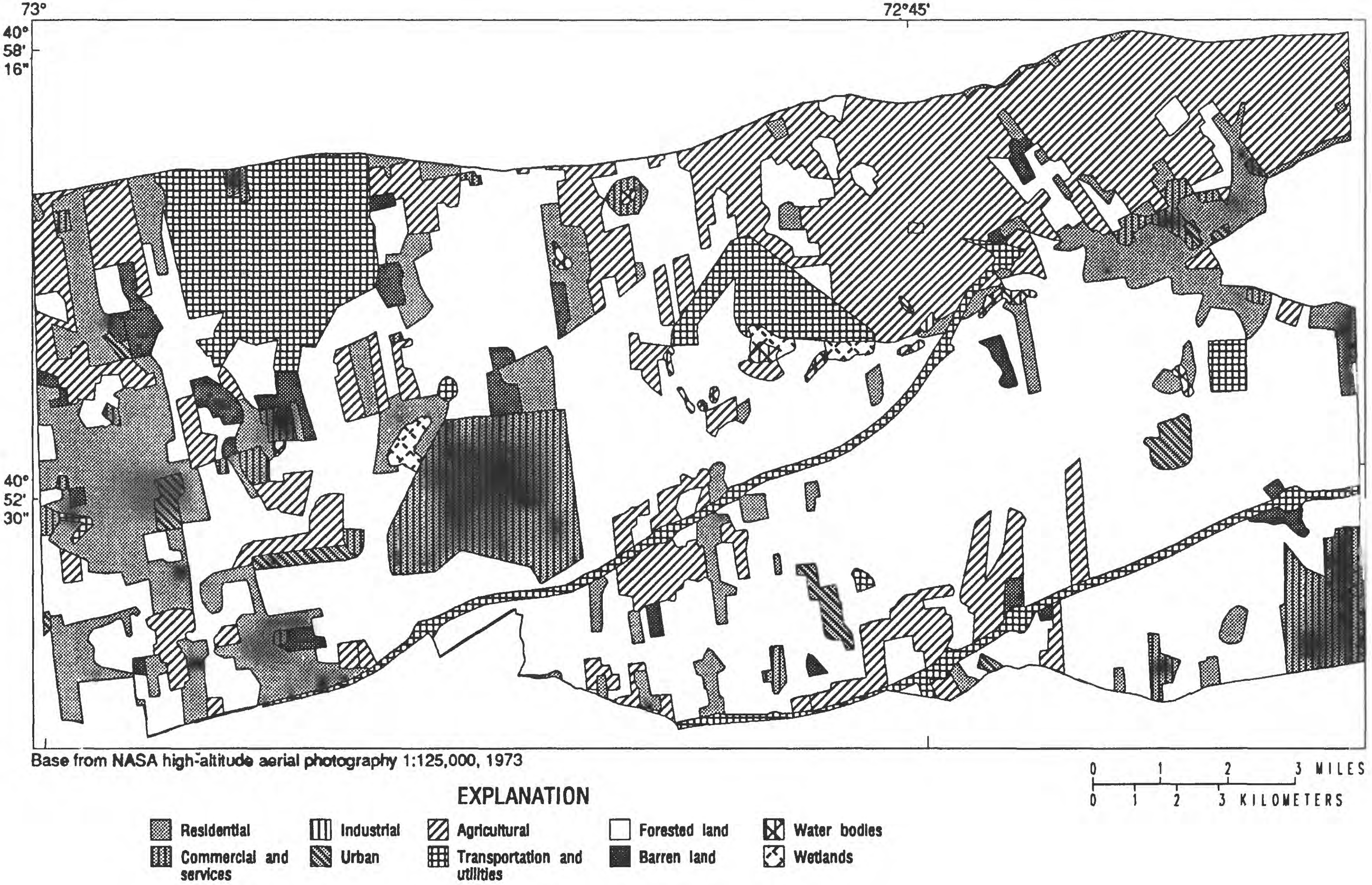
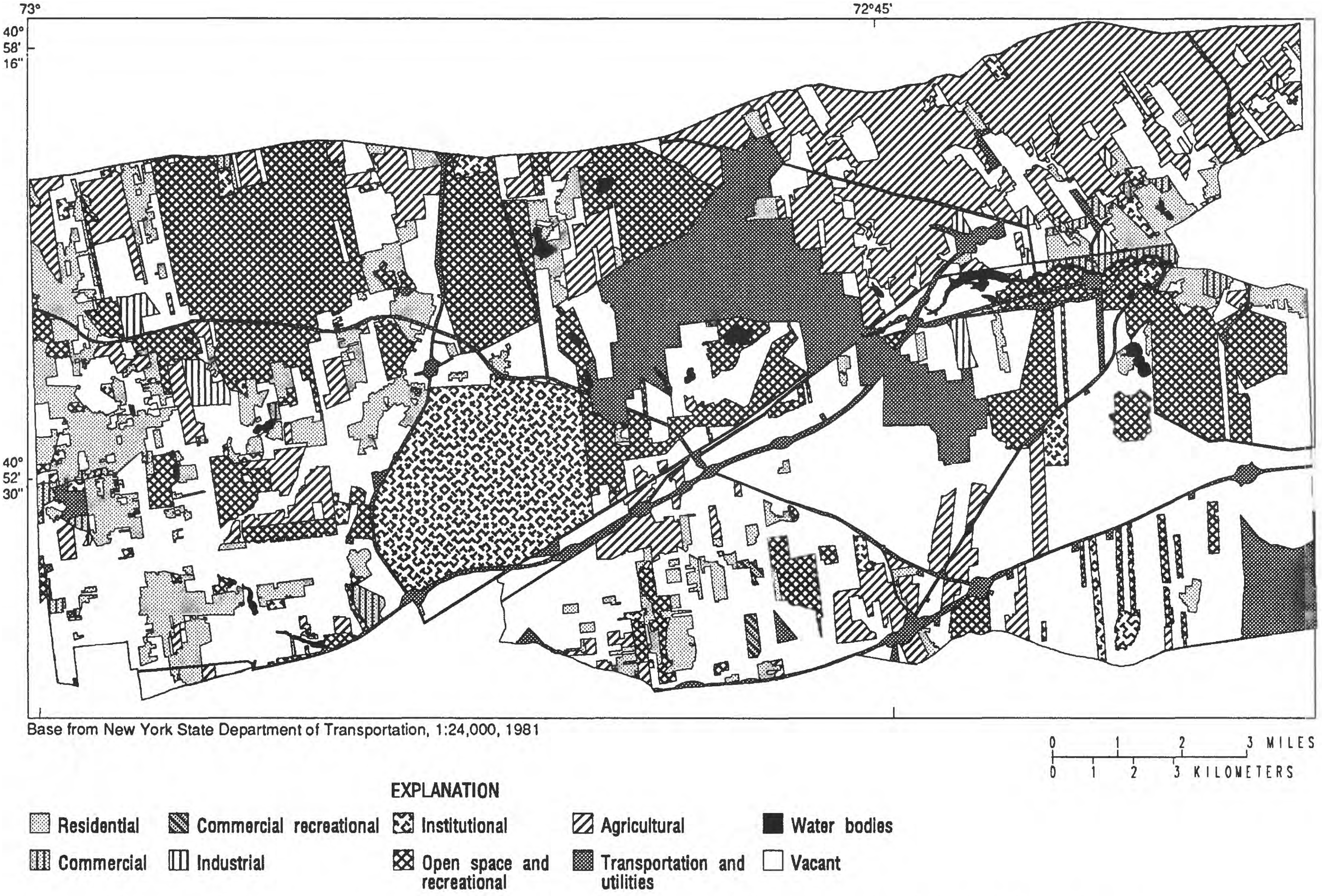
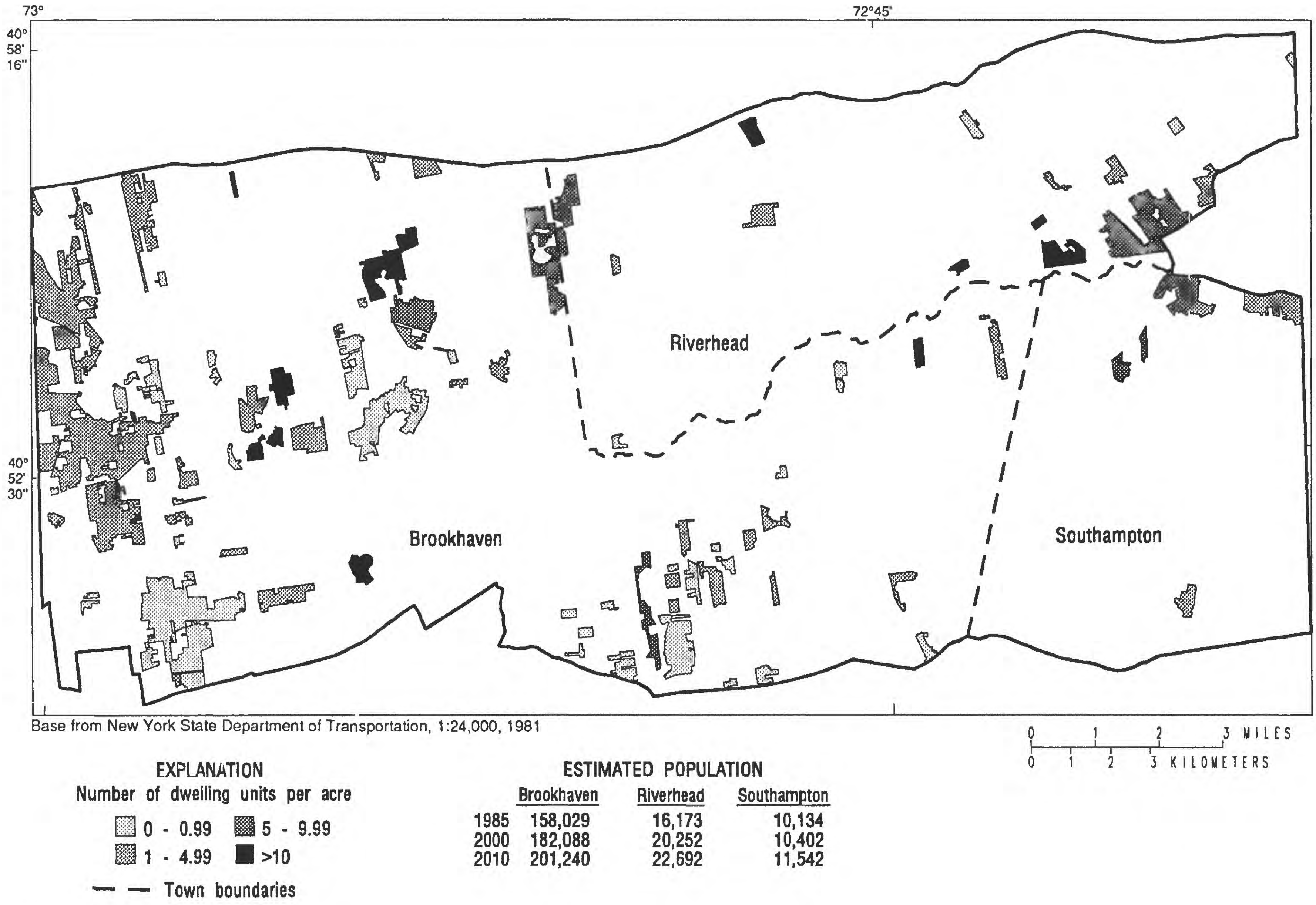


Figure 2A.--Land use in the study area in 1978 compiled at a scale of 1:250,000. (Location is shown in fig. 1. Data from Fegeas and others, 1983.)



Base from New York State Department of Transportation, 1:24,000, 1981

Figure 2B.--Land use in the study area in 1981 compiled at a scale of 1:24,000. (Location is shown in fig. 1. Data from Long Island Regional Planning Board, 1982.)



Base from New York State Department of Transportation, 1:24,000, 1981

ESTIMATED POPULATION

	Brookhaven	Riverhead	Southampton
1985	158,029	16,173	10,134
2000	182,088	20,252	10,402
2010	201,240	22,692	11,542

Figure 3.--1981 dwelling-unit density in residential areas (data from Long Island Regional Planning Board), with actual (1985) and projected (2000 and 2010) population estimates. (Population data provided by Peter Lambert, Long Island Regional Planning Board, written commun., 1989.)

Table 1.--Land use within the study area, by township

[Values are in percent. Locations are shown in fig. 1; mi^2 = square miles]

	Township (area)			Entire
Land-use category	Brookhaven (93.9 mi^2)	Southampton (27.3 mi^2)	Riverhead (41.4 mi^2)	study area (162.6 mi^2)
A. 1:250,000-scale 1973 data (fig. 2A)				
Residential	15.91	5.95	8.55	12.36
Commercial and services	7.10	8.95	1.88	6.08
Industrial	.00	.00	.09	.02
Urban	1.23	1.50	.23	1.02
Agricultural	11.69	2.79	53.85	20.93
Transportation/utilities	10.52	4.44	6.27	8.41
Forrested	49.84	75.06	26.91	48.24
Barren	3.14	1.06	.49	2.12
Water bodies	.19	.24	1.08	.42
Wetlands	.36	.00	.65	.38
B. 1:24,000-scale 1981 data (fig. 2B)				
Residential				
Number of dwelling units per acre:				
0.99 or fewer	5.17	0.03	1.34	3.33
1 to 4.99	5.51	1.46	3.33	4.27
5 to 9.99	.70	1.05	0.44	.62
10 or more	1.21	.29	0.81	.95
Commercial	1.10	.52	1.71	1.16
Commercial recreational	.24	.03	0.06	.16
Industrial	.92	.34	0.78	.79
Institutional	9.98	3.32	1.00	6.57
Open space/recreational	19.32	23.36	9.90	17.66
Agricultural	7.99	2.59	40.00	15.24
Transportation/utilities	6.64	9.72	19.42	10.41
Vacant	40.83	56.63	20.10	38.20
Water bodies	.39	.65	1.11	.62

Hydrogeologic Setting

The aquifer system beneath the study area (fig. 4) consists primarily of unconsolidated deposits of (1) sand and gravel, and (2) interbedded sandy clay, clayey sand, and silt of glacial, marine, and fluvial and deltaic origin. Cretaceous sediments unconformably overlie relatively impermeable crystalline bedrock and, along with the bedrock, dip southeastward at about 65 ft/mi (McClymonds and Franke, 1972). Local erosion of Cretaceous sediments by streams and glaciers left an irregular surface of moderate relief upon which sediments of Pleistocene age were deposited. These Pleistocene deposits form the upper glacial aquifer, which is the only part of the aquifer system that was of concern in this study; therefore, the deeper units are not discussed further.

The upper glacial aquifer consists of material deposited by Pleistocene glaciers as terminal moraines. Some of this material was reworked by glacial meltwater to form large outwash-plain deposits of stratified sand and gravel. Sediments that form the upper glacial aquifer are highly permeable and have an average horizontal hydraulic conductivity of 270 ft/d (Franke and Cohen, 1972). The estimated average vertical hydraulic conductivity is 27 ft/d. Within the study area, the upper glacial aquifer is 200 to 300 ft thick except where the underlying Magothy aquifer was eroded by rivers; glacial material deposited within these eroded valleys reaches thicknesses of 600 to 700 ft. The study area overlies the regional ground-water divide, where the direction of ground-water flow is vertically downward; thus, the study area is predominantly an area of recharge to the deep flow system as well as to the shallow (upper glacial) system. Local exceptions are in the vicinity of the Peconic River and Peconic Bay, at the eastern border of the study area (fig. 1), where recharge enters a shallow flow subsystem and discharges to these surface-water bodies (Krulikas, 1986).

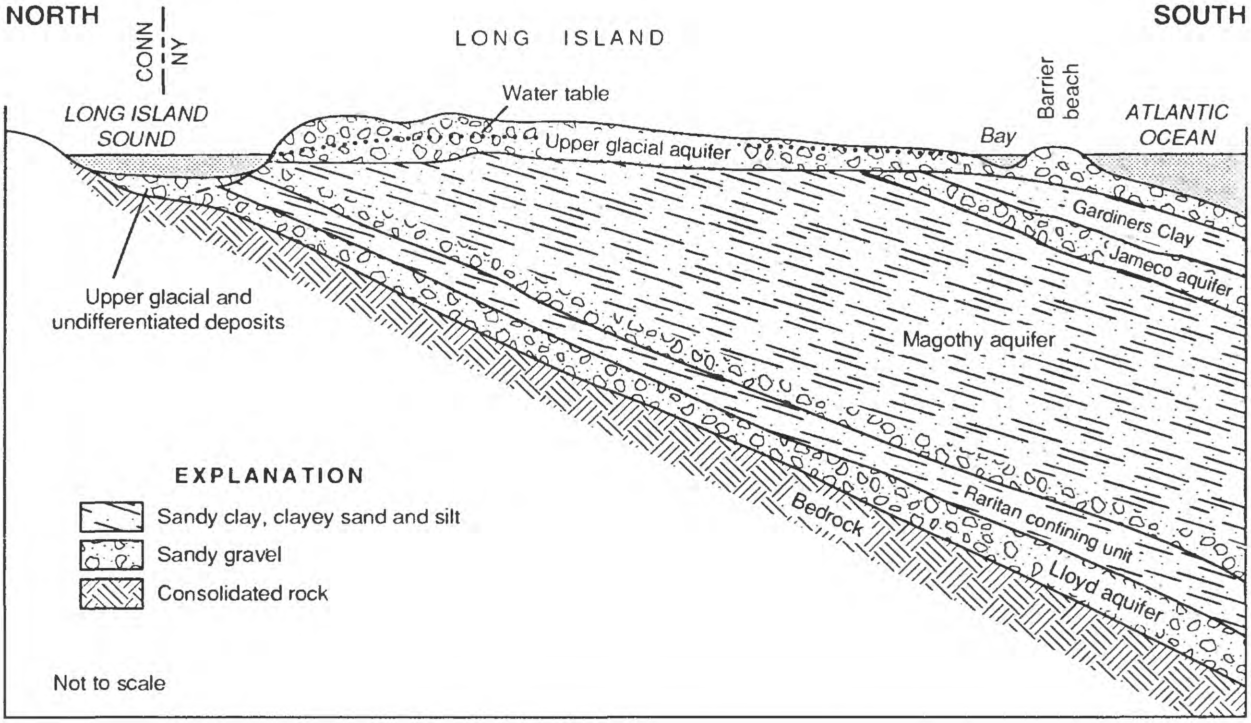


Figure 4.--Generalized geologic section of Long Island showing relative positions of the principal aquifers. (Modified from McClymonds and Franke, 1972, fig. 3.)

Ground-Water Quality

Water within the upper glacial aquifer is generally potable, except where it has been affected by contaminants from human activities. The parts of the study area that have been most severely affected include the densely developed areas in the Town of Brookhaven and agricultural areas in the Towns of Riverhead and Southampton. The following discussion focuses on nitrates and VOC's because they are indicative of human effects on ground-water quality.

Nitrate contamination within the study area is most prevalent in the agricultural areas of the Town of Riverhead. Elevated concentrations of nitrate from fertilizer applications and, possibly, septic effluent, have resulted in marginal (6 to 10 ppm) to poor (greater than 10 ppm) ground-water quality in these areas (Dvirka and Bartilucci, 1987). A nitrate concentration of 10 ppm is the current U.S. Environmental Protection Agency's drinking water standard. Elevated nitrate concentrations are also found in the densely populated western part of the study area, in the Town of Brookhaven (Long Island Regional Planning Board, 1986). Nitrate contamination in residential areas is attributed to discharges from septic systems and to lawn fertilizers. Nitrate concentrations in the predominantly vacant and undeveloped pine barrens of central and southeastern Brookhaven are generally less than 1 ppm (Dvirka and Bartilucci, 1987).

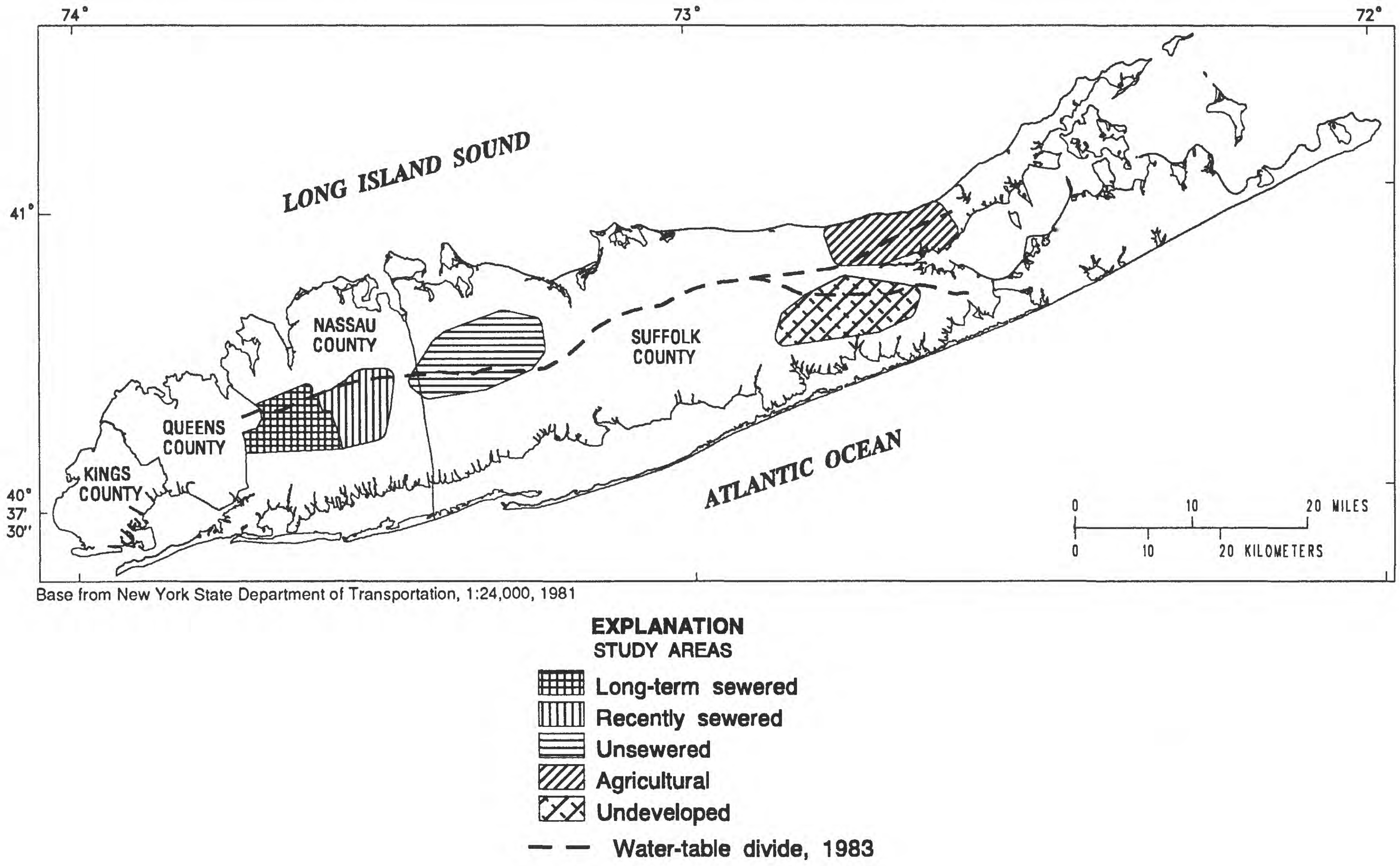
Dvirka and Bartilucci (1987) also report that shallow ground water within the study area generally has low concentrations of VOC's (0.01 to 0.30 μg/L for an individual VOC; 0.01 to 0.60 μg/L for total VOC's), and VOC's are undetected in large parts of the relatively undeveloped areas in and around the pine barrens. The few pockets of VOC contamination within the study area contain relatively low concentrations and are generally within the densely populated areas of Brookhaven and Riverhead; the contamination is attributed mainly to industrial discharges, landfills, wastewater-disposal systems, leaks and spills, and septic effluent.

STATISTICAL MODELS TO PREDICT THE CHEMICAL QUALITY OF SHALLOW GROUND WATER

The statistical models evaluated in this study were developed by previous investigators and are discussed in Eckhardt and others (1989b). These models describe the statistical relation between one or more variables that represent human activities and the detection (presence or absence) of VOC's in the underlying ground water. The following discussion gives a brief overview of the methods used to develop these models. Additional details regarding the methods of model development and data acquisition (water quality, land use, and population) are given in Eckhardt and others (1989b), Eckhardt and Helsel (1988), Leamond and others (1992), and S.J. Cauller (U.S. Geological Survey, written commun., 1991).

Development

The basis for development of these models is a network of 90 shallow observation wells in five representative land-use areas in Nassau and Suffolk Counties (S.J. Cauller, U.S. Geological Survey, written commun., 1991) (fig. 5). The land-use classification was based on predominant land use and historical sewering practices. Wells were located by a random-selection procedure, wherein a grid was superimposed over a map of each study area, and one well was selected from each grid cell (Leamond and others, 1992). This procedure ensured a uniform spatial distribution of samples and minimized the effects of spatial autocorrelation as described by Barringer and others (1990) and Hay and Battaglin (1990). No wells were screened deeper than 45 ft below the water table. Average ground-water flow rates (McClymonds and Franke, 1972) indicate that ground water withdrawn from these wells is probably less than 10 years old and would be expected to reflect the effects of human activities within this time period.



Base from New York State Department of Transportation, 1:24,000, 1981

Figure 5.--Location of the five land-use areas on Long Island, New York. (From Eckhardt and others, 1989b, p. 398.)

The statistical technique used in the development of these models, logistic regression, will predict a categorical (binary) response from one or more explanatory variables (SAS Institute Inc., 1988). Logistic regression computes maximum-likelihood estimates of regression coefficients that define the best linear fit between the binary response and the explanatory variables (Harrell and others, 1980). If the resulting regression (slope) coefficient for an explanatory variable is significantly different from zero at a confidence level of 95 percent ($\alpha = 0.05$), the correlation is statistically significant. In these models, the response variable was the detection or nondetection of VOC's at concentrations equal to or exceeding 1 µg/L in water from the 90 observation wells. The explanatory variables were (1) the percentage of residential land within a 1/2-mi radius of each well site, (2) the percentage of industrial and commercial land within that radius, or (3) the density of population within that radius.

The land-use and population data from which explanatory variables were computed were digitally automated as data layers in a geographic information system (GIS). GIS spatial analysis allowed the computation of the explanatory variables within the specified areas around each well. For each well, values of the explanatory variables were paired with the response variable (the detection or nondetection of any VOC's at 1 µg/L or greater) and input into a statistical routine that computes maximum-likelihood estimates for the linear regression coefficients.

Once established, the logistic regression models can be transformed to predict probabilities of VOC contamination through the logistic function

$$P(x) = 100\left(\frac{e^{(a + bx)}}{1 + e^{(a + bx)}}\right),$$

where P(x) is the probability, in percent, of a positive VOC detection, given the matrix x of explanatory variable(s) x, the vector b of slope parameters, and the scalar intercept a. Detailed discussions of logistic regression theory are given in Walker and Duncan (1967), Harrell and others (1980), Harrell and Lee (1985), Freeman (1987), and Helsel and Hirsch (1992).

To evaluate the probability of VOC detection within the shallow ground-water system of the study area, appropriate explanatory variables were quantified for each cell (1/4 mi^2) of a gridded overlay by means of a GIS. Resulting probabilities computed for each cell through the logistic function can then be displayed graphically to illustrate the spatial relation between the explanatory variable and the predicted presence of VOC's.

To verify model predictions, an independent data set of VOC analyses from more than 300 shallow wells throughout Nassau and Suffolk Counties was compared with the predicted VOC detection. Residuals resulting from the differences between observed and predicted VOC detection probabilities provide a measure of the unexplained variation in the data. Models that are properly fitted should provide residuals with a median of zero, a normal distribution, and no spatial trends. All models discussed in this report gave residuals with these characteristics when applied to the entire Nassau-Suffolk County area of Long Island.

Data Acquisition and Model Applications

This section discusses the acquisition of data, models that describe the probability of VOC detection within the shallow ground-water system of the five land-use areas (fig. 5), methods used to improve the fit of the models, and application of the models within the study area.

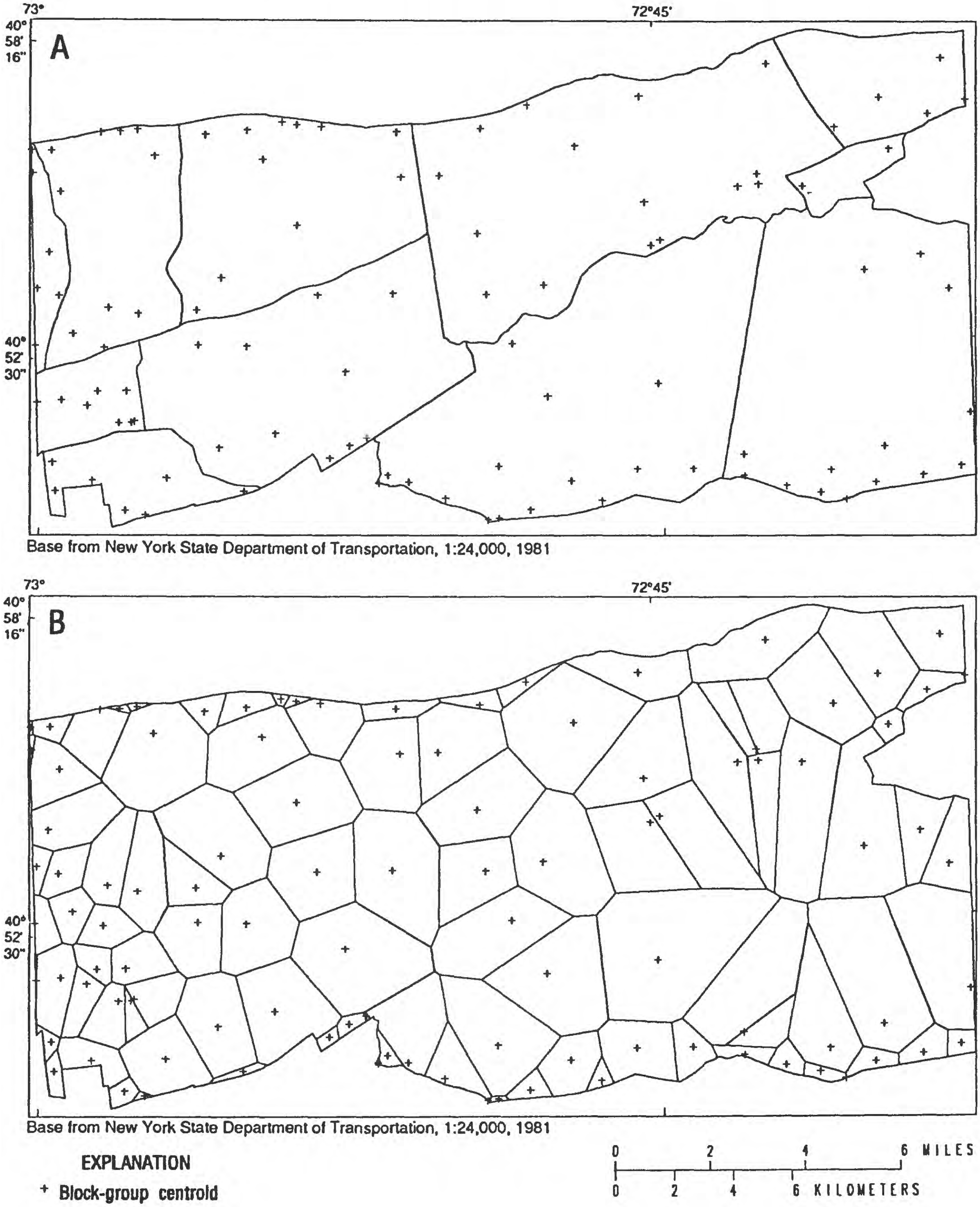
Population-Density Model

The first model examined in this study evaluates the relation between population density and the presence of VOC's. Population data were obtained in digital format from the U.S. Census Bureau and were represented in the form of census tracts and block-group centroids with the associated 1985 population totals. Census tracts are statistical subdivisions of counties that have a range of 2,000 to 8,000 inhabitants and an average of 4,000 inhabitants. Each census tract consists of blocks that generally are bounded by streets and other physical features and are numbered for identification. Block groups comprise all blocks in a census tract with the same first digit, and each block group contains a centroid that is assigned the population totals for all blocks within that group (U.S. Bureau of the Census, 1985). Population densities for each census tract were calculated as the sum of the population within each census tract, divided by the area (acres) of the tract. Census tracts and block-group centroids within the study area are depicted in figure 6A.

To increase the resolution of the population distribution within the study area, Thiessen polygons were generated around each block-group centroid by a GIS. The Thiessen-polygon procedure apportions the study area into polygons around each centroid such that every location in a polygon is nearer to its own centroid than to any other centroid. Thiessen polygons and block-group centroids within the study area are depicted in figure 6B. The population value for each centroid is then averaged over the area of its corresponding Thiessen polygon. This method results in an increase in the number of polygons from 11 to 105 and a decrease in average polygon size from 9,459 to 1,319 acres.

The census-tract model is summarized in table 2A. Intercept and explanatory-variable slope coefficients differ significantly from zero at a 95-percent confidence level, indicating that population density is a statistically significant explanatory variable for predicting the presence of VOC's in the five land-use areas shown in figure 5. The predictive ability of the model was assessed through the likelihood-ratio (goodness-of-fit) statistic, which is analogous to an overall F-statistic in multiple regression (Freeman, 1987). This statistic measures the amount of variation in VOC-detection data explained by the model and thereby provides a measure of how well the model describes VOC detection probability at the 90 wells in the five land-use areas. A likelihood ratio of 75.10 was obtained that provides a means for comparison of model fit relative to that of subsequent models. Relatively higher ratios are indicative of better fitting models.

The Thiessen-polygon model is summarized in table 2B. Intercept and explanatory-variable slope coefficients differ significantly from zero at a 95-percent confidence level, indicating that population density is a statistically significant explanatory variable for predicting the presence of VOC's in the five land-use areas (fig. 5). The likelihood-ratio (goodness-of-fit) test for this model was 86.28, an improvement over that of the census-tract model (75.10), indicating an improved fit by the Thiessen-polygon method.



Base from New York State Department of Transportation, 1:24,000, 1981

Base from New York State Department of Transportation, 1:24,000, 1981

Figure 6.--Population data for the study area: A. Block-group centroids and census tracts. B. Block-group centroids and Thiessen polygons.

Table 2.--Selected output from SAS CATMOD procedure for population-density models and probability equation based on logistic function

[POPDEN, population density, per acre; P, probability (in percent) of VOC detection at or above 1 microgram per liter]

A. Census-tract data

Effect	Parameter	Slope coefficient	Standard error	Chi-square	p-value
INTERCEPT	1	-1.92183	0.413932	21.56	0.0001
POPDEN	2	.227751	.0530237	18.45	.0001

Likelihood ratio (goodness of fit) = 75.10

Probability equation (based on logistic function):

$$P = 100\left(\frac{e^{(-1.92 + 0.23(\mathrm{POPDEN}))}}{1 + e^{(-1.92 + 0.23(\mathrm{POPDEN}))}}\right)$$

B. Thiessen-polygon data

Effect	Parameter	Slope coefficient	Standard error	Chi-square	p-value
INTERCEPT	1	-2.26998	0.469349	23.39	0.0001
POPDEN	2	.280623	.060009	21.87	.0001

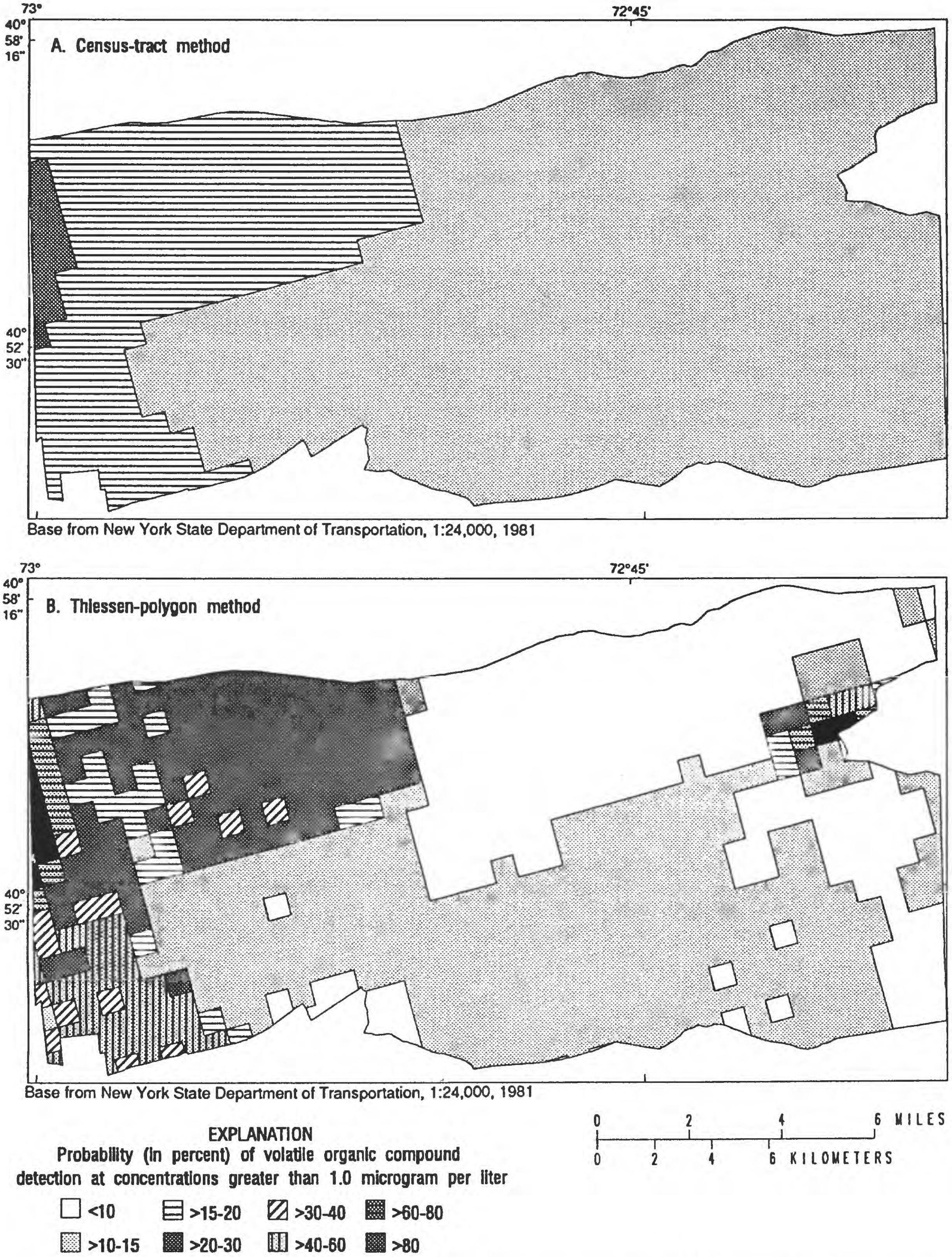
Likelihood ratio (goodness of fit) = 86.28

Probability equation (based on logistic function):

$$P = 100\left(\frac{e^{(-2.27 + 0.28(\mathrm{POPDEN}))}}{1 + e^{(-2.27 + 0.28(\mathrm{POPDEN}))}}\right)$$

Results of the census-tract-model application to the study area are displayed in figure 7A. The probability percentages of VOC detection in figure 7A are consistent with the general eastward decrease in population density across the study area. The relatively small range of detection-probability values (13 to 26 percent) and the corresponding small number of probability categories (3) are a function of the aggregation of population values over entire census-tract areas.

Results of the Thiessen-polygon model application to the study area are displayed in figure 7B. Comparison of these results with the census-tract model results (fig. 7A) indicates that the Thiessen-polygon procedure gives a larger range of probability values (9 to 91 percent) and finer resolution of VOC-detection probability. Several areas of low and high probability are revealed by the Thiessen-polygon model that are obscured in the census-tract model through generalization of population values over entire census-tract areas.



Base from New York State Department of Transportation, 1:24,000, 1981

Base from New York State Department of Transportation, 1:24,000, 1981

Figure 7.--Probability of VOC detection as predicted by population-density models based on: A. Census-tract method. B. Thiessen-polygon method.

Land-Use Models

The second model examined in this study relates the presence of VOC's in ground water to land use. Land-use data for 1973, compiled at a scale of 1:250,000, were obtained in digital format from the USGS (Fegeas and others, 1983). A multivariate logistic-regression model was developed from percentages of residential land and of industrial and commercial land around each of the 90 wells in the five land-use areas as the explanatory variables (Eckhardt and others, 1989b). Land use within the study area, compiled at a scale of 1:250,000, is depicted in figure 2B and summarized in table 1B.

To increase the resolution of land use within the study area, a 1981 data set that was compiled at a scale of 1:24,000 (Long Island Regional Planning Board, 1982) was used. The data were in the form of aerial photographs and were digitially automated into a GIS format. The same explanatory variables (residential and industrial/commercial land use) were used to generate a new logistic-regression model. The finer scale of this second data set decreased the smallest resolvable polygon size from 9.88 to 0.80 acres (fig. 2A and 2B).

The 1:250,000-scale model is summarized in table 3A. Intercept and explanatory-variable slope coefficients differ significantly from zero at a 95-percent confidence level, indicating that residential and industrial/ commercial land use are statistically significant explanatory variables for predicting the presence of VOC's in the five land-use areas shown in figure 5. The likelihood-ratio (goodness of fit) of this model was 71.98.

The 1:24,000-scale model is summarized in table 3B. Intercept and explanatory-variable slope coefficients differ significantly from zero at the 95-percent confidence level, indicating that residential and industrial/ commercial land use are significant explanatory variables for predicting the presence of VOC's in the five land-use areas (fig. 5). The attained significance levels (p-values) for the intercept and explanatory-variable slope coefficients were consistently lower than those of the 1:250,000 model, and the likelihood-ratio (goodness-of-fit) test for this model was 93.02, an improvement over that of the 1:250,000 model (71.98); both results indicate that the 1:24,000-scale data provided an overall better model fit than the 1:250,000-scale data.

In the 1:250,000 model, the slope coefficient associated with the residential land-use variable is 1.5 times greater than that associated with the industrial/commercial variable, whereas the slope associated with the residential land-use variable in the 1:24,000 model is nearly equal to that associated with the industrial/commercial variable. Spatial overlay of these land-use data through a GIS indicates that the majority of this difference between the two models is due to the misclassification of small (less than 9.88-acre) parcels of industrial/commercial land that are interspersed within residential areas and are unresolvable at the 1:250,000 scale, and thus were misclassified as residential (Siwiec and Stackelberg, 1989). Although most of the difference between slopes of the explanatory variables in the two logistic-regression equations is due to the failure to represent these small parcels of industrial/commercial land, some of the difference could result from differences between the two time periods represented by the 1973 and 1981 data sets.

Table 3.--Selected output from SAS CATMOD procedure for land-use models, and probability equation based on logistic function

[INDLU, percentage of commercial/industrial land within reference area; RESLU, percentage of residential land within reference area; VOC, volatile organic compound; P, probability (in percent) of VOC detection at concentration equal to or above 1 microgram per liter]

A. 1:250,000-scale 1973 data

Effect	Parameter	Slope coefficient	Standard error	Chi-square	p-value
INTERCEPT	1	-3.33117	0.920608	13.09	0.0003
INDLU	2	3.08451	1.12501	7.52	.0061
RESLU	3	4.6464	1.36754	11.54	.0007

Likelihood ratio (goodness of fit) = 71.98

Probability equation (based on logistic function):

$$P = 100\left(\frac{e^{(-3.33 + 3.08(\mathrm{INDLU}) + 4.65(\mathrm{RESLU}))}}{1 + e^{(-3.33 + 3.08(\mathrm{INDLU}) + 4.65(\mathrm{RESLU}))}}\right)$$

B. 1:24,000-scale 1981 data

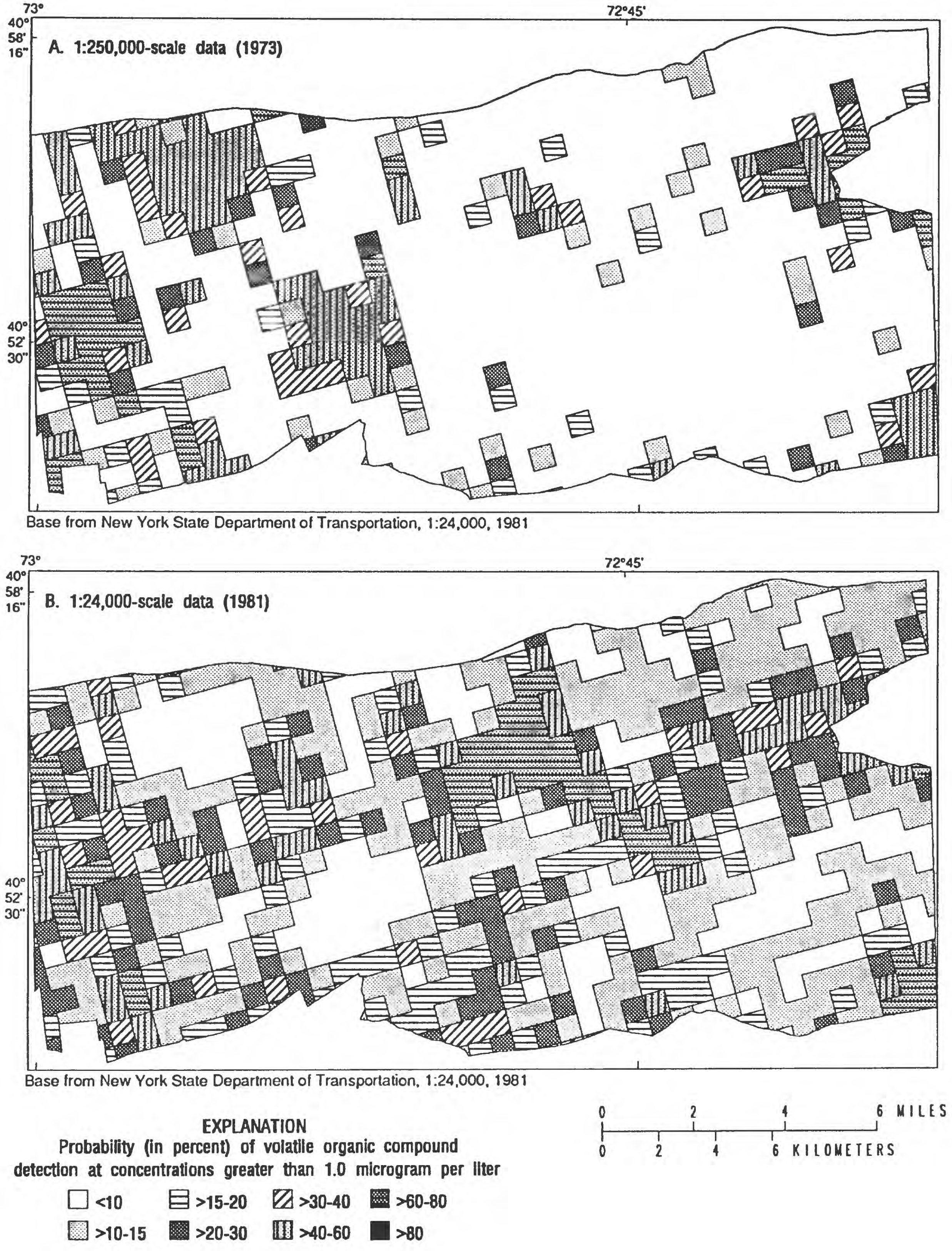
Effect	Parameter	Slope coefficient	Standard error	Chi-Square	p-value
INTERCEPT	1	-2.10336	0.483624	18.92	0.0001
INDLU	2	3.2206	1.15495	7.78	.0053
RESLU	3	3.52818	.091476	14.04	.0002

Likelihood ratio (goodness of fit) = 93.02

Probability equation (based on logistic function):

$$P = 100\left(\frac{e^{(-2.10 + 3.22(\mathrm{INDLU}) + 3.53(\mathrm{RESLU}))}}{1 + e^{(-2.10 + 3.22(\mathrm{INDLU}) + 3.53(\mathrm{RESLU}))}}\right)$$

Results of the 1:250,000-scale model application to the study area are displayed in figure 8A. The probability percentage values in figure 8A generally reflect the predominant land-use features across the study area; that is, high detection-probability values are associated with industrial or commercial areas and predominantly residential areas, whereas low probability values are associated with undeveloped and agricultural land or with land classified as other than residential or industrial/commercial (fig. 2B).


73°
72°45'
40° 58' 16"
A. 1:250,000-scale data (1973)
40° 52' 30"
73°
72°45'
40° 58' 16"
B. 1:24,000-scale data (1981)
40° 52' 30"
EXPLANATION
Probability (in percent) of volatile organic compound detection at concentrations greater than 1.0 microgram per liter
□ <10
▤ >15-20
▨ >30-40
▩ >60-80
░ >10-15
▦ >20-30
▥ >40-60
■ >80
0 2 4 6 MILES
0 2 4 6 KILOMETERS

Base from New York State Department of Transportation, 1:24,000, 1981

Base from New York State Department of Transportation, 1:24,000, 1981

Figure 8.--Probability of VOC detection as predicted by land-use models based on: A. 1:250,000-scale data. B. 1:24,000-scale data.

Results of the 1:24,000-scale model application to the study area are displayed in figure 8B. Comparison of these results with those of the 1:250,000-scale model (fig. 8A) shows increased probability of VOC detection at the finer resolution of the 1:24,000 data. Most areas that show low detection probability in the 1:250,000-scale model results have higher probability values in the 1:24,000 model results, in which the small parcels of industrial/commercial land within the residential areas are identified. Areas of high detection probability in the 1:250,000-scale model results also tend to be high in the 1:24,000-scale model results, except in the northwest and south-central parts of the study area, where the 1:250,000 model indicated high detection probability and the 1:24,000 model indicated low detection probability. These results are attributed primarily to the classification-scheme differences between the two land-use data sets. These areas are classified as commercial and transportation/utilities in the 1:250,000 data set, but as institutional and open space/recreational in the 1:24,000 data set (figs. 2A and B).

Analysis of Residuals

Analysis of residuals provides a means of verifying model predictions within the study area and evaluating the transferability of the models from the five land-use areas described by Leamond and others (1992) (fig. 5) to other areas with similar geohydrologic and climatic conditions.

Verification of Model Predictions

An independent set of data from VOC analyses from 33 shallow wells throughout the study area was compared with the predicted VOC detection to verify model predictions. Models that are properly fitted to the explanatory data within the study area should provide residuals with a median of zero, a normal distribution, and no spatial trends.

Residuals from the population-density models are displayed as boxplots in figure 9A and 9B. The smaller variability of residuals from the Thiessen-polygon model than from the census-tract model, and the smaller departure of the median from zero, indicates that the Thiessen-polygon model provided a better fit to observed water quality within the study area than the census-tract model. Residuals in both models are highly skewed and have medians less than zero, however, which indicates that both models are systematically over-predicting the probability of VOC detection within the shallow ground-water system of the study area. Spatial trends were not discernible in residuals from either population-density model.

Residuals from the land-use models are displayed in figures 9C and 9D. Residuals from the 1:250,000 model show less variability than those from the 1:24,000 model and have a median of about zero, whereas those from the 1:24,000 model are more highly skewed and have a median less than zero. These results indicate that the 1:24,000 model is systematically overpredicting the probability of VOC detection within the shallow ground-water system. Spatial trends were not discernible in residuals from either land-use model.

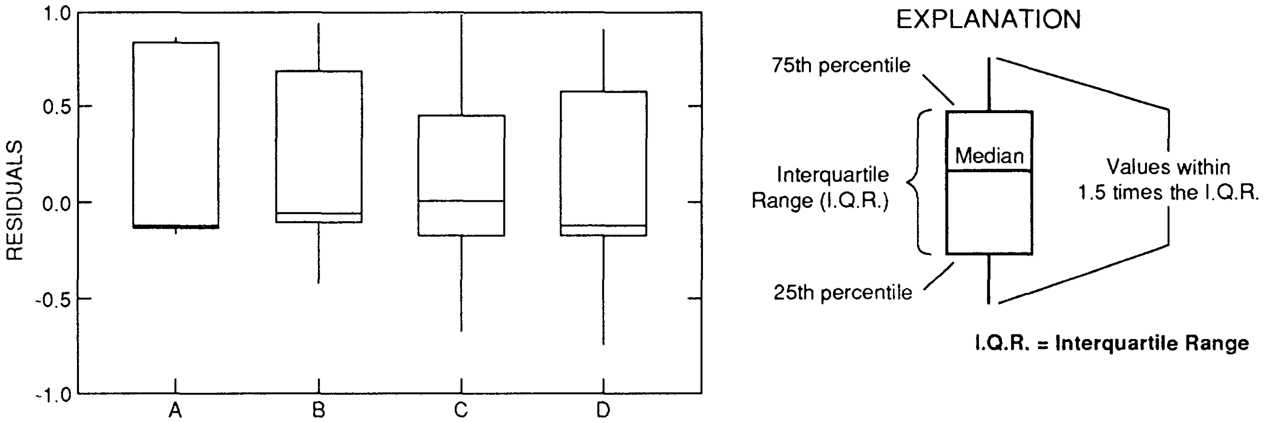


Figure 9. Box plots of residuals resulting from models: A. Census-tract population-density model. B. Thiessen-polygon population-density model. C. 1:250,000 land-use model. D. 1:24,000 land-use model.

Transferability of Models

Application of the models developed within the five land-use areas described by Leamond and others (1992) (fig. 5) to areas of similar land use, population density, hydrogeology, and climate provides a means for evaluating the transferability of these models to other regions with similar geohydrologic and climatic conditions (Helsel and Ragone, 1984).

Residuals resulting from the application of the population-density and land-use models to the SGPA (fig. 1) indicate systematic overprediction of the probability of VOC detection within the shallow ground-water system, even though the models are well fitted to the explanatory data within the five land-use areas. This systematic overprediction is attributed to differences between land-use percentages and population densities of the five land-use areas and those of the SGPA. The models were developed for areas that have predominant land-use patterns ranging from highly developed to relatively pristine and population density ranging from less than 1 to more than 20 per acre, whereas the SGPA is classified as predominantly vacant, open space/recreational, and agricultural (table 1A), and its population rarely exceeds 5 per acre. These differences in land-use and population-density characteristics make the models poorly fitted to the explanatory data within the SGPA and, therefore, led to systematic overprediction of the probability of VOC detection in the SGPA.

Land-use and population-density characteristics of the entire Nassau and Suffolk Counties area parallel those of the five land-use areas more closely than those of the SGPA do, therefore, application of the models to the entire two-county area would likely result in improved model fits. Results from models determined to be properly fit to explanatory data from the entire two-county area can allow examination of results from a specific area of interest, such as the SGPA. Alternatively, the logistic-regression models could be developed and verified within specific areas of interest, provided that sufficient water-quality and explanatory data are available, thereby avoiding the issue of model transferability. If the models are properly fitted and verified, they can be used to examine the relation between variables representing human activities and shallow ground-water quality and can provide maximum-likelihood estimates of shallow ground-water-quality conditions at less cost than would be required for a ground-water-sampling program. Additionally, the models can be used by decisionmakers to predict areas in which ground-water contamination is most likely to result from proposed development scenarios. In all applications, the reliability of the models depends on the availability of explanatory data bases, a proper fit to explanatory data in the areas of application, and accurate water-quality data for model development and verification.

SUMMARY

Logistic-regression modeling, coupled with geographical information system technology, provides a promising technique for evaluation of the possible effects of human activities and proposed development on the chemical quality of shallow ground water. Models presented in this study relate the presence or absence of volatile organic compounds within the shallow ground-water system to population density and land use in overlying areas. Results presented herein indicate that the predictive ability of the models is governed primarily by (1) the degree of resolution of the original explanatory data, and (2) the presence of appropriate ranges of explanatory data in areas of model application.

Increasing the degree of explanatory-data resolution consistently improved the statistical correlations between variables that represent human activities and shallow ground-water quality, as indicated by the appropriate statistical diagnostics. Methods used to improve the resolution of explanatory data included the use of (1) a GIS to generate Thiessen polygons for precise spatial representation of census data, and (2) large-scale, recent maps of land use in place of old, small-scale maps.

Application of models to areas with differing land-use and population-density characteristics resulted in a decreased ability to predict VOC detection within the shallow ground-water system, as indicated by an analysis of model residuals.

This modeling technique can help decisionmakers to predict where ground-water contamination is most probable and what the effects of proposed development on the chemical quality of shallow ground water will be. The models depend on (1) the availability of explanatory data bases at scales appropriate for the intended use of model results, (2) proper fit to explanatory data in areas of application, and (3) accurate water-quality data for model development and verification. An additional value of these models is that they can provide maximum-likelihood estimates of shallow ground-water quality at lower cost than would be required for a ground-water sampling program in areas where such data are scarce or absent, as long as explanatory variables can be quantified accurately.

REFERENCES CITED

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

Barton, Cynthia, Vowinkel, E.F., and Nawyn, J.P., 1987, Preliminary assessment of water quality and its relation to hydrology and land use, Potomac-Raritan-Magothy aquifer system, New Jersey: U.S. Geological Survey Water-Resources Investigations Report 87-4023, 79 p.

Cain, D.L., and Edelmann, Patrick, 1986, A reconnaissance water-quality appraisal of the Fountain Creek alluvial aquifer between Colorado Springs and Pueblo, Colorado, including trace elements and organic constituents: U.S. Geological Survey Water-Resources Investigations Report 86-4085, 45 p.

Chen, Hsiu-Hsiung, and Druliner, A.D., 1987, Nonpoint-source agricultural chemicals in ground water in Nebraska--preliminary results for six areas of the High Plains aquifer: U.S. Geological Survey Water-Resources Investigations Report 86-4338, 68 p.

Dvirka and Bartilucci Consulting Engineers, 1987, Suffolk County Comprehensive Water-Resources Management Plan, v. 1 and 2: Syosset, N.Y. [unpaginated]

Eckhardt, D.A., Flipse, W.J., Jr., and Oaksford, E.T., 1989a, Relation between land use and ground-water quality in the upper glacial aquifer in Nassau and Suffolk Counties, Long Island, New York: U.S. Geological Survey Water-Resources Investigations Report 86-4142, 35 p.

Eckhardt, D.A., Siwiec, S.F., and Cauller, S.J., 1989b, Regional appraisal of ground-water quality in five different land-use areas, Long Island, New York: U.S. Geological Survey Water-Resources Investigations Report 88-4220, p. 397-403.

Eckhardt, D.A., and Helsel, D.R., 1988, Statistical methods for a regional ground-water quality appraisal in different land-use areas, Long Island, New York [Abs.], *in* Proceedings of the Division of Environmental Chemistry: Los Angeles, Calif., American Chemical Society, p. 29.

Fegeas, R.G., Claire, R.W., Guptill, S.C., Anderson, K.E., and Hallam, C.A., 1983, USGS digitial cartographic data standards--land use and land cover digital data: U.S. Geological Survey Circular, 895-E, 21 p.

Franke, O.L., and Cohen, Philip, 1972, Regional Rates of ground-water movement on Long Island, New York, *in* Geological Survey Research 1972: U.S. Geological Survey Professional Paper 800-C, p. C271-C277.

Freeman, D.H., 1987, Applied categorical analysis: New York, Marcel Dekker, 318 p.

REFERENCES CITED (continued)

Grady, S.J., and Weaver, M.F., 1988, Preliminary appraisal of the effects of land use on water quality in stratified-drift aquifers in Connecticut, U.S. Geological Survey Toxic Waste--Ground-Water Contamination Program: U.S. Geological Survey Water-Resources Investigations Report 87-4005, 41 p.

Harrell, F.E., Lee, K.L., and McKinnis, R.A., 1980, Procedures for large regression problems requiring maximum likelihood estimation, *in* Proceedings of the fifth annual SAS users group international conference, Cray, N.C., SAS Institute, p. 199-202.

Harrell, F.E., and Lee, K.L., 1985, The practical value of logistic regression, *in* Proceedings of the tenth annual SAS users group international conference: Cray, N.C., SAS Institute, p. 1031-1036.

Hay, L.E., and Battaglin, W.A., 1990, Effects of land-use buffer size on Spearman's 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.

Helsel, D.R., and Ragone, S.E., 1984, Evaluation of regional ground-water quality in relation to land use, U.S. Geological Survey toxic waste--ground-water contamination program: U.S. Geological Survey Water-Resources Investigations Report 84-4217, 33 p.

Krulikas, R.K., 1986, Hydrologic appraisal of the Pine Barrens, Suffolk County, New York: U.S. Geological Survey Water-Resources Investigations Report 84-4271, 53 p.

Leamond, C.E., Haefner, R.J., Cauller, S.J., and Stackelberg, P.E., 1992, Ground-water quality in five areas of differing land use in Nassau and Suffolk Counties, Long Island, New York, 1987-88: U.S. Geological Survey Open-File Report 91-180, 86 p.

Long Island Regional Planning Board, 1982, Land Use 1981--Quantification and analysis of land use for Nassau and Suffolk Counties: Hauppauge, N.Y., 48 p.

_____ 1986, Special ground-water protection area project for the Oyster Bay pilot area and the Brookhaven pilot area: Hauppauge, N.Y., 100 p.

McClymonds, N.E., and Franke, O.L., 1972, Water-transmitting properties of aquifers on Long Island, New York: U.S. Geological Survey Professional Paper 627-E, 24 p.

New York State Department of Environmental Conservation, 1983, Draft Long Island groundwater management program: Albany, N.Y., 221 p.

REFERENCES CITED (continued)

_____ 1986, Final Long Island groundwater management program: Albany, N.Y., 206 p.

Rutledge, A.T., 1987, Effects of land use on ground-water quality in central Florida--preliminary results: U.S. Geological Survey Water-Resources Investigations Report 86-4163, 49 p.

SAS Institute Inc., 1988, SAS/STAT user's guide, release 6.03 edition: Cary, N.C., SAS Institute, 1,028 p.

Siwiec, S.F., and Stackelberg, P.E., 1989, Relating ground-water quality to land use--considerations of scale and data resolution, *in* Pederson, G.L., and Smith, M.M., (compilers), U.S. Geological Survey second national symposium on water quality--abstracts of the technical sessions, Orlando, Florida, November 12-17, 1989: U.S. Geological Survey Open-File Report 89-409, p. 90.

Trunzo, Caesar, and Bianchi, I.W., 1987, Proposal introduced to New York State Legislature, S. 2831, A. 3709, February 25, 1987, 7 p.

U.S. Bureau of the Census, 1985, Census geography--concepts and products: Factfinder for the nation, U.S. Bureau of the Census, CFF8 (Revised), 7 p.

Vowinkel, E.F., and Battaglin, W.A., 1989, Methods of evaluating the relation of ground-water quality to land use in a New Jersey Coastal Plain aquifer system, *in* Mallard, G.E., and Ragone, S.E., (eds.), U.S. Geological Survey toxic substances hydrology program--proceedings of the technical meeting, Phoenix, Arizona, September 26-30, 1988: U.S. Geological Survey Water-Resources Investigations Report 88-4220, p. 405-410.

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.