GROUND WATER SUSCEPTIBILITY TO ELEVATED NITRATE CONCENTRATIONS IN SOUTH MIDDLETON TOWNSHIP, CUMBERLAND COUNTY, PENNSYLVANIA

by

ELIZA L. GROSS

A thesis submitted to the

Department of Geography and Earth Science

In partial fulfillment of the

requirements for the degree of

Master of Science

In

Geoenvironmental Studies

Shippensburg University

Shippensburg, Pennsylvania

2008

Abstract

This study addresses factors responsible for ground water susceptibility to nitrate concentrations above 4 mg/L in South Middleton Township, Cumberland County, Pennsylvania. High concentrations of nitrate in ground water are problematic due to the adverse health impacts that are caused by consumption of drinking water containing elevated concentrations of nitrate. Studies suggest that these include methemoglobinemia in infants and non-Hodgkin's lymphoma in individuals partaking in the long-term consumption of water with nitrate concentrations greater than 4 mg/L. A review of the literature indicates that similar projects have commonly been conducted at national and regional levels, and this presents the need for a similar study to be performed at a local scale in order to increase knowledge regarding ground water quality at the local level.

Water quality data for 2001 were obtained from South Middleton Township for 190 privately owned domestic drinking water wells. Explanatory data regarding anthropogenic and hydrogeologic variables closely representing the landscape in 2001 were obtained for analysis and compiled in relation to 500-meter, 1,000-meter, and 1,500-meter buffers around wells. Statistical methods used to determine explanatory variables best at predicting nitrate concentrations exceeding a threshold of 4 mg/L in ground water included univariate analysis and logistic regression analysis. Models associated with each of the three buffers sizes were calculated and test statistics were analyzed in order to choose a final model. Final models for the three different buffer sizes yielded different variables, thus showing how differing variables will become statistically significant at various scales. These methods yielded a final model associated with the 500-meter buffer that included the variables of total nitrogen inputs and percentage of silt in soil. This model produced statistically significant results with model significance p-values less than 0.05, a p-value of 0.0752 for the Hosmer-Lemeshow goodness-of-fit test statistic, a maximum rescaled r-square value of 0.3502, and a percent concordance of 79.0. Conversely, the model did not have a predictive power that was great enough to determine the probability of elevated nitrate concentrations occurring across the entire township. The Pearson residual statistic was calculated for the final model, and mapping of the residuals revealed areas of poor prediction in the northern and south-central portions of the township.

The main difference between this study and other studies that have been performed is that a majority of the study area was located on karst terrain, the study was performed at the local level, and there may have been spatial autocorrelation issues associated with the dependent data. The predictive power of the correlations was not strong enough to predict nitrate concentrations exceeding 4 mg/L throughout the township. Therefore, there is a need for future research within the township involving a similar study that divides the study area by physiographic province or lithologic unit, that addresses a larger study area, or that utilizes different buffer sizes for explanatory variables.

The statistical significance of the correlations indicates that total nitrogen inputs and percentage of silt in soils impact ground water quality within the township. Findings associated with the study include differences in scale among variables and the applicability of these types of studies at the local scale. The meaning of these results is useful to local officials in charge of water and land management and enables the improvement of knowledge and awareness concerning the occurrence of nitrate in ground water within the township.

Table of Contents

Table of Contents (continued)

List of Figures

List of Figures (continued)

List of Figures (continued)

List of Tables

Chapter 1

Introduction

Ground water is an important natural resource utilized by over half of the people in the United States for drinking water (Nolan *et al.*, 2002). Contaminants in ground water commonly come from the land surface due to anthropogenic impacts, and some aquifers are more susceptible than others to elevated concentrations of contaminants (Canter *et al.*, 1987). In particular, nitrate is the most ubiquitous of ground water contaminants, since its chemical composition allows it to readily travel with surface runoff and penetrate ground water resources (Canter *et al.*, 1987). High concentrations of nitrate in ground water are problematic due to the adverse health impacts that are caused by consumption of drinking water containing elevated concentrations of nitrate (Canter, 1997). These health impacts especially impact newborns and infants, which is primarily why the United States Environmental Protection Agency (USEPA) established a drinking water standard of 10 mg/L for nitrate in 1992 (Canter, 1997). In addition, a 1996 study has suggested that there may be an increased risk for non-Hodgkin's lymphoma

associated with long-term consumption of water containing nitrate concentrations greater than 4 mg/L (Ward *et al.*, 1996).

The probability of high concentrations of nitrate occurring in ground water serves as an informative resource for officials in charge of water and land management. Communities containing large numbers of households obtaining drinking water from domestic wells are most at risk because these wells are typically more shallow than public supply wells and are not routinely monitored for water quality (Hitt & Nolan, 2005). Shallow wells are at risk for elevated nitrate concentrations because shallow ground water is more susceptible to nitrate occurrence than deep ground water (Hitt & Nolan, 2005). Therefore, since South Middleton Township, Cumberland County, Pennsylvania contains a significant number of households using domestic wells, it is important for this community to be aware of factors impacting elevated nitrate concentrations in ground water.

1.1 Background

Statistical vulnerability assessments regarding elevated nitrate concentrations in ground water have typically been performed at the national or regional level (Gurdak & Qi, 2006). Those studies involving the land area encompassed by South Middleton Township include a 2005 national scale study performed by Hitt and Nolan and a 2005 regional study performed by Greene *et al.* for the Mid-Atlantic region of the eastern United States. The lack of local level studies presents a need for these types of studies to be performed at a larger scale so that the data are more useful to local planning officials. South Middleton Township presents a feasible study area due to the substantial number of households using domestic wells and data availability regarding domestic wells within the township. A ground water study addressing nitrate concentrations would improve

current knowledge regarding ground water attributes within the township through the usage of the municipality's ground water quality data and explanatory, or independent, variables collected for the township and surrounding areas.

1.2 Purpose and Scope

The purpose of this study is to create a statistical model using logistic regression analysis in order to define those variables that best predict the probability of the occurrence of recent (2001) nitrate concentrations above 4 mg/L within South Middleton Township. Logistic regression analysis was utilized instead of other statistical methods, such as multiple linear regression, because of its ability to predict the probability of elevated nitrate concentrations occurring within the township rather than determining actual nitrate concentrations (Helsel & Hirsch, 1992). Predictions of actual nitrate values within a township provide management officials with predictive concentration values, while predicted probabilities present the chance of an event occurring. Therefore, the predicted probability of the occurrence of elevated nitrate concentrations in relation to a threshold is more useful to officials in charge of water and land management because decision-makers can draw more conclusions from a predicted probability than from a predicted value (Focazio *et al.*, 2002). Elements regarding risk and uncertainty issues associated with environmental phenomena, such as elevated nitrate concentrations in ground water, can be better interpreted through predicted probability maps that display the possibility of an occurrence (Focazio *et al.*, 2002).

A model was developed using logistic regression analysis and represents the relationship between concentrations of nitrate occurring above 4 mg/L and anthropogenic and hydrogeologic explanatory variables. Nitrate concentration data consists of 190 samples collected across South Middleton Township from December 2000 to March

2001. Anthropogenic variables are those mostly derived from human activities, and for this study, these include land cover, total nitrogen inputs, onsite waste disposal, and population density. Hydrogeologic variables are typically a result of the natural environment, and in this study, they consist of bedrock type, soil texture, soil hydrologic group, and surface depression and sinkhole densities. The final model is based on corresponding explanatory variables obtained from existing and constructed Geographic Information System (GIS) raster data.

Chapter 2

Literature Review

2.1 Nitrate and Ground Water

Nitrate (NO_3^-) forms in the environment from nitrogen (N), which is a nutrient used for plant growth (Makuch and Ward, n.d.). The four primary forms of nitrogen include organic nitrogen, ammonia nitrogen (NH_3), nitrite (NO_2^-), and nitrate (Canter *et al.*, 1987). Organic nitrogen is converted to nitrate through a process called nitrification (Makuch and Ward, n.d.; Canter *et al.*, 1987). Nitrification involves an aerobic reaction that is principally carried out by obligate autotrophic organisms, which are organisms that are able to synthesize their own food from simple organic material (Canter *et al.*, 1987). Through this process, microorganisms transform organic nitrogen into inorganic ammonium, nitrifying bacteria convert ammonium ions to nitrite, and nitrite is converted to nitrate by another bacterial form (Makuch and Ward, n.d.; Canter *et al.*, 1987).

Nitrogen enters the landscape via both nonpoint and point sources. Nonpoint sources include contamination areas of large extent (Winter *et al.*, 1998). For example, when nitrogen fertilizer and nitrogen-containing manures are applied to agricultural fields

in order to increase crop yields, these fields are considered nonpoint sources of nitrogen contamination (Makuch and Ward, n.d.). Once nitrogen is applied to the agricultural landscape and has undergone nitrification, the resulting nitrate can be readily used by plants since it is water soluble, thus causing it to be absorbed easily by plant roots (Makuch and Ward, n.d.). In addition, nitrate ions are not adsorbed to soil particles since both nitrate ions and soils have negative charges; therefore, nitrate is very mobile in both saturated and unsaturated soils (Canter *et al.*, 1987). In some cases, nitrate may not be absorbed by plants because it is applied to the landscape before crops are planted or after crops are harvested (Makuch and Ward, n.d.). Also, it may not be absorbed because there is an excess amount that cannot absorbed by crops that have already met their nitrate needs (Makuch and Ward, n.d.). If nitrate is not absorbed by plants, its mobility will cause it to readily enter ground water through rain or flood water seepage, and this is especially pertinent in areas with permeable soils (Makuch and Ward, n.d.).

In addition, septic tank systems can serve as nitrate nonpoint sources (Canter *et al.*, 1987). These systems collect wastewater, provide a tank for solids to settle out, and allow the separated effluent to percolate into the geology through a subsurface drainage system (Canter *et al.*, 1987). When septic tank systems are designed, built, maintained, or situated inadequately, they are more susceptible to leaching excessive nitrate into soils, thus threatening ground water quality (Canter *et al.*, 1987; Makuch and Ward, n.d.). Furthermore, even large densities of properly functioning septic tanks can cause an overabundance of nitrate to be released into soils, and septic tank systems situated in highly permeable soils can also cause nitrate to be released too rapidly (Canter *et al.*, 1987). When these instances occur, the effluent is not exposed to the removal

mechanisms associated with the soils because the soil is overloaded or the effluent is percolating too quickly through the soil (Canter *et al.*, 1987).

Conversely, point sources represent a single point of discharge, such as a small area with a concentration of livestock or a facility burning fossil fuels (Winter *et al.*, 1998; Canter *et al.*, 1987; Driscoll & Lambert, 2003). Sometimes livestock are held in small feedlots or barnyards, and these facilities can result in large amounts of animal waste being concentrated in a small area (Makuch and Ward, n.d.; Winter *et al.*, 1998; Canter *et al.*, 1987). This occurrence may lead to an overabundance of nitrate leaching through soils (Makuch and Ward, n.d.; Winter *et al.*, 1998; Canter *et al.*, 1987).

Also, facilities burning fossil fuels release nitrogen emissions, which are deposited on land and water surfaces as nitrate in precipitation (Driscoll & Lambert, 2003). The deposition of these emissions across the landscape can cause nitrate to easily enter surface runoff (Canter *et al.*, 1987). Therefore, nitrogen deposition can cause nitrate to percolate through soils with the surface runoff, thus impacting ground water quality (Canter *et al.*, 1987).

Once nitrate reaches the land surface and leaches into ground water, it is capable of traveling significant distances as long as the lithologic materials are permeable and contain dissolved oxygen (Canter *et al.*, 1987). This process becomes hindered when nitrate is not capable of reaching ground water supplies, which occurs through immobilization and denitrification (Canter *et al.*, 1987; Knox & Moody, 1991). Immobilization occurs when growing bacteria absorb nitrate. Bacteria will only absorb nitrate if there is a sufficient amount of organic matter available in the soil, which serves as a carbon food source for bacteria (Knox & Moody, 1991; Canter *et al.*, 1987). Denitrification occurs when there is a limited amount of oxygen in the environment;

therefore nitrate becomes substituted for oxygen by bacteria (Knox & Moody, 1991). This biological process is performed mainly by heterotrophs, which are organisms that require carbon for growth and development (Canter *et al.*, 1987). In the absence of oxygen, nitrate becomes an electron acceptor as heterotrophic bacteria respire organic matter (Canter *et al.*, 1987). Nitrate is converted into gaseous nitrogen through this process (Canter *et al.*, 1987).

Absence of oxygen in soil is often caused when the soil has a high moisture content; therefore, a soil that retains or stores moisture, such as a clay soil or hydric soils in wetlands, will lack oxygen (Canter *et al.*, 1987; Knox & Moody, 1991). In addition, since clay does not allow water to pass through it easily, a clay soil will store water along with any nitrate within the water, thus delaying nitrate from leaching into ground water (Canter *et al.*, 1987; Knox & Moody, 1991). On the other hand, a soil that allows water to easily pass through it, such as a sandy soil with more available oxygen, will not store moisture or retain nitrate within water; thus sandy soils are capable of allowing nitrate to leach more quickly into ground water without the occurrence of denitrification (Knox & Moody, 1991).

Excess concentrations of nitrate in ground water can have negative impacts on drinking water quality, thus leading to the identification of nitrate as a primary water contaminant (Makuch and Ward, n.d .; Killingstad *et al.*, 2002). Therefore, the Safe Drinking Water Act of 1974 required the EPA to set a drinking water standard for nitrate to which public water purveyors must adhere (Killingstad *et al.*, 2002; Makuch and Ward, n.d .; Nolan *et al.*, 2002). The maximum contaminant level (MCL) set by the EPA is 10 mg/L, but household or domestic wells used by many property owners are not regulated or monitored (Focazio *et al.*, 2006).

The drinking water standard of 10 mg/L is based on the harmful impacts that elevated nitrate concentrations in drinking water can have on infants (Makuch and Ward, n.d.). A certain bacteria found in infants will cause nitrate to convert to nitrite in an infant's digestive system, thus diminishing the ability of an infant's blood to carry oxygen (Makuch and Ward, n.d .; Killingstad *et al.*, 2002). This state results in methemoglobinemia due to the inadequate supply of oxygen in the infant's blood, and this condition is sometimes fatal (Makuch and Ward, n.d .; Nolan *et al.*, 2002; Killingstad *et al.*, 2002). In addition, a 1996 study suggests that there may be an increased risk for non-Hodgkin's lymphoma associated with long-term consumption of water containing nitrate concentrations greater than 4 mg/L (Ward *et al.*, 1996). Although the study presents findings showing an increased risk for non-Hodgkin's lymphoma, the significance of the risk was not great enough to be unquestionable (Ward *et al.*, 1996).

2.2 Similar Studies

Similar studies performed in the United States regarding ground water susceptibility to elevated nitrate concentrations in relation to various explanatory factors primarily include those performed by the United States Geological Survey (USGS) and were carried out at national, regional, and local scales (Figure 2.1 and Table 1). Interestingly, different types of explanatory variables were utilized in these studies to analyze ground water vulnerability to nitrate (Table 2). On the other hand, each of the studies utilized a logistic regression model in order to predict the probability of nitrate concentrations in ground water exceeding a certain threshold. Notably, statistically significant variables utilized in final models varied slightly among studies (Table 2). The variance in significant variables included in final models may be due to the fact that each study was performed at a different scale, such as national or regional, and this may have

enabled the studies covering smaller areas to pick up on local trends that may not be significant over larger areas.

Figure 2.1. Various locations of similar studies performed (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Rupert, 2003; Gardner & Vogel, 2005; Greene *et al.*, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006; LaMotte & Greene, 2007).

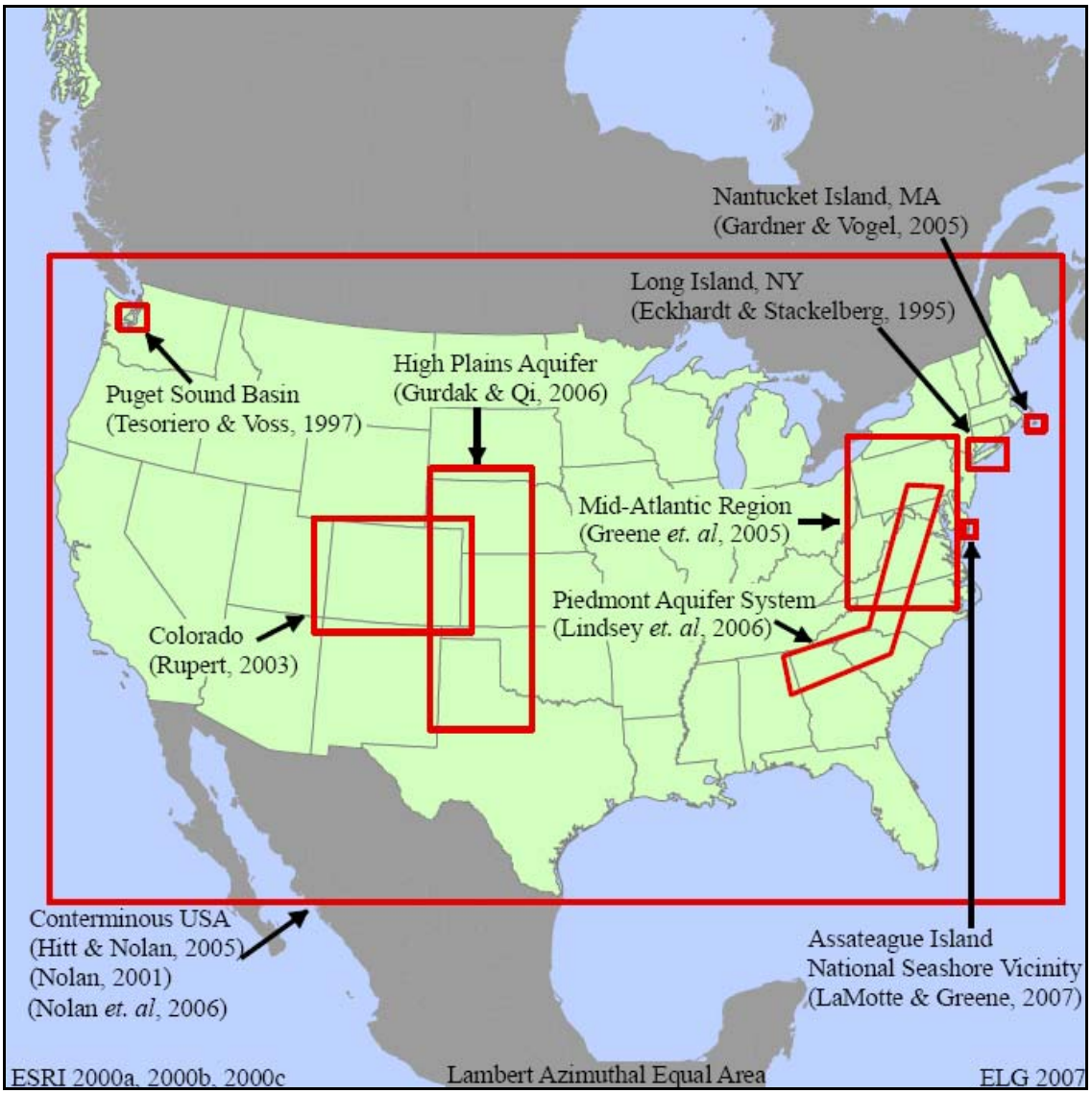


Table 1. Characteristics regarding various studies performed (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Rupert, 2003; Gardner & Vogel, 2005; Greene *et al.*, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006; LaMotte & Greene, 2007).

Study Characteristics	Study									
	Eckhardt & Stackelberg, 1995	Tesoriero & Voss, 1997	Nolan, 2001	Nolan *et. al*, 2002; Hitt & Nolan, 2005	Rupert, 2003	Gardner & Vogel, 2005	Greene *et. al*, 2005	Gurdak & Qi, 2006	Lindsey *et. al*, 2006	LaMotte & Greene, 2007
Study Area	Five Areas on Long Island, New York, USA	Puget Sound Basin in Washington, USA	Conterminous USA	Conterminous USA	Colorado, USA	Nantucket Island, Massachusetts, USA	Mid-Atlantic Region of USA	High Plains Aquifer in central USA	Piedmont Aquifer System of Eastern USA	Watershed adjacent to Assateague Island National Seashore, Maryland, and Virginia, USA
Study Area Size	285 to 570 km² (total of 5 study areas)	35,000 km²	9,629,091 km²	9,629,091 km²	269,837 km²	124 km²	466,198 km²	450, 660 km²	240,869 km²	1,179 km²
Number of Wells	90	1,967	900	1,280	655	69	927	336	260	529
Ratio of Study Area Size to Number of Wells	3 to 6 km² per well	18 km² per well	10,699 km² per well	7,523 km² per well	412 km² per well	2 km² per well	503 km² per well	1,341 km² per well	926 km² per well	2 km² per well
Threshold	3 mg/l	3 mg/l	4 mg/l	4 mg/l	2 mg/l, 5 mg/l, and 10 mg/l	2 mg/l	1 mg/l through 10 mg/l	4 mg/l	4 mg/l	3 mg/l
Contributing Area Buffer Radius	805 m	3,200 m	500 m	500 m	500 and 2,000 m	305 m	1,500 m	500 m	500 m	1,300 m
Number of Explanatory Variables	4	6	11	12	7	4	11	10	9	11
Number of Statistically Significant Explanatory Variables in Final Model	2	3	6	6	3	1	2	4	4	2

Table 2. Variables utilized in the various studies with statistically significant variables used in final models indicated (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Rupert, 2003; Gardner & Vogel, 2005; Greene *et al.*, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006; LaMotte & Greene, 2007).

	Variable	Study Utilizing Variable									
		Eckhardt & Stackelberg, 1995	Tesoriero & Voss, 1997	Nolan, 2001	Nolan *et. al*, 2002; Hitt & Nolan, 2005	Rupert, 2003	Gardner & Vogel, 2005	Greene *et. al*, 2005	Gurdak & Qi, 2006	Lindsey *et. al*, 2006	LaMotte & Greene, 2007
Anthropogenic Data	land use or land cover	O	O	O	O	X	O	O	O	O	O
	nitrogen inputs - atmospheric deposition							X	X	X	X
	nitrogen inputs - fertilizer applications			O	O	X		X	X	X	X
	nitrogen inputs - manure applications							X	X	X	X
	nitrogen inputs - total fertilizer and manure applications and atmospheric deposition			X	X					O	
	population density	O	X	O	O			X			X
	septic tanks - number of						X				
	well depth - depth of well or sampling depth		O	X	X		X				
Hydrogeologic Data	geology - presence or absence of rock fracture			O	X						
	geology - surficial geology		O					O			X
	ground water - depth to water table			O	O		X	X	O	X	X
	ground water - recharge		X								
	ground water - specific conductivity	X									
	hydrogeomorphic regions					X					
	precipitation - mean annual precipitation			X	X						
	soil - artificially drained soils			X	X						
	soil - available water capacity					O			X		
	soil - flood frequency of								X		
	soil - hydrologic soil groups		X	O	O	X		X		X	O
	soil - layer depth	X						X			X
	soil - organic matter			X	X	O		X	O	O	X
	soil - texture				O	O		X	O	O	X
	soil - universal soil loss factor								X		
	Explanation:										
	X and **O** indicate variable was utilized in study										
	O indicates variable was statistically significant and utilized in final model										

Of the studies examined, the one completed at the largest scale was performed by Gardner and Vogel in 2005 for Nantucket Island, Massachusetts with a study area of 124

km². Conversely, the studies performed at the smallest scale were USGS studies completed by Nolan in 2001 and by Nolan *et al.* in 2002 and Hitt and Nolan in 2005 for the conterminous United States. Other studies regarding elevated nitrate concentrations in ground water include study areas consisting of several watersheds of various sizes: the Mid-Atlantic region of the United States, the state of Colorado, and a group of five small areas on Long Island, New York (Tesoriero & Voss, 1997; Gurdak & Qi, 2005; Lindsey *et al.*, 2006; LaMotte & Greene, 2007; Greene *et al.*, 2005; Rupert, 2003; Eckhardt & Stackelberg, 1995).

Previous studies were completed with varying sample sizes and various ratios of study area size to sample size. The study utilizing the smallest number of wells to determine explanatory variables impacting elevated nitrate concentrations in ground water was a 1995 study performed by Eckhardt and Stackelberg for five small areas on Long Island, New York. Eckhardt and Stackelberg's 1995 study utilized 90 wells for a study area ranging from 285 to 570 km², which means that the ratio of study area size to number of wells was 3 to 6 km² per well. On the other hand, the study using the largest number of wells was the 1997 study performed by Tesoriero and Voss for Puget Sound Basin, Washington. This study utilized 1,967 wells to determine explanatory variables most responsible for impacting elevated nitrate concentrations in ground water for a study area of 35,000 km², thus establishing a ratio of 18 km² per well (Tesoriero & Voss, 1997). Furthermore, the study with the largest ratio of study area size to sample size was the 2001 study performed by Nolan for the conterminous United States with a ratio of 10,699 km² per well. Conversely, the study with the smallest ratio of study area size to sample size was the 2007 study completed by LaMotte and Greene for a watershed adjacent to Assateague Island National Seashore in the states of Maryland and Virginia with a ratio

of 2 km² per well. Among the ten studies discussed, each study utilized logistic regression analysis in order to predict the probability of nitrate concentrations exceeding certain thresholds.

Logistic regression analysis is applicable for these types of studies because it is capable of identifying a dichotomous response between independent and dependent variables, such as predicting the presence of nitrate concentrations above a specified threshold (Gurdak & Qi, 2006; Greene *et al.*, 2005). When nitrate concentrations in milligrams per liter are put into classes based upon a specific threshold value, the dataset is converted from a continuous variable into a categorical variable (Greene *et al.*, 2005). For example, based on a determined threshold value of 4 mg/L, a dataset containing nitrate concentrations in mg/L would have all nitrate values below 4 mg/L reclassified as zeros to represent nonevents, while all concentrations equal to or exceeding 4 mg/L would be reclassified as ones to represent events. This reclassification of nitrate concentrations according to a specific threshold value to create a variable in binary format presents a need for researchers to understand thresholds and to establish scientifically sound reasoning for selecting specific threshold values (Greene *et al.*, 2005).

Each of the studies of interest used similar threshold values in order to convert continuous nitrate concentration datasets into categorical binary datasets (Table 1). Studies performed by Eckhardt and Stackelberg (1995), Tesoriero and Voss (1997), and LaMotte and Greene (2007) used threshold values of 3 mg/L when creating categorical datasets. The threshold value of 3 mg/L was chosen for each of these studies because background or natural concentrations of nitrate in the environment are typically below 3 mg/L (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; LaMotte and Greene,

2007). Therefore, separating data categorically according to a threshold of 3 mg/L enables nitrate concentrations that are thought to be elevated due to anthropogenic activities to be classified as events. Additionally, studies completed by Nolan (2001), Nolan *et al.* (2002), Hitt and Nolan (2005), Gurdak and Qi (2006), and Lindsey *et al.* (2006) utilized a threshold level of 4 mg/L. A threshold of 4 mg/L was chosen for these studies because it is well over reported background levels, other studies commonly used the same threshold, and the threshold has been associated with adverse human health impacts (Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006). The 2003 study performed by Rupert used thresholds of 2, 5, and 10 mg/L in order to cover a broad range of nitrate concentrations. The threshold of 2 mg/L was chosen because background concentrations are generally less than 2 mg/L in the study area (Rupert, 2003). The 5 mg/L threshold was chosen because it is one-half the MCL established by the EPA (Rupert, 2003). Finally, the 10 mg/L threshold was chosen because it is the reported drinking water standard (Rupert, 2003). A 2005 study performed by Greene *et al.* examined thresholds of 1 mg/L through 10 mg/L and determined 10 mg/L as the maximum threshold for the study because it is the MCL established by the EPA.

Each of these studies used statistical analyses with different combinations of explanatory variables in order to determine which variables were the best predictors of elevated nitrate concentrations (Table 2). The studies examined anthropogenic data such as land use and land cover, various nitrogen inputs, population density, number of septic tanks, and depth of sampled wells (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Rupert, 2003; Gardner & Vogel, 2005; Greene *et al.*, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006; LaMotte &

Greene, 2007). In addition, the studies focused on hydrogeologic data such as geologic factors, ground water characteristics, hydrogeomorphic regions, mean annual precipitation, and various soil attributes. Also, each study included some sort of land use or land cover variable. This variable most often described the amount of agricultural land located within a specific vicinity of each well or sample point.

After explanatory variables were chosen for each study, the explanatory data were extracted from all of the datasets in relation to a specific buffer representing a contributing area around each well or sample point (Table 1). Buffers differed from study to study and sometimes even from variable to variable within each study. A majority of the studies utilized a contributing area buffer of 500 meters because a 500-meter radius was assumed by many USGS studies to be generally related to the recharge area of each well, but this radius was not intended to be exactly the same as each well's recharge area (Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006). Eckhardt and Stackelberg (1995) chose a data extraction radius of 805 meters (0.5 miles) because ground water within the Long Island, New York study area was known to move less than 805 meters in 6 years, thus allowing any explanatory datasets created within that 6-year window of time to be utilized for the study. Similarly, Gardner and Vogel (2005) utilized a 305-meter (1,000-foot) radius for explanatory variable extraction because ground water within the Nantucket Island, Massachusetts study area was known to move less than 1,000 feet in 2.5 years, thus allowing the usage of 1999 land use data to examine 2001 nitrate concentrations.

On the other hand, studies performed by Tesoriero and Voss (1997), Rupert (2003), Greene *et al.* (2005), and LaMotte and Greene (2007) utilized a wide range of radii for explanatory variable extraction in order to determine the optimal radius that best

fit each study's nitrate concentration data through analysis of logistic regression models. For example, Greene *et al.* (2005) examined radii ranging from 500 to 4,000 meters in 500-meter increments and chose the best-fit model by determining that a radius of 1,500 meters maximized specific test statistics for nitrate concentrations above a threshold of 3 mg/L. Also, Rupert (2003) utilized similar methods to determine a best-fit model for agricultural and urban land cover types and yielded an optimum buffer size of 2,000 meters for agricultural land cover and an optimum buffer size of 500 meters for urban land cover, thus utilizing varying radii for different land cover characteristics.

Once explanatory variables were extracted according to specific buffer sizes and examined through statistical analyses, the most statistically significant variables were included in the final model for each study (Table 2). Explanatory variables in each study's final model for all but one study included those associated with land cover or land use (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Rupert, 2003; Gardner & Vogel, 2005; Greene *et al.*, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006; LaMotte & Greene, 2007). Other important anthropogenic explanatory variables that were statistically significant and included in final models were: fertilizer applications, total nitrogen inputs from fertilizer and manure applications and atmospheric deposition, population density, and well depth (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Lindsey *et al.*, 2006). Statistically significant hydrogeologic explanatory variables included presence or absence of rock fracture, surficial geology, depth to water table, available water capacity of soil, hydrologic soil groups, organic matter in soil, and texture of soil (Tesoriero & Voss, 1997; Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Rupert, 2003; Greene *et al.*, 2005; Gurdak & Qi, 2006; Lindsey *et al.*, 2006;

LaMotte & Greene, 2007). Interestingly, variables in final models for studies performed for the contiguous United States included total nitrogen inputs from fertilizer and manure applications and atmospheric deposition, well depth, mean annual precipitation, artificially drained soils, and organic matter in soils (Nolan, 2001; Nolan *et al.*, 2002; Hitt & Nolan, 2005). On the other hand, variables included in the final model for the study with the smallest study area were number of septic tanks, well depth, and depth to water table (Gardner & Vogel, 2005).

Many of the studies included validation of final logistic regression models, and most presented maps depicting the predicted probability of elevated nitrate concentrations occurring in ground water for each study area. Studies conducted by Nolan *et al.* (2002) and Hitt and Nolan (2005), Rupert (2003), Greene *et al.* (2005), and Gurdak and Qi (2006) validated final models with an independent dataset. Lindsey *et al.* (2006) attempted model validation using a subset of the original dataset, but this proved to be unsuccessful because an inadequate number of well data points were used to validate the model.

Studies performed by Eckhardt and Stackelberg (1995), Tesoriero and Voss (1997), Nolan *et al.* (2002) and Hitt and Nolan (2005), Rupert (2003), Greene *et al.* (2005), Gurdak and Qi (2006), and LaMotte and Greene (2007) resulted in maps depicting the probability of elevated nitrate concentrations exceeding a specific concentration in ground water. Conversely, the study completed by Lindsey *et al.* (2006) omitted predictive maps because additional data would need to be collected in a future study to accurately predict nitrate concentrations for the study area.

Chapter 3

Study Area

3.1 Location

South Middleton Township, Cumberland County, is located in south-central Pennsylvania. The township encompasses approximately 127 km² (49 mi²) with a 2000 population of 12,939 (Figure 3.1) (USGS, 2001; South Middleton Township, n.d.). The township was established in 1810 when it was divided from the area known as Middleton, thus forming both North and South Middleton Townships (South Middleton Township, n.d.). South Middleton Township is bordered on the north by Carlisle Borough, North Middleton, and Middlesex Townships, on the east by Monroe Township, on the south by York and Adams Counties, and on the west by Dickinson Township. In addition, South Middleton Township surrounds the Borough of Mount Holly Springs, but the borough is not part of the township.

Figure 3.1. Location of South Middletown Township within Pennsylvania with major streams, major roads, and populated places (PennDOT, 2007a; PennDOT, 2007b; PennDOT, 2007c; USGS, 1999b).

3.2 Topography

The topography within South Middleton Township is unique because the area encompasses sections of three different physiographic provinces (Figure 3.2). A majority of the township, including its entire northern half, lies in the Great Valley, which is the easternmost valley of the Ridge and Valley physiographic province (Thornbury, 1965). The township's lowest elevation of 136 meters is located in the Great Valley, or the Cumberland Valley as it is locally known, where the Yellow Breeches Creek exits the

township (Figure 3.3) (South Middleton Township, 1999). The southern part of the township is located on South Mountain, and this signifies the northernmost ridge of the Blue Ridge physiographic province (Thornbury, 1965). The township's highest elevation is located within this province at 481 meters, thus giving the township a relief of 345 meters (South Middleton Township, 1999). Finally, a small portion of the southeastern tip of the township is located in the Gettysburg-Newark Lowland, which is part of the Piedmont physiographic province, and elevations within this small area of the township remain similar to those within the Blue Ridge portion of the township (Thornbury, 1965).

Figure 3.2. Physiographic provinces of South Middletown Township (PennDOT, 2007c; PGS, 1998).

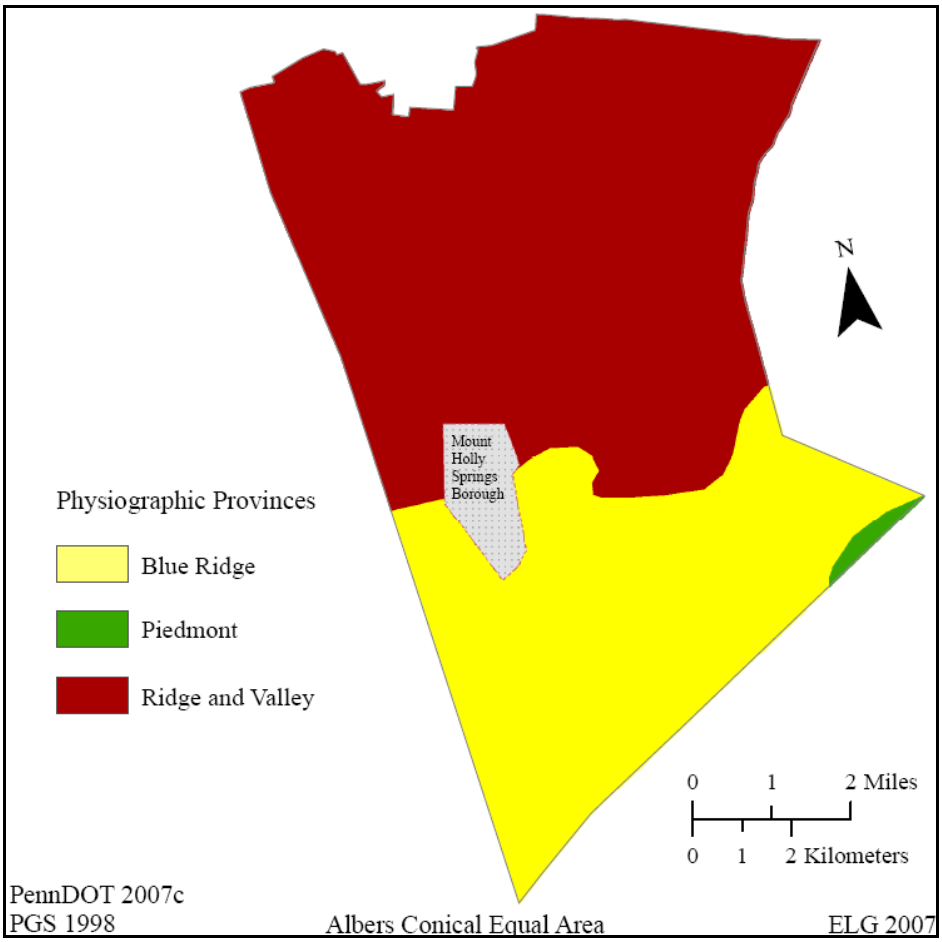
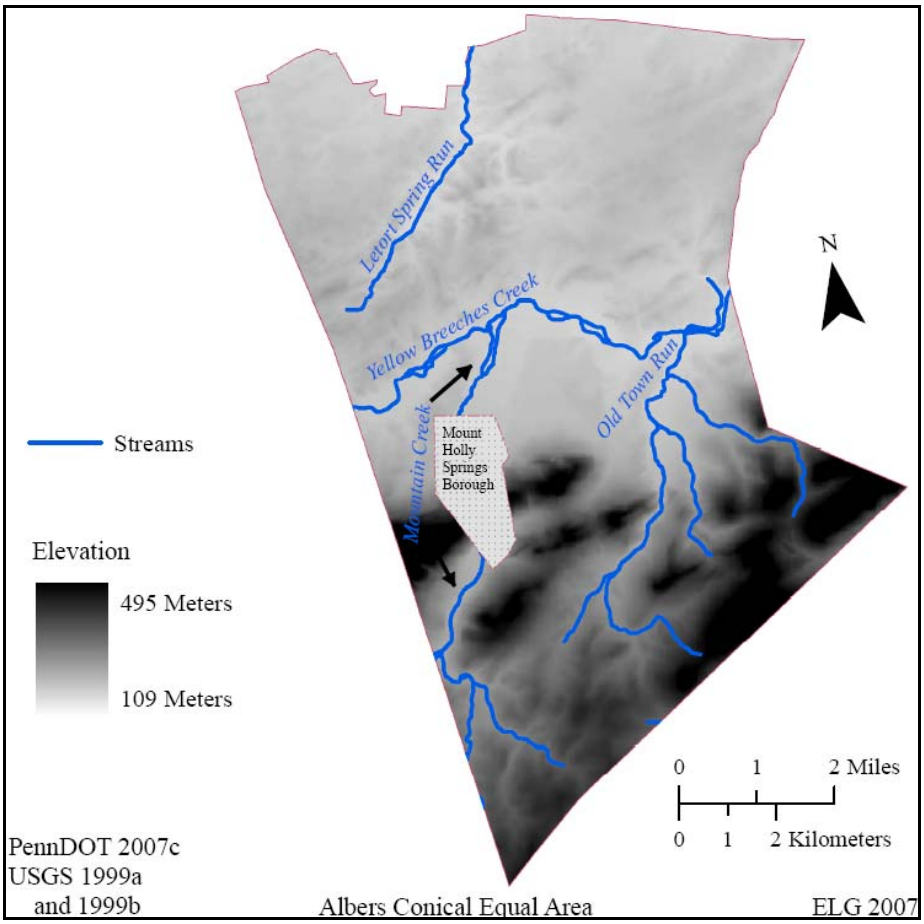


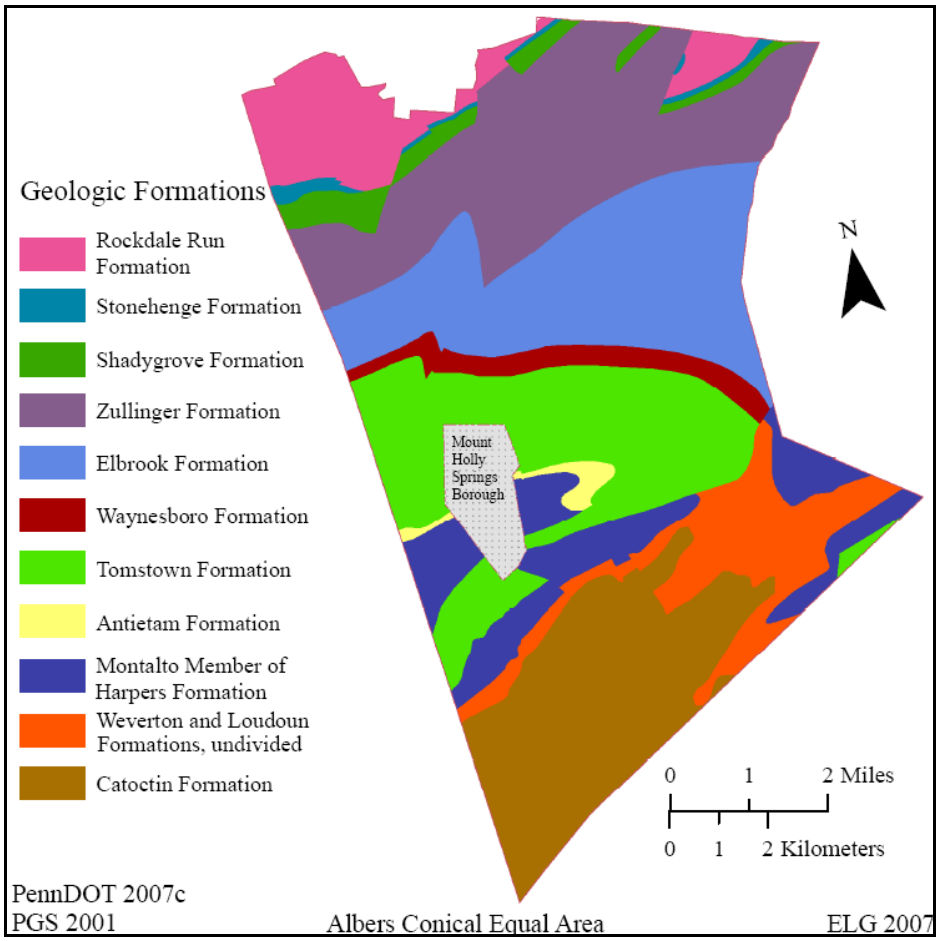
Figure 3.3. Topography of South Middletown Township with streams and elevation (PennDOT, 2007c; USGS, 1999a; USGS, 1999b).



3.3 Geology

The unique topography of South Middleton Township is directly influenced by its underlying geologic characteristics. South Middleton Township encompasses geologic formations from the Catoctin formation in the southeast portion of its boundary to the Rockdale Run Formation in the northwest (Figure 3.4). Geologic formations are discussed from oldest to youngest moving northwest through the township.

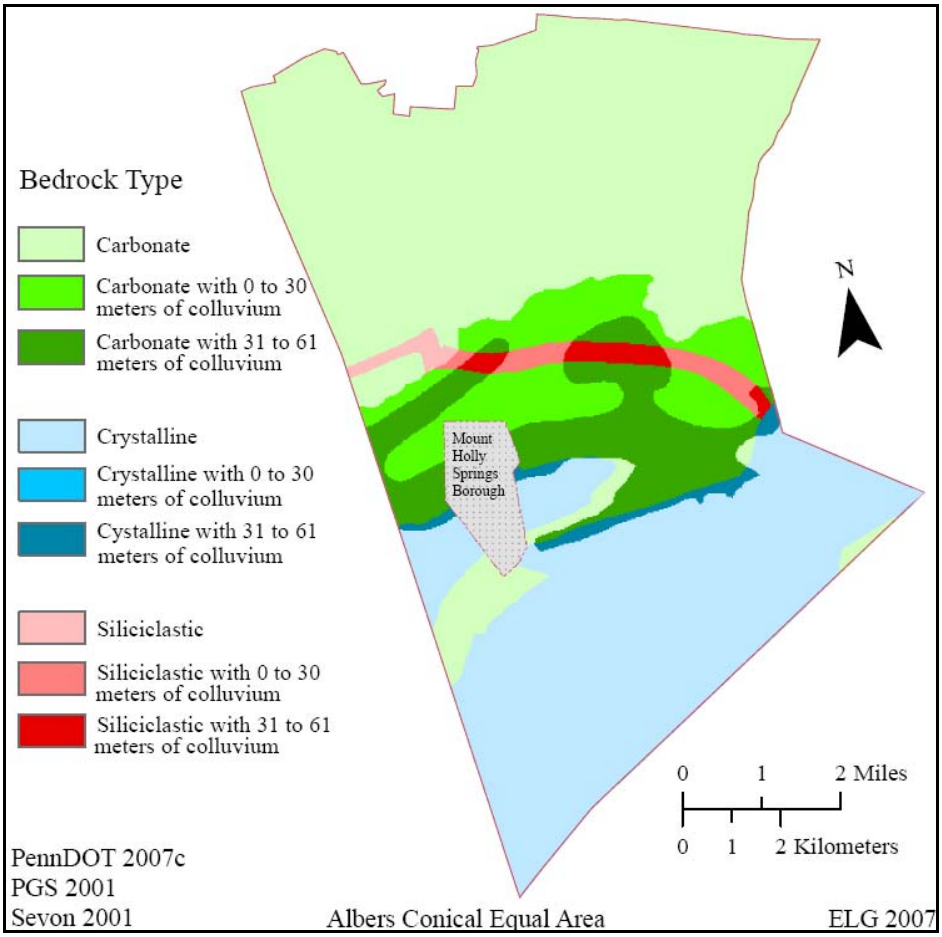
Figure 3.4. Geologic formations and faults of South Middletown Township (PennDOT, 2007c; PGS, 2001).



The Precambrian Catoctin Formation of South Mountain is composed of metarhyolite and metabasalt (Root, 1968). The lower Cambrian Chilhowee Group lies unconformably atop the Catoctin Formation and includes the Weverton, Harpers, and Antietam Formations (Root, 1968). These formations consist of rough clastics overlain by a carbonate lithology of limestone and dolomite with interbedded mudstones (Root, 1968; Shirk, 1980; Way, 1986). Also, the immense and well-bedded lower Cambrian Tomstown Formation flanks the Chilhowee Group, and it is composed of limestone and medium to dark gray dolomite (Shirk, 1980; Root, 1968). This formation forms a rolling lowland that is entirely covered at the base of South Mountain by a thick colluvium and

alluvium stratum that was deposited during the Tertiary and Quaternary time periods due to mass wasting processes and the heavily loaded streams that once ran down the slopes of the mountain (Figure 3.5) (Root, 1968; Becher & Root, 1981b; Shirk, 1980). This stratum reaches a maximum thickness of 61 meters in South Middleton Township (Sevon, 2001).

Figure 3.5. Generalized bedrock types in South Middletown Township with colluvium stratum indicated (PennDOT, 2007c; PGS, 2001; Sevon, 2001).



Next, the lower Cambrian Waynesboro Formation borders the Tomstown Formation and consists of carbonate limestone and dolomite at its central portion with resistant sandstone ridges at its edges that consist of shale and siltstone (Root, 1968; Shirk, 1980). The middle Cambrian Elbrook Formation, which consists of limestone and more resistant shale interbedded with dolomite, can be found to the west of the

Waynesboro Formation (Root, 1968; Shirk, 1980). The northern section of the upper Cambrian Conococheague Group, which includes the Zullinger and Shadygrove Formations, runs parallel to the Elbrook Formation and also consists primarily of limestone and dolomite (Root, 1968; Shirk, 1980). To the west, the lower Ordovician Beekmantown Group consists of the Stoufferstown, Stonehenge, Rockdale Run, and Pinesburg Station Formations, which are primarily made up of limestone and dolomite with interbedded clay and chert (Shirk, 1980; Root, 1968). Notably, the Beekmantown Group is sometimes over 3 miles wide, and it is considered to be the focal point of the Cumberland Valley's carbonate region (Shirk, 1980).

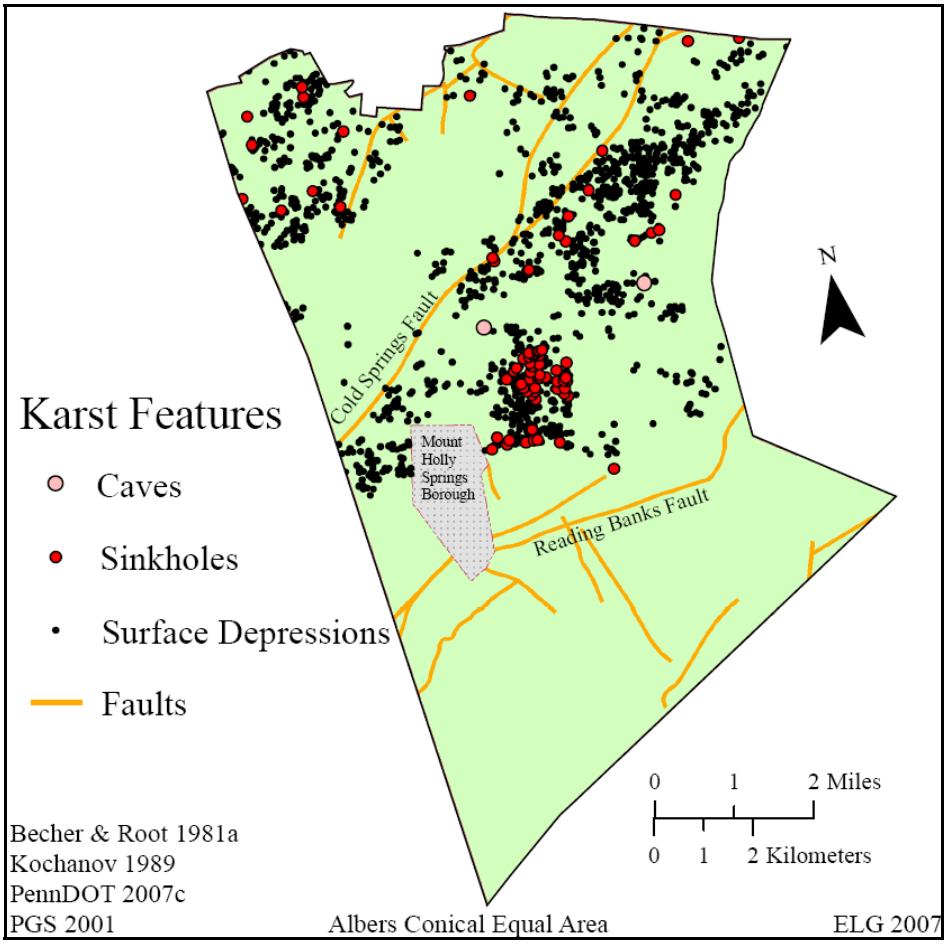
3.4 Hydrogeology

Ground water contamination issues are especially pertinent in areas possessing limestone and dolomite bedrock because the dissolution of these lithologic features results in the creation of karst terrain (Winter *et al.*, 1998). Carbonate waters that are produced by the solution of limestone and dolomite have high ionic strength and are a result of the dissolution process that creates enlarged fractures and solution holes in bedrock (Winter *et al.*, 1998; South Middleton Township, 1999). When solution holes become enlarged, ground water flow rates increase, thus ground water will flow across a larger surface area of exposed bedrock (Winter *et al.*, 1998). The increased flow further stimulates the dissolution process, and over time, surface depressions, sinkholes, or caves may form (Winter *et al.*, 1998). As bedrock is dissolved and is no longer quite as capable of supporting the land surface, surface depressions of various sizes will form (Winter *et al.*, 1998). When the bedrock becomes dissolved to the point that it can no longer support the land surface, the surface will cave in and form a sinkhole (Winter *et al.*,

1998). As the dissolution process continues, underground caves will form in the bedrock over time (Winter *et al.*, 1998).

South Middleton Township contains various karst features, such as surface depressions, sinkholes, and caves that have formed as a result of the weathering of the limestone and dolomite bedrock located in the portion of the township known as the Cumberland Valley (Figure 3.6) (Kochanov, 1989). Overall, the township has 1,274 surface depressions, 73 sinkholes, and 2 caves (Kochanov, 1989). A majority of surface depressions and sinkholes are located in the central part of the township close to Mount Holly Springs Borough, and located slightly north of this cluster of features are both of the township's caves. Another cluster of surface depressions and sinkholes is located just north of the caves and extends in a northeast band across the township. One last cluster of surface depressions and sinkholes can be found in the northwestern corner of South Middleton Township.

Figure 3.6. Locations of surface depressions, caves, sinkholes, and faults within South Middleton Township (Becher & Root, 1981a; Kochanov, 1989; PennDOT, 2007c; PGS, 2001).



The sinkholes and other solution-enlarged cracks that intersect the land surface of the township form infiltration paths to the ground water, and these lead to contamination issues due to pollutants in precipitation or surface runoff that quickly reach ground water (Winter *et al.*, 1998). For example, when surface runoff from a farm field enters a nearby sinkhole, nitrate is capable of quickly reaching the ground water by traveling with the surface runoff. In addition, malfunctioning septic tanks located in karst landscapes are capable of releasing raw sewage into fractures, thus causing contaminants to quickly enter the ground water (South Middleton Township, 1999). Since septic tanks can easily contaminate ground water supplies in karst terrain, ground water movement is an

important consideration for an area such as South Middleton Township where both septic tanks and privately owned domestic wells are widely used in residential areas (South Middleton Township, 1999).

Ground water movement in karst terrain is difficult to predict, but it can be assumed that ground water moves along fault lines, fractures, and through other weak areas in the bedrock (Winter *et al.*, 1998; South Middleton Township, 1999). South Middleton Township contains numerous faults, sinkholes, and surface depressions, thus indicating areas of structural weakness that impact the direction of ground water flow (Figure 3.6) (South Middleton Township, 1999). A prominent fault within the township, known as the Reading Banks Fault, is located in the township's central portion where the crystalline bedrock of South Mountain meets the carbonate bedrock of the Great Valley. Another noticeable fault, the Cold Springs Fault, follows a northeast path through the north-central portion of the township. Interestingly, both of the township's caves and a large amount of surface depressions and sinkholes can be found between the Reading Banks Fault and Cold Springs Fault. Although the movement of ground water is difficult to predict in the valley, it is assumed that ground water movement in the foothills of South Mountain is very different than ground water movement within the karst terrain of the Great Valley. Ground water in mountainous terrain is typically discharged at the base of steep slopes, at the edges of flood plains, or directly to valley streams (Winter *et al.*, 1998).

3.5 Land Cover and Planning

Land cover within South Middleton Township follows the expected pattern with developed lands located primarily in the flat Great Valley and forested lands located predominantly in steeper South Mountain region (Figure 3.7). The township is largely

made up of agricultural lands with 43.1% (55 km²) of the township consisting of pastures and cultivated lands (Table 3 and Figure 3.8). Most of the agricultural lands are located in the northern portion of the township where the abundance of limestone and dolomite in the Great supply a suitable structure and mineral content for the creation of prime agricultural soils (South Middleton Township, 1999). The gentle slopes and deep, well-drained soils of this part of the township facilitate an environment that is attractive to both farmers and developers (South Middleton Township, 1999). Forested lands make up 40.2% (51 km²) of the township, and a majority of this land cover type is located in the southern, more mountainous portion of the township where the crystalline bedrock and steeper slopes provide an unsuitable landscape for agricultural lands or development. Developed land is also primarily located in the Great Valley and constitutes 14.8% (19 km²) of the township. Developed lands are located within proximity of Boiling Springs and the boroughs of Mount Holly Springs and Carlisle. South Middleton Township contains few wetlands and open water areas; wetlands make up 1.5% (2 km²) of the township while open water constitutes 0.4% (1 km²) of the land area. These areas are generally located in the central part of the township where the Yellow Breeches Creek flows through the study area.

Figure 3.7. Land cover patterns in South Middletown Township (PennDOT, 2007c; USGS, 2001).

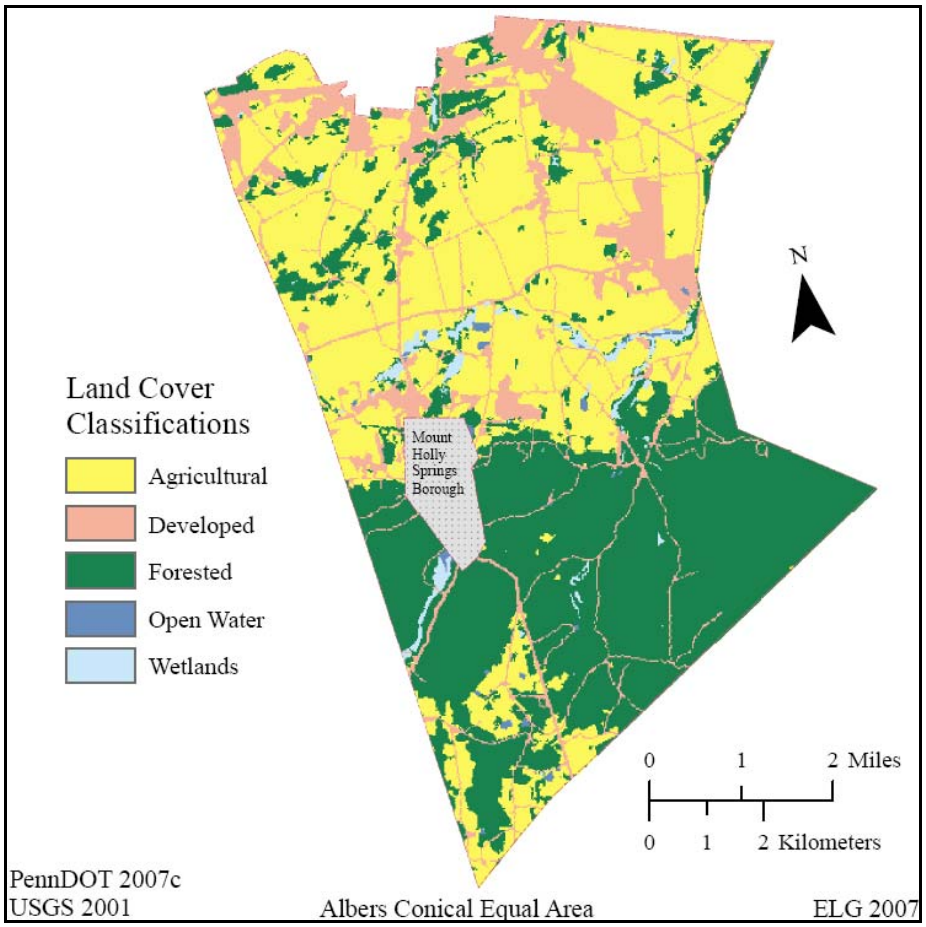
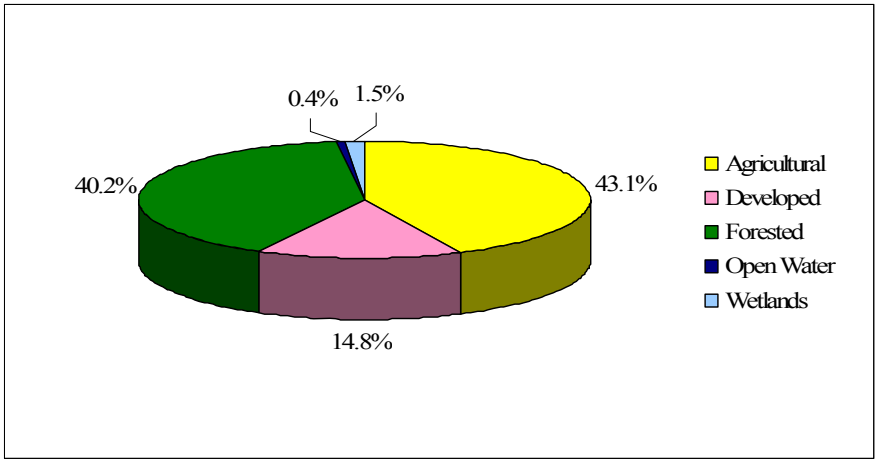


Table 3. Percentage and area of different land cover types within South Middleton Township (USGS, 2001).

Land Cover Classifications	Area (km^2)	Area (mi^2)	Area (Percent)
Agricultural	55	21	43.1%
Developed	19	7	14.8%
Forested	51	20	40.2%
Open Water	1	0	0.4%
Wetlands	2	1	1.5%
Total	**127**	**49**	**100%**

Figure 3.8. Graphical representation of percentage of different land cover types within South Middleton Township (USGS, 2001).



While there were currently few residential areas under development in 1999, the possibility for future residential development in agricultural areas at the time was high due to the moderate slopes possessed by these areas that are ideal for development (South Middleton Township, 1999). Many residential areas within the township utilize onsite waste disposal methods such as septic systems or sand mounds (Figure 3.9) (South Middleton Township, 1999). Only small portions of the township surrounding Boiling Springs, Mt. Holly Springs Borough, and Carlisle Borough had access to sewer lines as of 2001; therefore, a majority of land parcels within the township utilize some type of onsite waste disposal method (Cumberland County Planning Commission, 2001). Both current and future development within proximity of well-drained soils presents an increased potential for negative ground water impacts due to the nitrate-rich manures and fertilizers commonly applied to agricultural landscapes and the septic systems used by residential areas that release nitrate into the soil (Makuch and Ward, n.d.; South Middleton Township, 1999). The karst terrain allows contaminants to easily reach ground water supplies that are accessed by private wells, which are quite common within the township (Figure 3.10). As of 2001, a similar number of land parcels were serviced

by public water suppliers as were serviced by public sewer lines with a few more areas near Boiling Springs and Mount Holly Springs Borough being serviced by sewer lines. Ultimately, the fertilizers and manures being applied to large areas of agricultural lands and the high densities of parcels with onsite waste disposal within the township have the potential to negatively impact the quality of drinking water being accessed by private wells within areas containing many karst landforms.

Figure 3.9. Parcels in South Middleton Township that were serviced by public sewer or that utilized onsite waste disposal methods in 2001 (Cumberland County Planning Commission, 2001; PennDOT, 2007c).

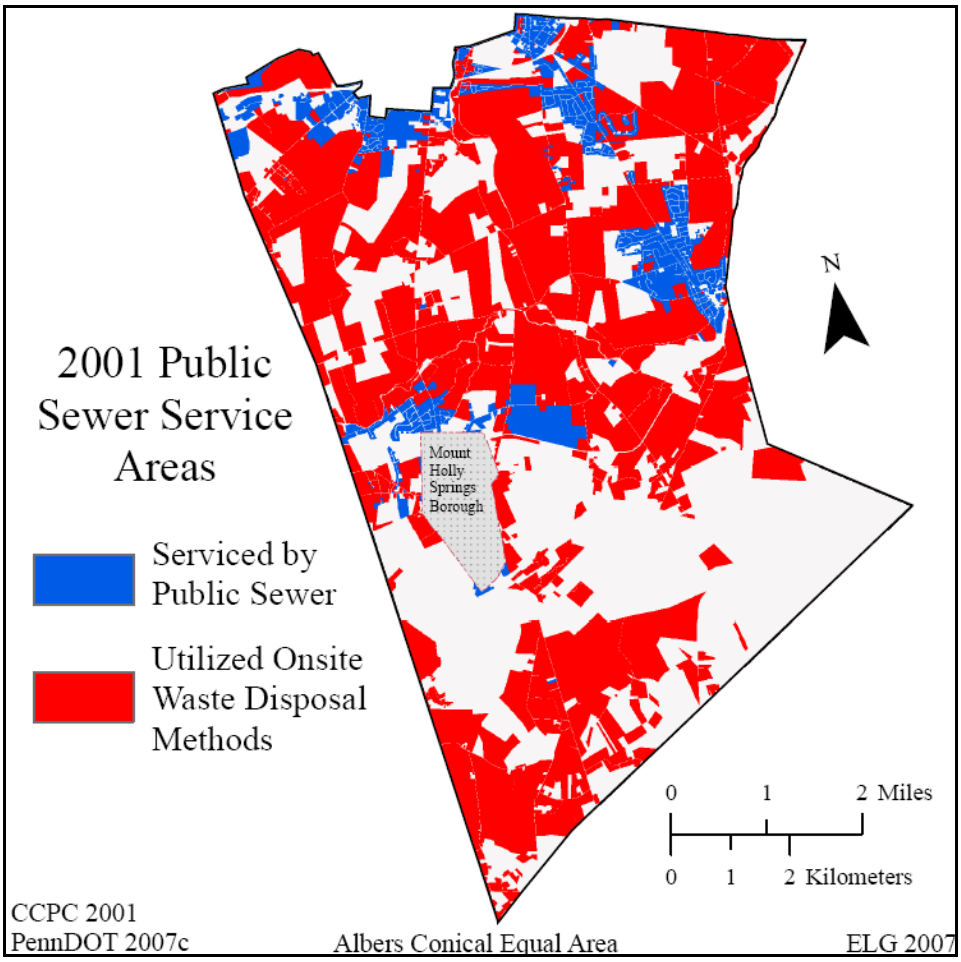
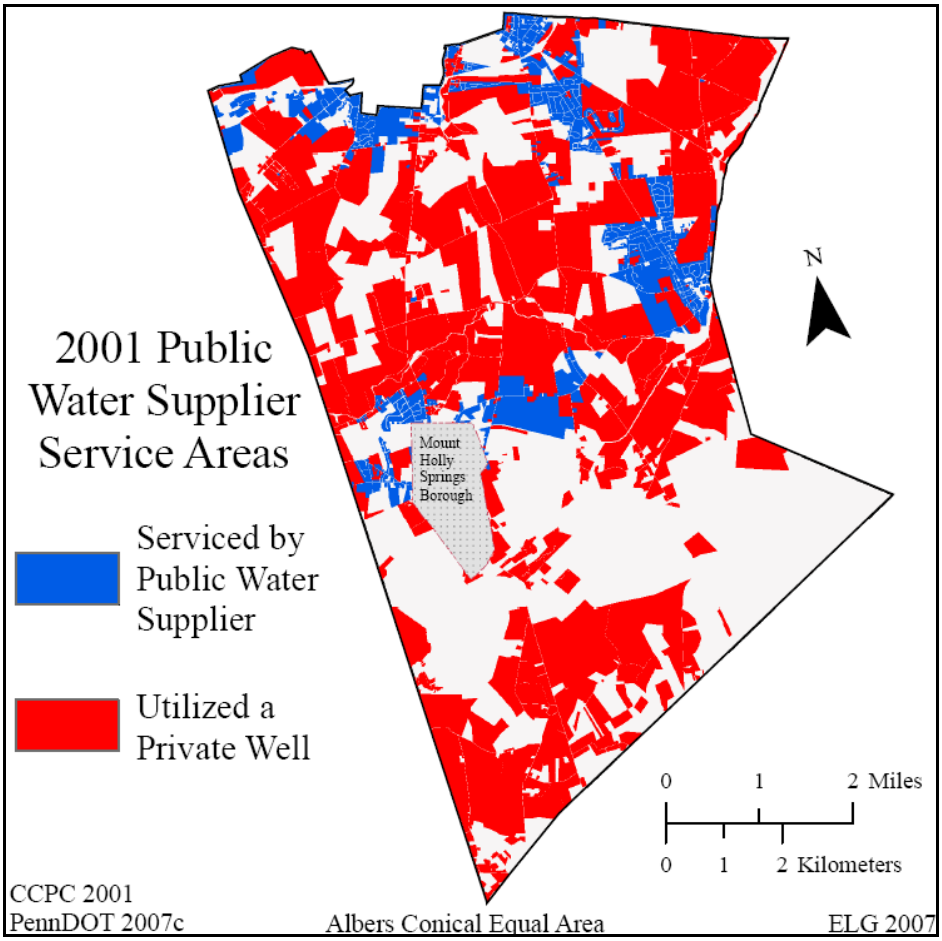


Figure 3.10. Parcels in South Middleton Township that were serviced by a public water supplier or that utilized a private well in 2001 (Cumberland County Planning Commission, 2001; PennDOT, 2007c).



Chapter 4

Methods

Methods associated with this project include data compilation and extraction, development of logistic regression models, and evaluation of the final model's performance. Primarily, in order to assess elevated nitrate concentrations occurring in ground water using logistic regression, a dependent variable and explanatory variables were identified and compiled. The dependent variable consists of concentrations of nitrate as nitrogen for 190 privately owned domestic drinking water wells due to the potential health risks associated with elevated nitrate concentrations and the availability of this data for the study area (Canter, 1997; South Middleton Township, 2001).

Explanatory variables utilized in the study include both anthropogenic and hydrogeologic variables, including land cover, nitrogen inputs from atmospheric deposition and from farm and non-farm fertilizer and manure applications, onsite waste disposal, population density, bedrock type, soil texture, soil hydrologic group, surface depression and sinkhole densities, and percent slope. These data were compiled for South Middleton Township and surrounding municipalities and extracted according to

500-meter, 1,000-meter, and 1,500-meter buffers surrounding each well. These processes were performed utilizing a GIS, which consists of computer hardware and software and data management and analytic techniques that are used to compile, analyze, and display geographic data.

The resulting data were analyzed using univariate statistical analysis and logistic regression analysis. A different logistic regression model was created for each buffer size, and a final model was selected based on the maximization of test statistics. The final model was then evaluated in order to determine how well the final model fit the nitrate concentration data.

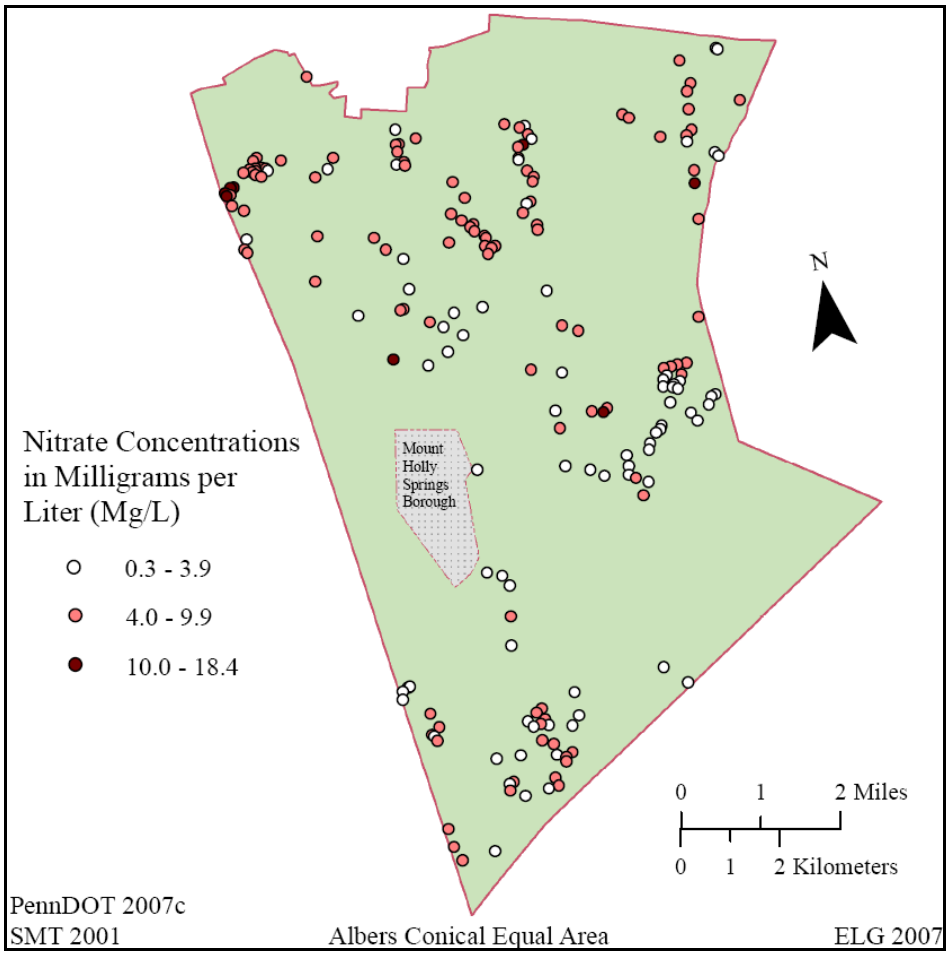
4.1 Data Description

4.1.1 Dependent Variable

The dependent variable consists of a ground water quality dataset with nitrate concentrations in mg/L for 190 privately owned domestic drinking water wells in South Middleton Township (Figure 4.1) (Appendix A) (South Middleton Township, 2001). The ground water quality data for the wells were collected over a 51-day time period with the first well sample collected on December 26, 2000 and the last water quality sample collected on March 6, 2001 (South Middleton Township, 2001). Grab samples were collected from various kinds of taps, such as taps located outside or indoors (South Middleton Township, 2001). Grab samples collected indoors typically came from kitchens, houses, barns, or garages (South Middleton Township, 2001). Some grab samples were also collected from bathrooms, hydrants, kitchen sinks, pressure tanks, and pressure taps (South Middleton Township, 2001). Since all nitrate concentrations were obtained from domestic wells and well depths were unknown, it was assumed that all of the wells were more shallow than public supply wells, which typically penetrate much

deeper aquifers that are not as susceptible to elevated nitrate concentrations (Hitt & Nolan, 2005).

Figure 4.1. Location of wells with associated nitrate concentrations in South Middleton Township (PennDOT, 2007c; South Middleton Township, 2001).



Ground water quality samples were collected through the township in compliance with Act 537, the Pennsylvania Sewage Facilities Act (South Middleton Township, 2001; PADEP, 2006). The Pennsylvania Sewage Facilities Act was enacted in 1968 by the Pennsylvania Department of Environmental Resources (PADER), which is now known as the Pennsylvania Department of Environmental Protection (PADEP), in order to mitigate sewage disposal issues and prevent future problems (PADEP, 2006). The Pennsylvania Sewage Facilities Act requires that municipalities plan for and monitor

community and individual sewage systems within their jurisdictions through the submittal of plans, authorization of grants, requirement of permits for sewage systems, and permission for state departments to administer rules, regulations, standards, and procedures (PADEP, 2006). In order to meet the planning and monitoring requirements of the Pennsylvania Sewage Facilities Act, the On-Lot Septic Ordinance was created in 2000 and implemented throughout South Middleton Township (South Middleton Township, 2000). The township's On-Lot Septic Ordinance authorizes the inspection of all on-lot sewage disposal systems by an authorized agent (South Middleton Township, 2000). The associated inspections are permitted to include physical inspection of any property of interest, attainment of sewage disposal system samples, and obtainment of surface water, well, or other ground water samples (South Middleton Township, 2000). The township's need to comply with the Pennsylvania Sewage Facilities Act and the township's On-Lot Septic Ordinance reveal the premise for the creation of the dataset being used as a dependent variable for this study.

The original ground water quality dataset obtained from the township did not include geographic coordinates that would have enabled each sampled well to be represented as a point on a map. Instead, the dataset included land parcel identification values, so the data were related to a South Middleton Township parcel polygon dataset obtained from the Cumberland County Planning Commission (2001) using parcel identification values. Next, the centroid for each land parcel containing ground water quality data was generated in order to create points representing each well from which the grab samples were taken.

Also, the generated locations for each well were validated using PAMAP 2003 orthoimages of South Middleton Township (USGS, 2004). The USGS (2004) PAMAP

2003 data are 2-foot pixel resolution orthoimages collected and distributed through a joint collaboration of the Pennsylvania Geological Survey (PGS) and the USGS. Data validation using the orthoimages utilized the basic assumption that domestic drinking water wells are not located beneath private homes, garages, or other large buildings. Therefore, when any of the wells that were plotted as a parcel centroid overlapped buildings on the 2003 orthoimages, the points were moved to the nearest area not overlapping a building according to the 2003 orthoimages. The original ground water quality dataset included samples for 200 wells, but 10 samples were omitted from the study dataset. Three of the wells were located on land parcels that were not within South Middleton Township's political boundary. The rest of the omitted data had land parcel identification numbers associated with more than one parcel. Instead of judging which of the two parcels with the same identification numbers with which to associate a well, the data were omitted from the study in order to reduce data inaccuracy. Parcels ranged in size from less than 1 km^2 to 5.4 km^2, so wells within larger parcels had a higher susceptibility of being inaccurately placed (Cumberland County Planning Commission, 2001). Since the average parcel size was less than 1 km^2, it was assumed that the placement of a majority of the well data points was fairly accurate (Cumberland County Planning Commission, 2001).

4.1.1.1 Variability

Nitrate concentrations in ground water vary according to natural processes, such as changing seasons or varying rainfall amounts (Reese & Lee, 1998). In Pennsylvania, the most significant amount of ground water recharge occurs in October through November and March through April (Reese & Lee, 1998). The ground is typically frozen throughout the winter months and evapotranspiration by plants occurs in large amounts

during the summer, thus lessening the amount of ground water recharge occurring during the winter and summer seasons (Reese & Lee, 1998). When there is a less significant amount of ground water recharge occurring, this means that there are fewer opportunities for nitrate to be transported to ground water supplies (Canter, 1997). Rainfall amounts are also an important factor when considering nitrate concentrations in ground water (Reese & Lee, 1998). Since nitrate travels readily with water, large amounts of precipitation percolating into the soil and ground water can cause elevated nitrate concentrations in ground water supplies (Canter, 1997). In addition, the ground does not always freeze significantly during the winter months in Pennsylvania, so precipitation is still capable of infiltrating the subsurface and impacting ground water quality during this time period (Canter, 1997).

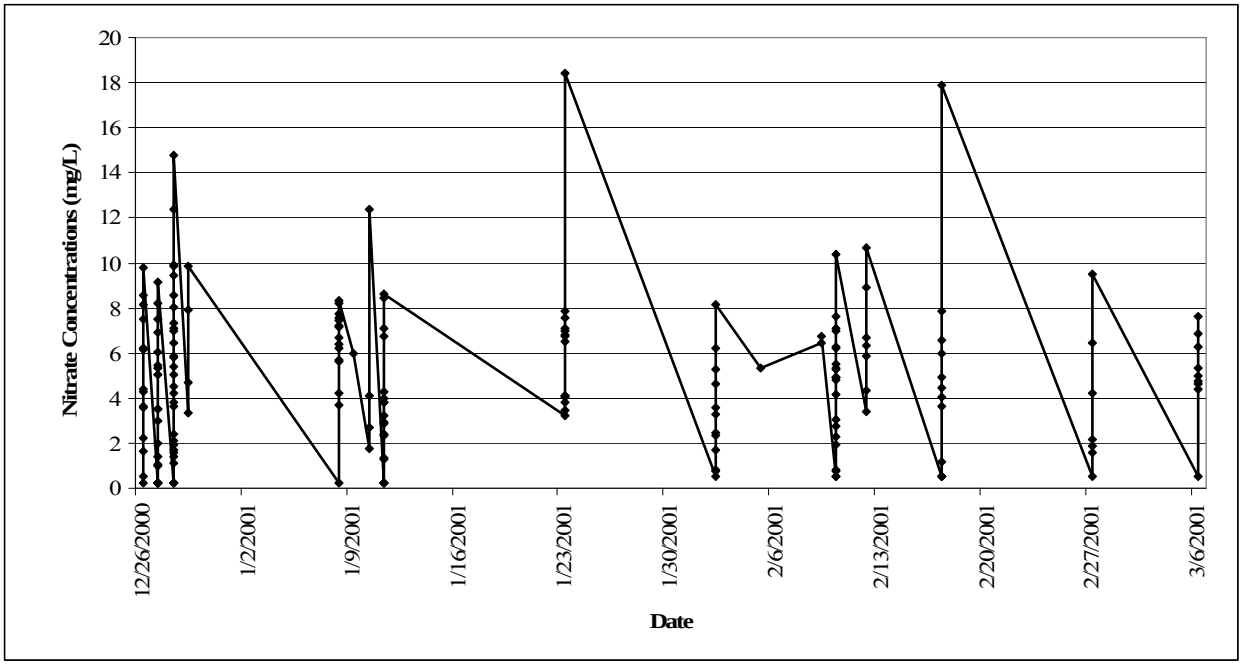
Since the nitrate concentration data for South Middleton Township were collected in the winter months from December through March, it can be assumed that nitrate concentrations were lower at this time of year than they typically would have been during the spring or summer months (Reese & Lee, 1998). In addition, drought conditions were not reported in Cumberland County during the sample period, although some areas received precipitation amounts that were slightly below average from December 2000 through March 2001 (Table 4) (PADEP, 2000a). By the end of March 2001, those areas experiencing slightly below average precipitation amounts once again had average rainfall accumulations or were experiencing amounts somewhat higher than normal (PADEP, 2001a; PADEP, 2001b). Since rainfall amounts deviated little from average in Cumberland County from December 2000 to March 2001, it can be assumed that ground water quality within the county was not substantially impacted by large amounts of precipitation percolating into the soil and water table.

Table 4. Average monthly precipitation and departure from normal for Cumberland County, Pennsylvania from December 2000 though March 2001 (PADEP, 2000b; PADEP, 2001b).

Date	Cumberland County Average Monthly Precipitation (cm)	Cumberland County Departure from Normal (cm)
December 2000	7.6	-0.3
January 2001	5.6	-1.3
February 2001	2.8	-3.8
March 2001	11.7	3.6

Variability in nitrate concentrations will also occur due to nitrogen-rich fertilizers and manures being applied to agricultural landscapes in the fall and spring. Since the nitrate concentration data for South Middleton Township were collected in the winter months, from December 2000 through March 2001, impacts from fertilizers and manures should not have caused increased nitrate concentrations during this time. The portion of the data collected in March is most susceptible to these impacts. According to the collected data, nitrate concentrations in well samples collected at the beginning of March seem to be on the rise, but these concentrations are not very different from concentrations in samples collected at the end of December or beginning of January (Figure 4.2). The sample with the highest nitrate concentration of 18.4 mg/L was collected in mid-January, while the twelve samples with the lowest nitrate concentration of 0.25 mg/L were collected from the end of December through the beginning of January.

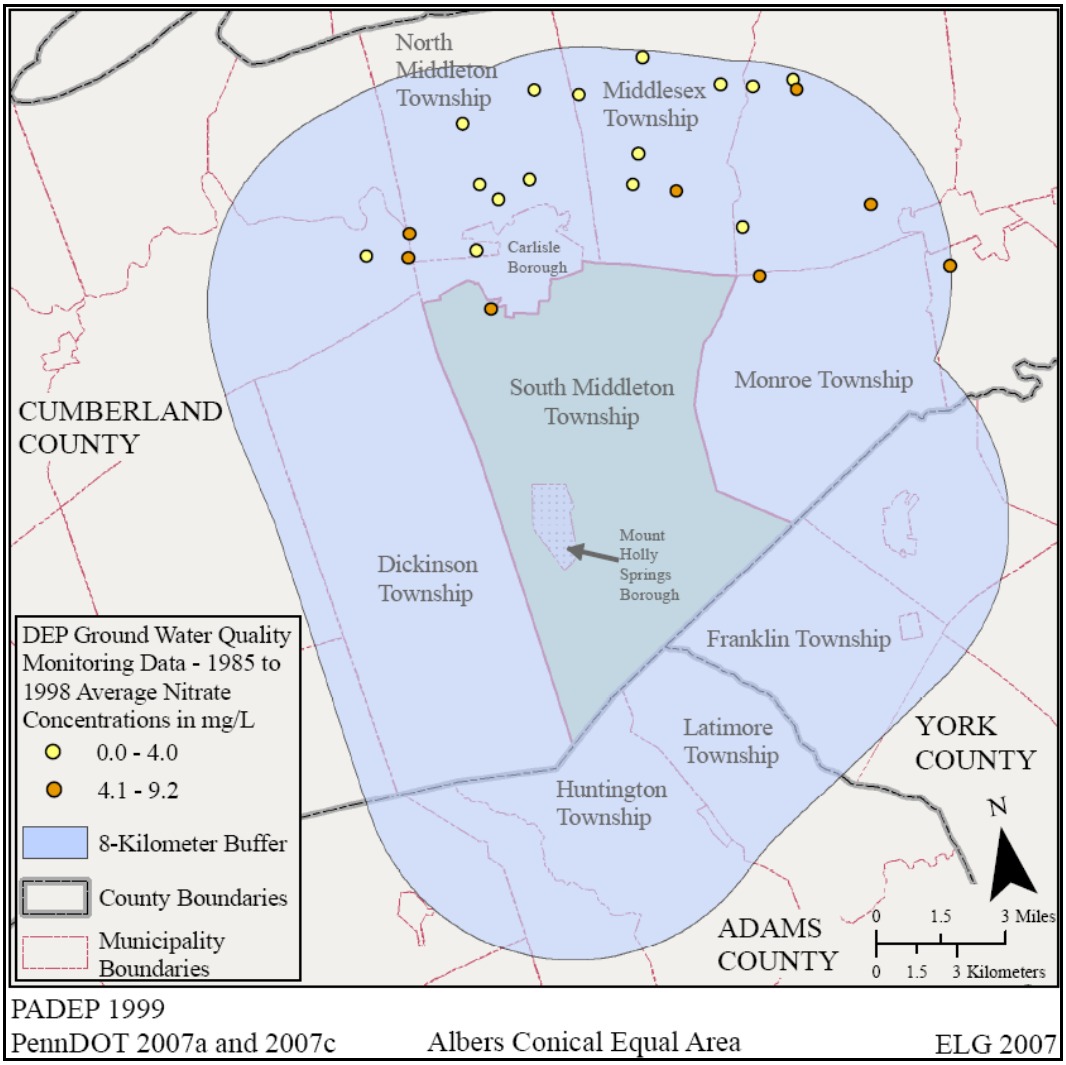
Figure 4.2. Nitrate concentrations for well samples collected in South Middleton Township from December 2000 through March 2001 (South Middleton Township, 2001).



4.1.1.2 Threshold

In order to analyze elevated nitrate concentrations in South Middleton Township, a nitrate concentration threshold exceeding the average local concentration was used. Data were obtained from the PADEP (1999) regarding average nitrate concentrations in ground water from 1985 to 1998. These data were collected as part of the state's Ambient and Fixed Station Network Monitoring Program in order to serve as a general observation of ground water quality (Reese & Lee, 1998). A total of 23 domestic wells located within 8 kilometers (5 miles) of the township were selected (Figure 4.3). Average nitrate concentrations from 1985 to 1998 associated with each well were assessed, and 8 of the wells had average nitrate concentrations exceeding 4 mg/L (PADEP, 1999). The average nitrate concentration among the wells was 3.1 mg/L (PADEP, 1999).

Figure 4.3. Wells extracted from Pennsylvania's Ambient and Fixed Station Network Monitoring Program that are located within an 8-kilometer buffer of South Middleton Township (PADEP, 1999; PennDOT, 2007a; PennDOT, 2007c).



It was decided to use a threshold of 4 mg/L nitrate for the study in order to indicate elevated nitrate concentrations in ground water caused by anthropic impacts since it exceeds the average local concentration according to the PADEP (1999) data. In addition, a 1996 study conducted by Ward *et al.* suggested that there is an increased risk for non-Hodgkin's lymphoma associated with long-term consumption of water containing nitrate concentrations greater than 4 mg/L. Furthermore, the value of 4 mg/L

is similar to the median value of 4.7 mg/L that is associated with the sample dataset used for the study.

4.1.1.3 Summary Statistics

Out of the 190 wells in the dependent variable dataset, there are 113 samples with nitrate concentrations greater than or equal to the 4 mg/L threshold, thus accounting for 59 % of the dataset (Table 5). Also, 4% of the samples in the dataset, representing 7 of the 190 wells, have nitrate concentrations greater than the MCL of 10 mg/L. A total of 12 wells within the dataset have nitrate concentrations of 0.3 mg/l, which is the minimum value within the dataset. A well with a nitrate concentration of 18.4 mg/L represents the dataset's maximum. The mean of the nitrate concentration dataset is 4.9 mg/L. In addition, the dataset's skewness value of 0.9 indicates positive skewness, and the kurtosis value of 2.1 shows stretching of the dataset's distribution since these values deviate from 0, which is a value indicating normality. Well water quality data were converted into a categorical variable by classifying all events or nitrate concentrations equal to and greater than 4 mg/L as ones and all nonevents or nitrate concentrations less than 4 mg/L as zeros (Appendix A).

Table 5. Summary statistics of nitrate concentrations in South Middleton Township (South Middleton Township, 2001).

Statistic	Value
Number of Samples	190
Minimum	0.3
Maximum	18.4
Median	4.7
Mean	4.9
Sample Variance	10.3
Standard Deviation	3.2
Kurtosis	2.1
Skewness	0.9
Standard Error	0.2
Number of Samples above 4 mg/L threshold	113
Percent of Samples above 4 mg/L threshold	59%
Number of Samples above 10 mg/L MCL	7
Percent of Samples above 10 mg/L MCL	4%

4.1.2 Independent Explanatory Variables

Explanatory variables utilized in the study include both anthropogenic and hydrogeologic data (Table 6). Anthropogenic data consist of land cover, total nitrogen inputs from atmospheric deposition and from farm and non-farm fertilizer and manure applications, onsite waste disposal, and population density. Hydrogeologic data include bedrock type, soil texture, soil hydrologic group, and surface depression and sinkhole densities. All of the data obtained represent the land surface as closely to the sample collection dates as possible, according to the years in which the available data were collected. These data were compiled for South Middleton Township and for some areas within surrounding municipalities.

Table 6. Explanatory variables utilized in the study (USGS, 2001; Ruddy *et al.*, 2006; Cumberland County Planning Commission, 2001; US Census Bureau, 2000a; US Census Bureau, 2000b; PGS, 2001; NRCS, 2004a; NRCS, 2004b; NRCS, 2004c; Kochanov, 1989; USGS, 1999a).

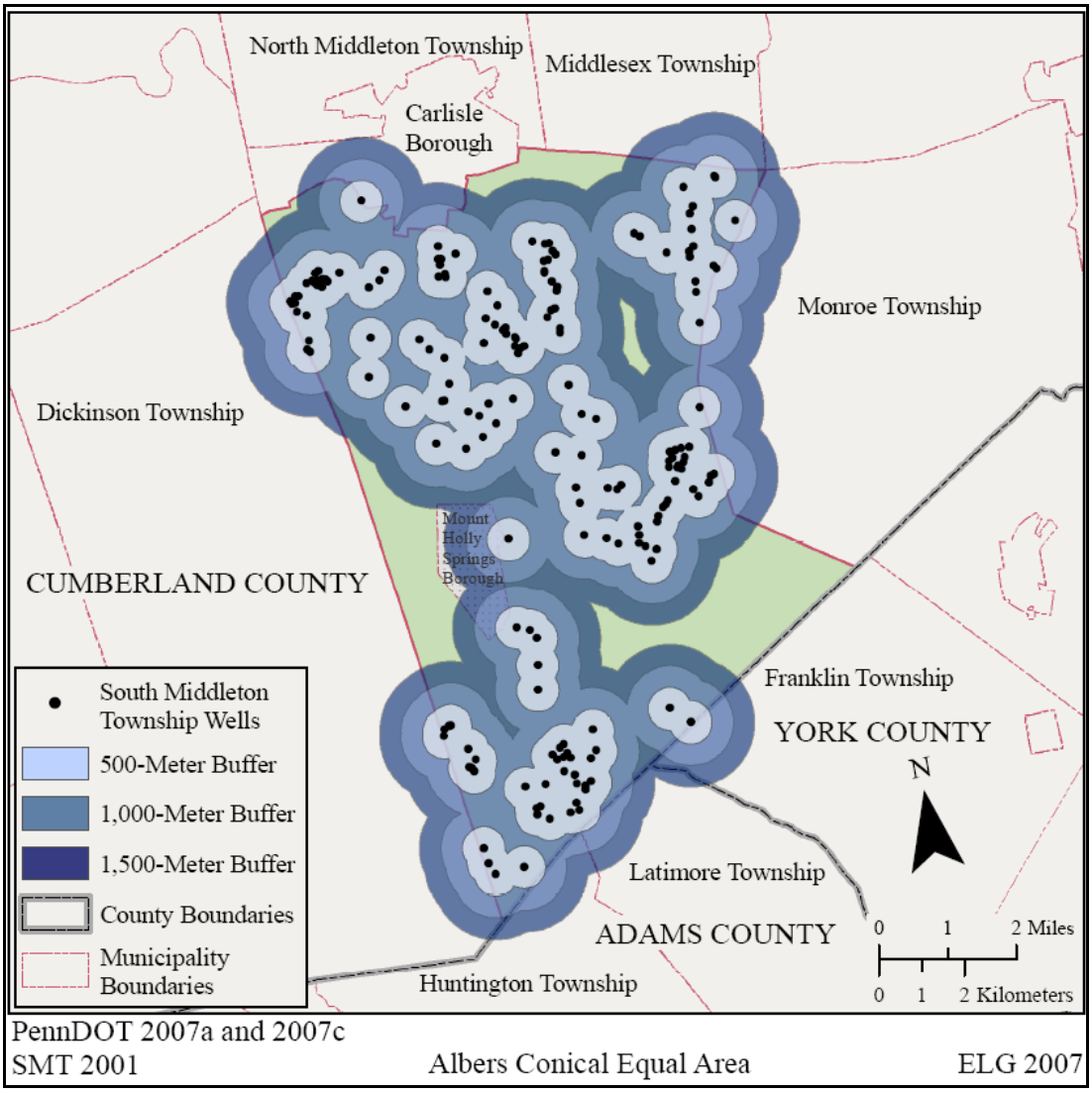
Variables		Level of Data	Source	Date
Anthropogenic Data	Land Cover	30-Meter Raster	US Geological Survey	2001
	Total Nitrogen Inputs from Atmospheric Deposition, Farm and Non-Farm Fertilizer Applications, and Manure Applications	30-Meter Raster	US Geological Survey; Ruddy *et. al*, 2006	2000 and 1997
	Onsite Waste Disposal	Polygon Land Parcel Data	Cumberland County Planning Commission	2001
	Population Density	Polygon Census Block Data	US Bureau of the Census	2000
Hydrogeologic Data	Bedrock Type	Polygon State Data	Pennsylvania Geological Survey	2001 (based off of the 1980 "Geologic Map of Pennsylvania")
	Soil Texture	Polygon County Data	US Department of Agriculture - Natural Resource Conservation Service	2004
	Soil Hydrologic Group	Polygon County Data	US Department of Agriculture - Natural Resource Conservation Service	2004
	Sinkhole and Surface Depression Densities	Point Data	Kochanov, 1989	Documented Since 1985

All explanatory datasets were compiled in shapefile or raster formats. All shapefiles were converted to raster datasets, and all raster datasets were created as 30-meter digital raster datasets in order to maintain consistency among data. All datasets were required to be in integer grid, or discrete raster, format before data extraction because attributes for an integer grid are stored in a value attribute table (VAT). Data cannot be extracted with an Arc Macro Language (AML) from a raster dataset without an associated VAT. Many of the datasets were floating-point grids, or continuous rasters, which do not have an associated VAT since the raster cells in a floating-point grid can have any value within a specific range of values. Therefore, the floating-point grids had to be converted to integer grids for data extraction purposes through the reclassification of variables.

Once each of the explanatory variables were represented as 30-meter integer grid raster datasets, all data within 500-meter, 1,000-meter, and 1,500-meter buffers of each well were extracted using various forms of an AML obtained from Hitt and Nolan (2005). The AMLs initiated an automated extraction process that obtained data according to 500-meter, 1,000-meter and 1,500-meter buffers surrounding the data points; therefore, three AMLs were run for each explanatory dataset since data was to be extracted for three different buffer sizes. The final output from the AMLs was converted to a table displaying a fraction value representing the portion of each unique variable from the explanatory datasets falling within a specific buffer of each well.

Although portions of the 500-meter, 1,000-meter, and 1,500-meter buffers extended into neighboring municipalities, the explanatory data were extracted in order to produce unique variables for each well describing the land area within each buffer for every well (Figure 4.4). Three different buffer sizes were chosen in order to determine which buffer size best fit the nitrate concentration dataset through logistic regression analysis. Once the data within the buffers were extracted, they were converted into tables, and a final dataset was compiled. The final dataset included fractional values for each type of variable representing the percentage of that variable type that fell within the buffer surrounding each well.

Figure 4.4. South Middleton Township wells with 500-meter, 1,000-meter, and 1,500-meter buffers (PennDOT, 2007a; PennDOT, 2007b; South Middleton Township, 2001).



Since ground water movement within South Middleton Township has not been documented in detail and is difficult to determine, the different buffer sizes were necessary in order to attempt to define the contributing area for each well in relation to its associated nitrate concentration. Although contributing areas will vary from well to well depending on environmental factors, the buffers utilized in the study were meant to be broadly associated with the recharge area within proximity of each well, but the buffers were by no means intended to precisely define well recharge areas. The processes used

to obtain and compile explanatory data before the data were extracted according to the three buffer sizes are described in detail in the following sections.

4.1.2.1 Anthropogenic Data

Anthropogenic datasets regarding South Middleton Township include variables that are a result of human impacts across the landscape. The data were obtained from various sources and were compiled as 30-meter digital raster datasets so that data regarding wells from the dependent dataset could be extracted. These data include land cover, nitrogen inputs from atmospheric deposition and from farm and non-farm fertilizer and manure applications, onsite waste disposal, and population density.

4.1.2.1.1 Land Cover

The 2001 land cover dataset for South Middleton Township was created by the USGS Multi-Resolution Land Characteristics Consortium (Figure 4.5). Land cover classifications within the township include agricultural, developed, forested, open water, and wetlands. The original data represented land cover for twelve different classifications located within South Middleton Township and were classified according to Level 2 data classifications (Figure 4.5). These different classifications included four different intensities of development, open water, deciduous, evergreen, and mixed forest types, pasture/hay and cultivated crops, and woody and emergent herbaceous wetlands (USGS, 2001). Land cover classifications were aggregated to Level 1 classifications in order to minimize the number of variables being used for statistical analysis and because it was similar to the aggregation of land use classifications used by Greene *et al.* (2005) in their study of the Mid-Atlantic region.

Figure 4.5. Level 1 land cover classifications for South Middleton Township (PennDOT, 2007c; USGS, 2001).

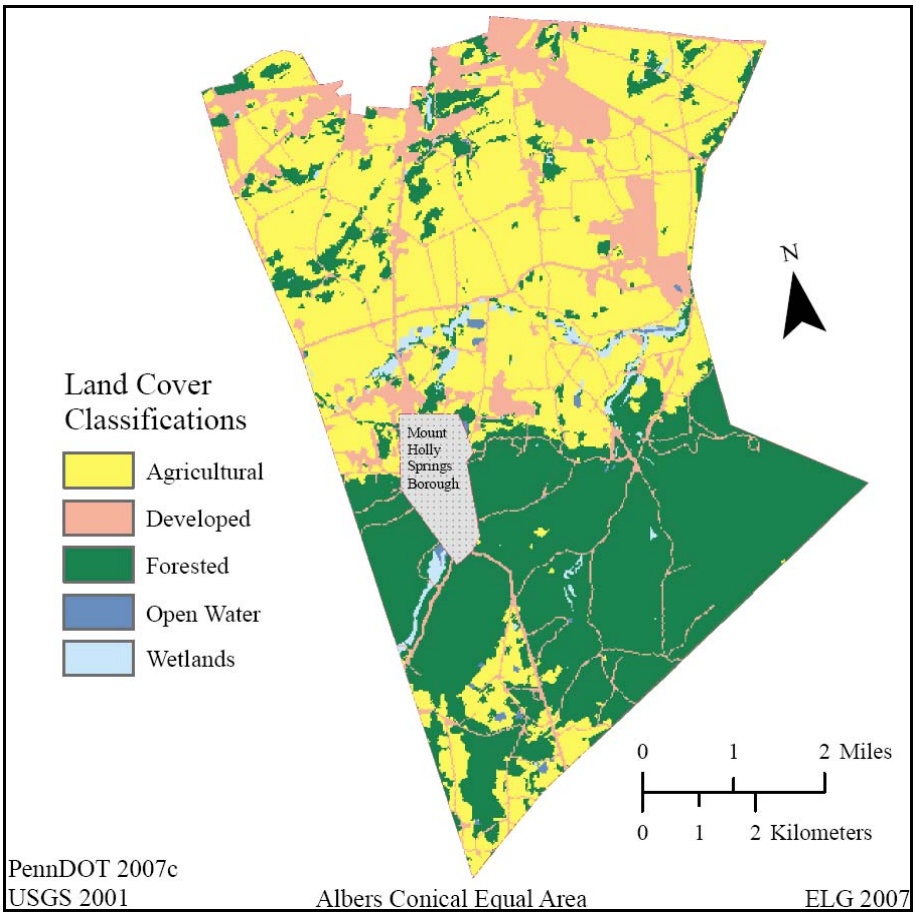
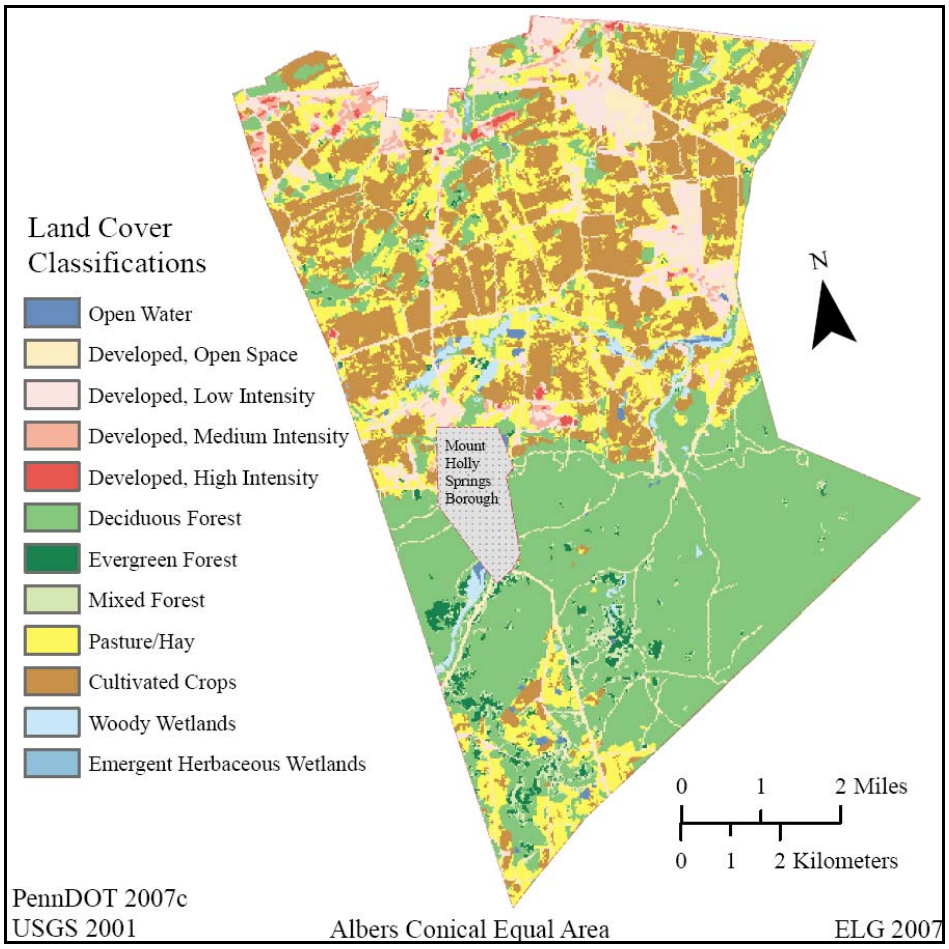


Figure 4.6. Level 2 land cover classifications for South Middleton Township (PennDOT, 2007c; USGS, 2001).

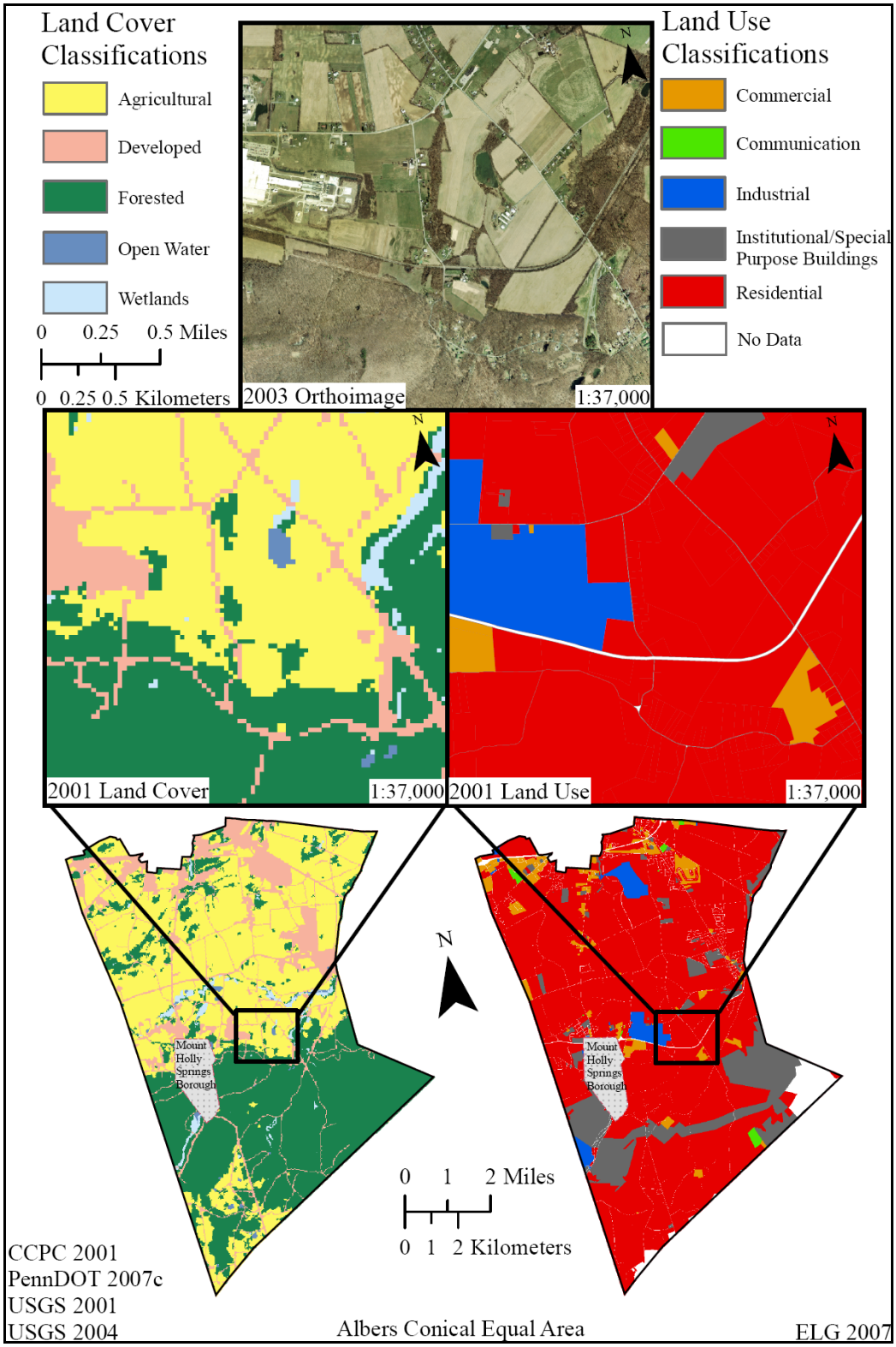


Therefore, instead of land cover data consisting of twelve different variables, data aggregation enabled land cover data to account for only five variables. All of the different intensities of development were combined to create the development variable. Pasture/hay and cultivated crops were merged to form an agricultural land cover variable. Mixed forest types and deciduous and evergreen forest types were aggregated to produce a forested variable. The open water classification remained unchanged. Also, the wetlands variable was aggregated from the woody wetlands and emergent herbaceous wetlands classifications. Once these data were extracted for the three different buffers, the final datasets consisted of the percentage of agricultural, developed, forested, open

water, and wetland land cover areas within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

For this study, 30-meter land cover raster data were utilized instead of polygon parcel data that were available for the township because they were more accurate (Figure 4.7) (USGS, 2001; Cumberland County Planning Commission, 2001). A 2001 polygon parcel dataset was obtained from the Cumberland County Planning Commission, and these data and the 2001 land cover data from the USGS were compared to 2003 high-resolution orthophoto images from the USGS (2004). Ultimately, the 2001 parcel land use data represent the landscape much differently than the 2001 raster land cover data (Figure 4.7). Parcels defined by 2001 land use encompass large areas of land, thus defining the entire area as residential, commercial, etc. when a large portion of the land owner's property may have been forested. In addition, large portions of Michaux State Forest in the central portion of South Middleton Township were defined in the 2001 land use dataset as residential, when the area was more than likely state forest land leased from the state government. Additionally, the 2001 land use dataset was not complete for the entire township since roads or large parcels in the southern portion of the township had no land use data associated with them. Therefore, it was determined that the 2001 land cover data would best describe the land area representing each well's recharge area.

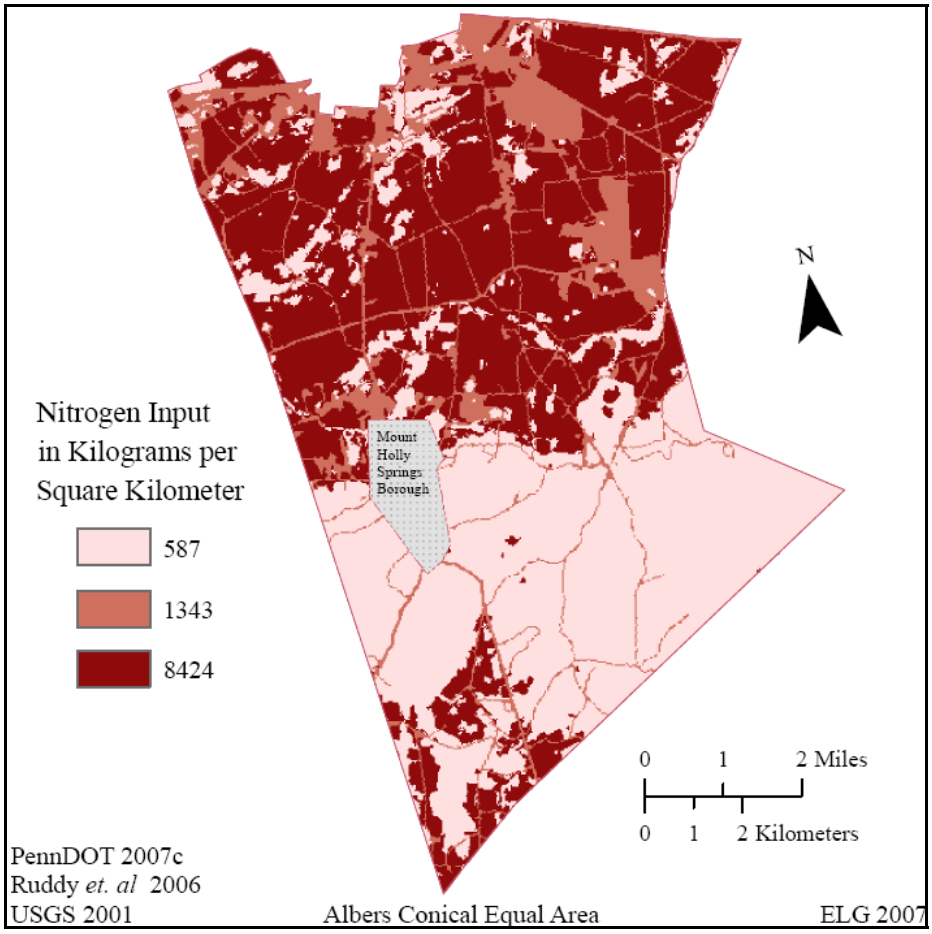
Figure 4.7. Comparison among land cover data, land use data, and a high resolution orthoimage (Cumberland County Planning Commission, 2001; PennDOT, 2007c; USGS, 2001; USGS, 2004).



4.1.2.1.2 Total Nitrogen Inputs

Data regarding nitrogen inputs from atmospheric deposition and non-farm and farm fertilizer and manure applications were also obtained for South Middleton Township and utilized to create a total nitrogen input dataset (Figure 4.8). Non-farm fertilizers typically consist of fertilizers applied to gardens or lawns in residential areas by property owners to provide nutrients for garden plants or to maintain a lawn's thickness and color (Ruddy *et al.*, 2006). On the other hand, farm fertilizers and manures are applied to agricultural fields in order to increase crop yields and provide nutrients for crops (Makuch and Ward, n.d.; Ruddy *et al.*, 2006).

Figure 4.8. Estimated total nitrogen input across the landscape from 2000 atmospheric deposition, 2000 farm and non-farm fertilizers applications, and 1997 manure applications in South Middleton Township (PennDOT, 2007c; Ruddy *et al.*, 2006; USGS, 2001).



County level values for nitrogen input from 2000 atmospheric deposition, 2000 non-farm fertilizer use, 2000 farm fertilizer use, and 1997 livestock manure in kilograms (kg) were obtained from a 2006 report completed by Ruddy *et al.* (2006). Applying county level data to a township is not as accurate as collecting data specifically regarding the township, but these were the data readily available. The data for livestock manure were the most recently available compiled by Ruddy *et al.* (2006) from the Census of Agriculture, and they reflect the total of unconfined and confined livestock. These nitrogen input values were then applied to the landscape by converting them from kilograms to kilograms per 30 m^2 and associating them with their proper land use classifications from the 2001 USGS 30-meter land cover raster dataset, thus utilizing methods similar to those used by Ruddy *et al.* (2006) in order to allocate nitrogen inputs across the landscape .

Using the methods associated with the 2006 study performed by Ruddy *et al.*, it was assumed that nitrogen inputs from atmospheric deposition occur evenly across the landscape, thus the nitrogen input value of .582 kg was applied to every 30-meter raster cell occurring within the township and surrounding municipalities in Cumberland County. Different values of .585 kg for Adams County and .528 kg for York County were applied to raster cells falling within municipalities in those specific counties, since nitrogen input values differed from county to county (Table 7). Next, the same process was completed for nitrogen inputs from farm fertilizer, non-farm fertilizer, and manure, but instead of applying these values across the entire landscape, they were only applied to their associated land cover classifications defined by Ruddy *et al.* (2006) (Figure 4.6). Nitrogen inputs for both farm fertilizer and manure were applied to all raster cells classified as the pasture/hay or cultivated crops land cover classifications since the 2001

USGS 30-meter land cover raster dataset does not always accurately discriminate between pasture/hay or cultivated crop land cover types across the landscape (Ruddy *et al.*, 2006). Likewise, nitrogen inputs for non-farm fertilizers were applied to all raster cells classified as the four different developed (open space, low intensity, medium intensity, high intensity) land cover classifications.

Table 7. Nitrogen input values for Adams, Cumberland, and York Counties (Ruddy *et al.*, 2006).

Counties	Nitrogen Input Values in kilograms per 30-meter raster cell (kg/0.001 km²)			
	Manure	Farm Fertilizer	Non-Farm Fertilizer	Atmospheric Deposition
Adams	2.581	3.552	0.570	0.585
Cumberland	4.100	2.899	0.626	0.582
York	2.452	2.979	1.175	0.528

The final total nitrogen input dataset was created by summing all nitrogen inputs from atmospheric deposition and non-farm and farm fertilizer and manure applications across the landscape. The total nitrogen inputs dataset reflects the total amount of nitrogen deposited and applied to the landscape through the averaging of the 2000 atmospheric deposition data, 2000 non-farm and farm fertilizer data, and 1997 manure data. Once these data were extracted for the three different buffers, the final datasets consisted of the total sum of kilograms of nitrogen per square meter applied to the landscape through atmospheric deposition, fertilizer and non-farm fertilizer applications, and manure applications within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

4.1.2.1.3 Onsite Waste Disposal

Although the 2001 land use parcel data obtained from the Cumberland County Planning Commission were not used as an explanatory variable for the study, the dataset

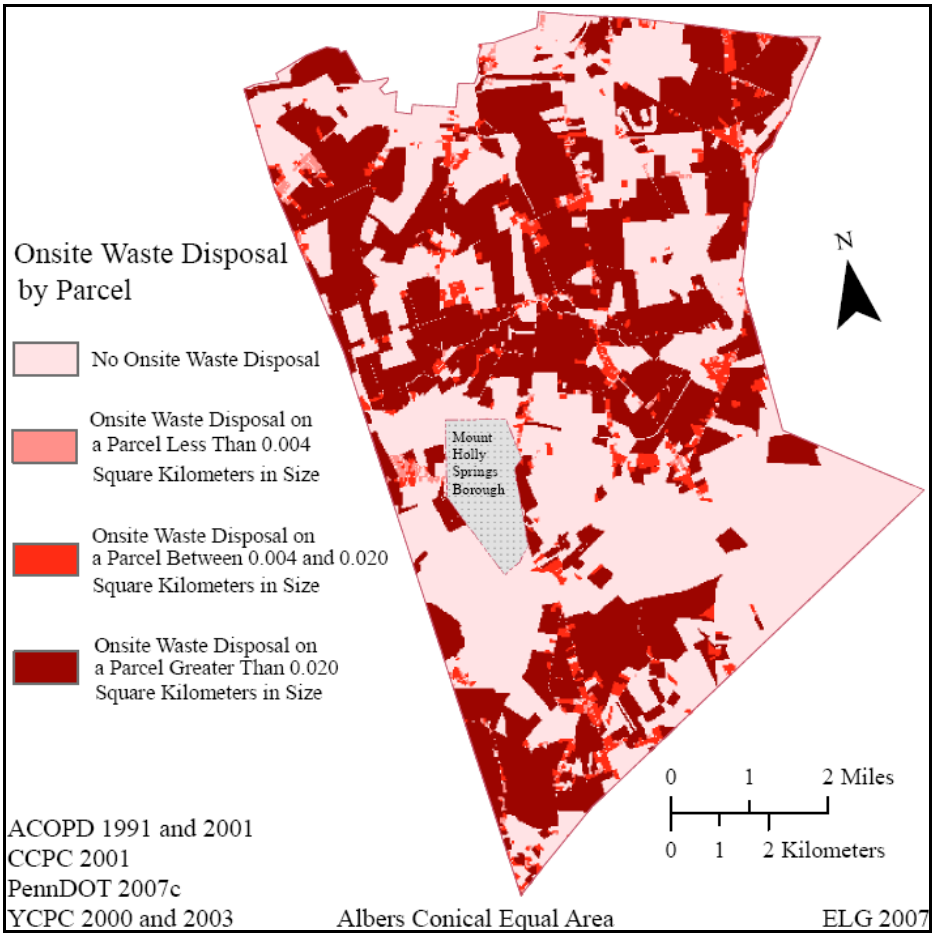
itself was still utilized since one attribute within the dataset indicated the type of waste disposal method that each parcel within the county used in 2001. Different waste disposal types included public sewer, sandmound, and septic. Parcels with either sandmound or septic waste disposal types were classified as parcels utilizing onsite waste disposal methods. Slightly different methods were used for those areas within buffers falling outside of Cumberland County.

For example, 2000 parcel data for York County were procured from the county's planning commission, but the dataset did not contain information regarding waste disposal methods. Therefore, York County's 2003 Water Management plan was obtained in order to retrieve information regarding waste disposal methods in northern York County. According to the plan, as of 2003 none of the areas of interest were within an existing community water system service area (York County Planning Commission, 2003). Consequently, those areas in buffers falling within residential parcels in northern York County were classified as using onsite waste disposal methods. Parcel data for Adams County could not be obtained, so residential parcels of interest in Adams County were digitized according to 2003 USGS high resolution orthophoto imagery. Next, Adams County's 1991 Comprehensive Plan and 2001 Water Supply and Wellhead Protection Plan were analyzed, and it was determined that those parcels of interest falling within Adams County also utilized onsite waste disposal methods (Adams County Office of Planning and Development).

Once the data for Adams, Cumberland, and York Counties were merged, those parcels utilizing onsite waste disposal methods were categorized according to size (Figure 4.9). Parcels using onsite waste disposal methods that were less than 0.004 km^2 (less than 1 acre) in size, between 0.004 and 0.020 km^2 (between 1 and 5 acres) in size,

and greater than 0.020 km² (greater than 5 acres) in size were divided categorically by those four different area groupings. The parcels were divided in this manner in order to determine if the presence of onsite waste disposal methods on parcels of various sizes had an impact on nitrate concentrations in ground water in 2001. Once these data were extracted for the three different buffers, the final datasets consisted of the percentage of land area not using onsite waste disposal methods and using onsite waste disposal methods on parcels less than 0.004 km², parcels between 0.004 and 0.020 km², and parcels greater than 0.020 km² within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

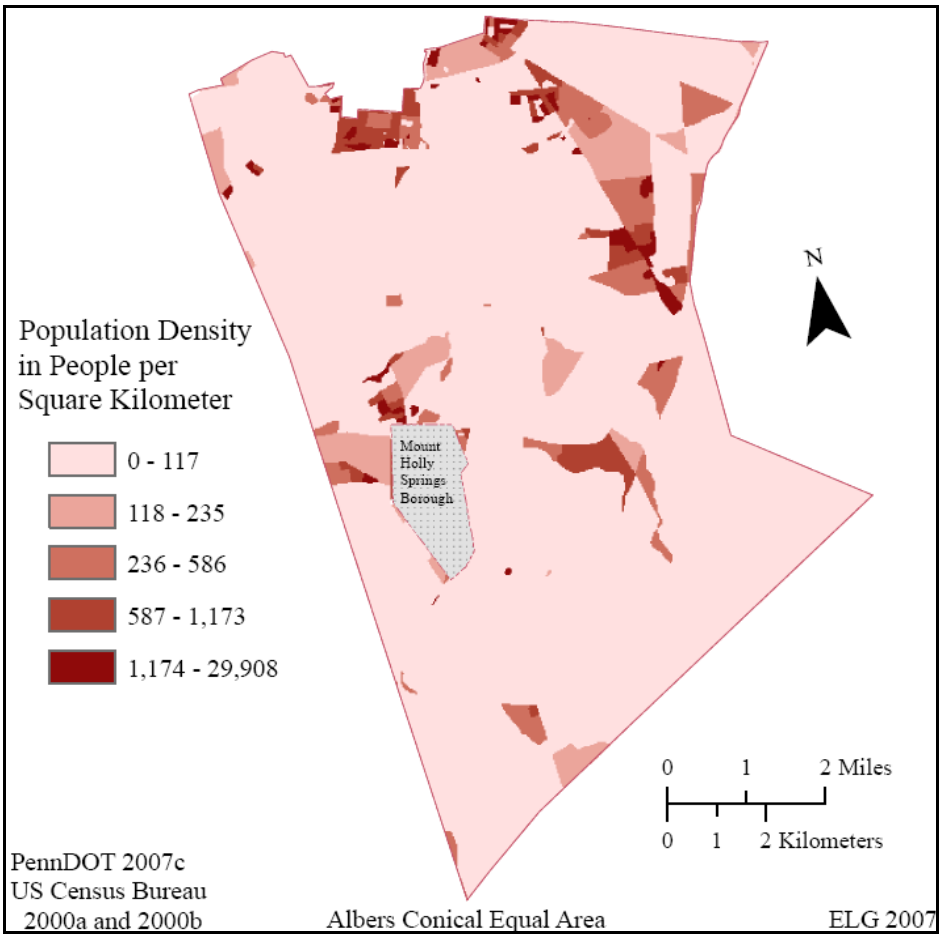
Figure 4.9. Land parcels of various sizes where onsite waste disposal methods were utilized in South Middleton Township in 2001 (ACOPD, 1991; ACOPD, 2001; Cumberland County Planning Commission, 2001; PennDOT, 2007c; YCPC, 2000; YCPC, 2003).



4.1.2.1.4 Population Density

Data regarding 2000 population in South Middleton Township and surrounding municipalities were obtained from the US Census Bureau for the census block level in order to create a population density dataset (Figure 4.10). The 2000 population data were related to 2000 census block spatial data that were also obtained from the US Census Bureau. Once these data were related, the area of each census block was calculated, and the total population for each block was divided its corresponding area, thus yielding population density data at the census block level. When obtaining 2000 population density at the census block level, it was assumed that the population was evenly distributed throughout the census block for the purposes of this study. The resulting floating-point grid with values ranging from 0 to 29,908 people per km^2 was converted to an integer grid by dividing the data categorically using a quantile classification so that each category contained an equal amount of features. A total of 25 classes were used to divide the data into different categories in order to minimize distortion, and each category was represented by its median value. Once these data were extracted for the three different buffers, the median values were averaged for each buffer in order to obtain an average population density for each buffer surrounding every well. Therefore, the final datasets consisted of the average population of the land area falling within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

Figure 4.10. Population density for 2000 by census block in South Middleton Township (PennDOT, 2007c; US Census Bureau, 2000a; US Census Bureau, 2000b).



4.1.2.2 Hydrogeologic Data

Hydrogeologic data regarding South Middleton Township included variables that are a result of natural phenomena occurring in the environment. The data were obtained from various sources and were compiled as 30-meter digital raster datasets so that data regarding wells from the dependent dataset could be extracted. These data included bedrock type, soil texture, soil hydrologic group, and surface depression and sinkhole densities.

4.1.2.2.1 Bedrock Type

A dataset regarding different bedrock types for South Middleton Township was produced to be utilized as an independent variable in the study (Figure 4.11). The dataset

was created according to the primary lithology attribute classification in a 2001 bedrock geology dataset obtained from the Pennsylvania Geological Survey (PGS). Primary lithology attributes were grouped in relation to several bedrock types, such as carbonate, crystalline, and siliciclastic (Table 8). These groupings were performed based on geologic groups that are typically utilized in studies completed by the USGS. For USGS studies, complex geologic formations are typically grouped according to major physiographic provinces and generalized rock types in order to identify general areas in which the chemical composition of ground water is expected to differ (Risser & Siwiec, 1996). This type of geologic grouping is important for this study because certain generalized bedrock types, such as carbonate bedrock types, are generally more susceptible to elevated nitrate concentrations than others. Once these data were extracted for the three different buffers, the final datasets consisted of the percentage of carbonate, crystalline, and siliciclastic bedrock types within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

Figure 4.11. Bedrock types in South Middleton Township (PennDOT, 2007c; PGS, 2001)

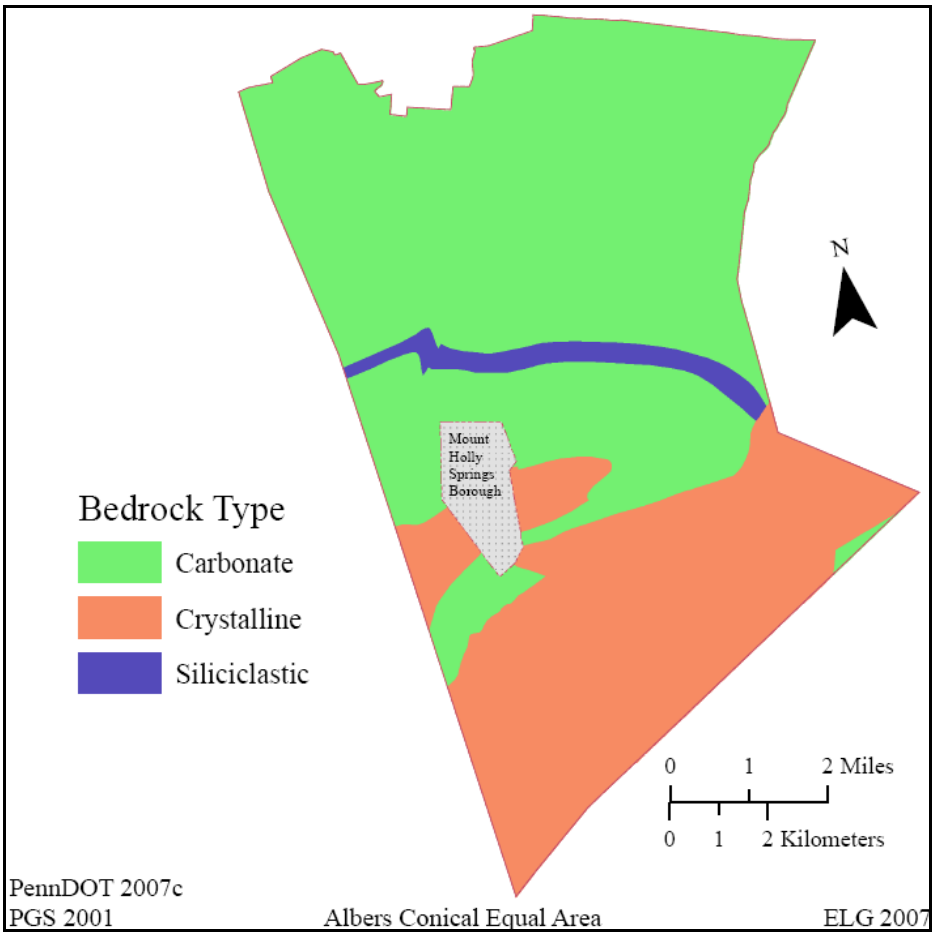


Table 8. Grouping of the primary lithology attribute by bedrock type in order to create the bedrock type dataset (PGS, 2001).

Bedrock Type	Primary Lithology Attribute Classifications
Carbonate	Argillaceous limestone
	Calcareous shale
	Dolomite
	Limestone
	Limestone conglomerate
Crystalline	Diabase
	Greenstone schist
	Metabasalt
	Metarhyolite
	Quartz conglomerate
	Quartzite
Siliciclastic	Shale
	Silty mudstone

4.1.2.2.2 Soil Texture

Furthermore, datasets describing soil texture within South Middleton Township were created. Data were obtained from the US Department of Agriculture's Natural Resources Conservation Service (NRCS) (2004a; 2004b; 2004c). The soils data are the most detailed level of soil geographic data developed by the National Cooperative Soil Survey (NRCS, 2004a; NRCS, 2004b; NRCS, 2004c). Three of the datasets depict the percentages of sand, silt, or clay in soils within the study area (Figures 4.12, 4.13, and 4.14). Once these data were extracted for the three different buffers, the final datasets consisted of the percentages of sand, silt, and clay in soils located within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

Figure 4.12. Percentage of sand in soils in South Middleton Township (NRCS, 2004a; NRCS, 2004b; NRCS, 2004c; PennDOT, 2007c).

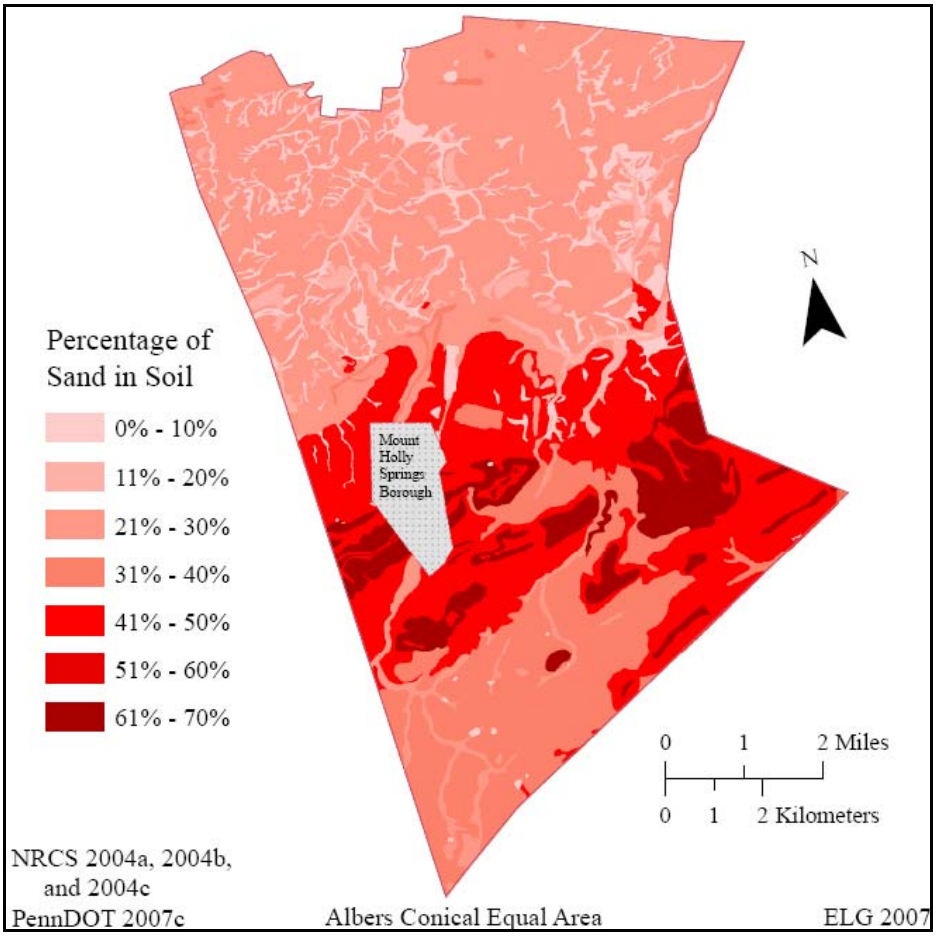


Figure 4.13. Percentage of silt in soils in South Middleton Township (NRCS, 2004a; NRCS, 2004b; NRCS, 2004c; PennDOT, 2007c).

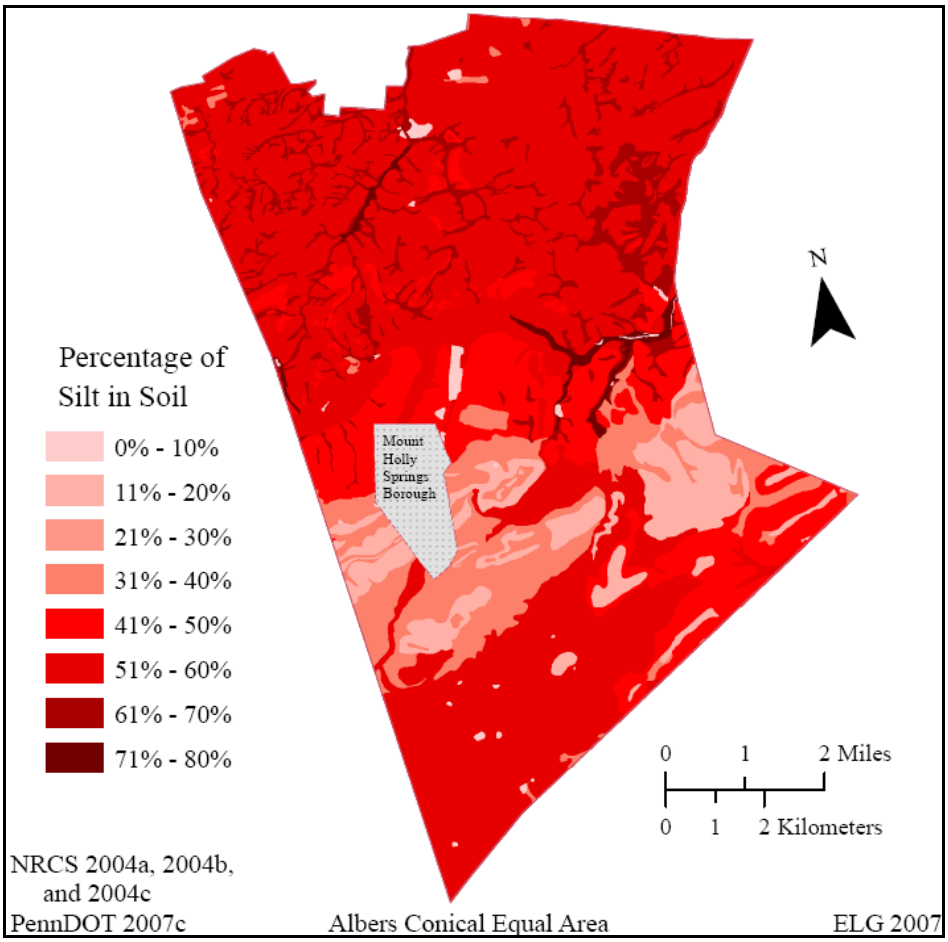
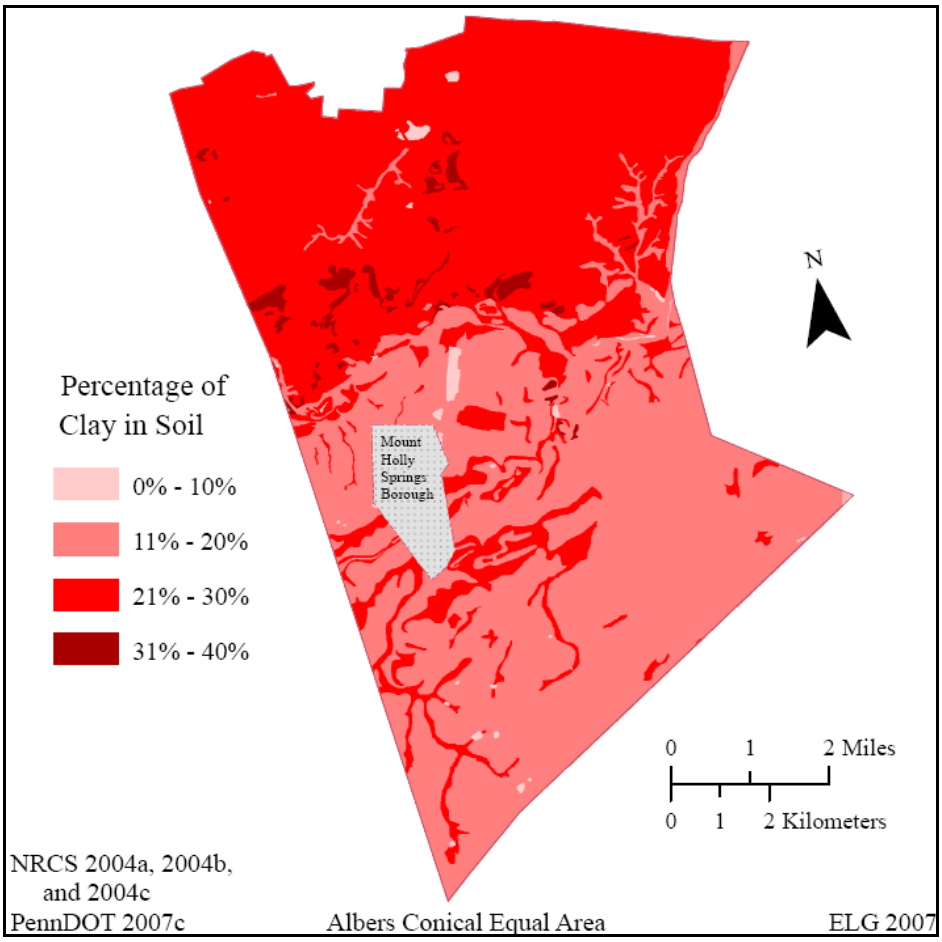


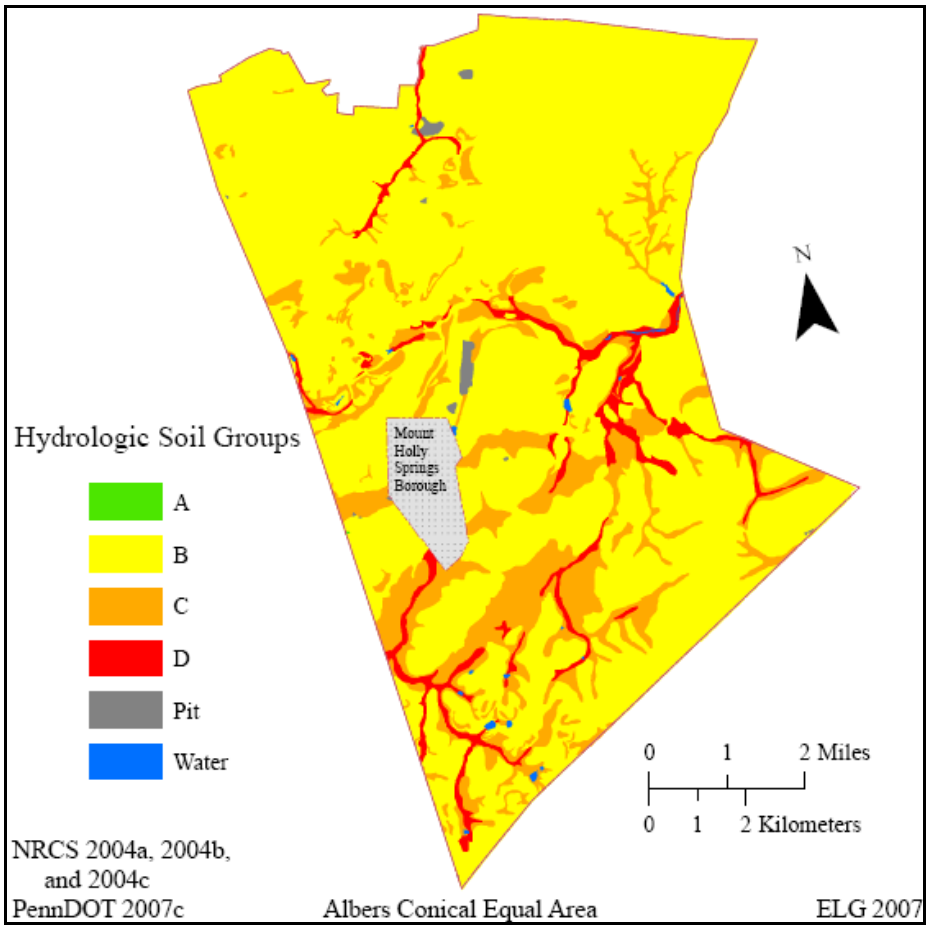
Figure 4.14. Percentage of clay in soils in South Middleton Township (NRCS, 2004a; NRCS, 2004b; NRCS, 2004c; PennDOT, 2007c).



4.1.2.2.3 Soil Hydrologic Group

Data regarding soil hydrologic groups were obtained from the US Department of Agriculture's Natural Resources Conservation Service (NRCS) (2004a; 2004b; 2004c). Just like the soil texture data, this data are the most detailed level of soil geographic data developed by the National Cooperative Soil Survey (NRCS, 2004a; NRCS, 2004b; NRCS, 2004c). The dataset depicts each area's hydrologic soil group classification, which are grouped as A, B, C, or D (Figure 4.15).

Figure 4.15. Hydrologic soil groups A, B, C, and D in South Middleton Township (NRCS, 2004a; NRCS, 2004b; NRCS, 2004c; PennDOT, 2007c).



Hydrologic soil group classifications are based on the runoff potential of each soil type, so the A grouping has the smallest runoff potential with sandy soils and the D grouping has the greatest runoff potential with clay soils (NRCS, 1986). Groupings B and C have moderate runoff potentials, with the B grouping having silt loam soils and the C grouping having sandy clay loam soils (NRCS, 1986). Other groupings included pit and water, and these were combined into one category labeled "Other" for data extraction. Once these data were extracted for the three different buffers, the final datasets consisted of the percentage of soils classified as hydrologic soil groups A, B, C, D, or "Other" within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

4.1.2.2.4 Sinkhole and Surface Depression Densities

In addition to the other hydrogeologic datasets obtained for the study area, sinkhole and surface depression density datasets were also created for analysis (Figures 4.16 and 4.17). Point data for sinkholes and surface depressions were obtained from the PGS (Kochanov, 1989). Next, a GIS was utilized in order to create a continuous density surface for each dataset, which enabled the interpretation of the point values in relation to their distribution over the study area. This method provided datasets depicting areas where sinkholes and surface depressions are most densely located within South Middleton Township. The resulting floating-point grids with values ranging from 0 to 14 sinkholes per km^2 and 0 to 60 surface depressions per km^2 were converted to integer grids by dividing the data categorically using a quantile classification so that each category contained an equal amount of features. A total of 25 classes were used to divide the data into different categories in order to minimize distortion, and each category was represented by its median value. Once these data were extracted for the three different buffers, the median values associated with each buffer around each well were averaged in order to obtain average sinkhole and surface depression densities surrounding every well. The final datasets consisted of the average densities of sinkholes and surface depressions within 500-meter, 1,000-meter, and 1,500-meters buffers of each of the 190 wells.

Figure 4.16. Sinkhole density in South Middleton Township (Kochanov, 1989; PennDOT, 2007c).

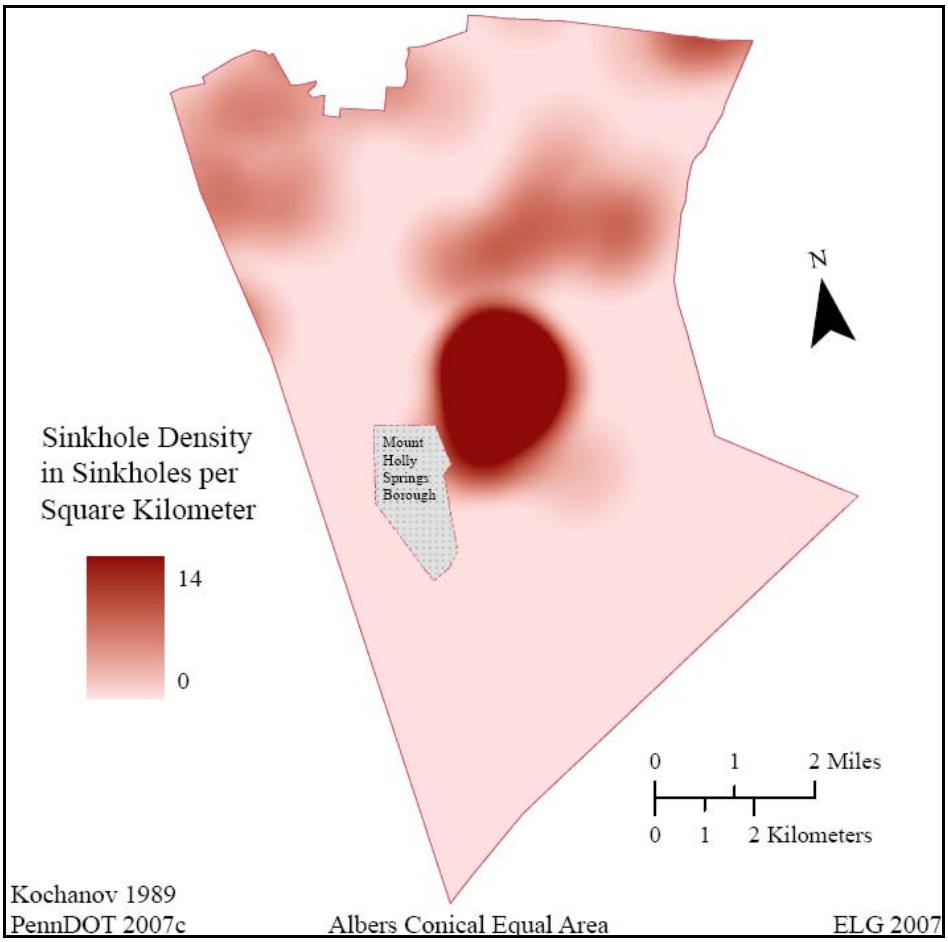
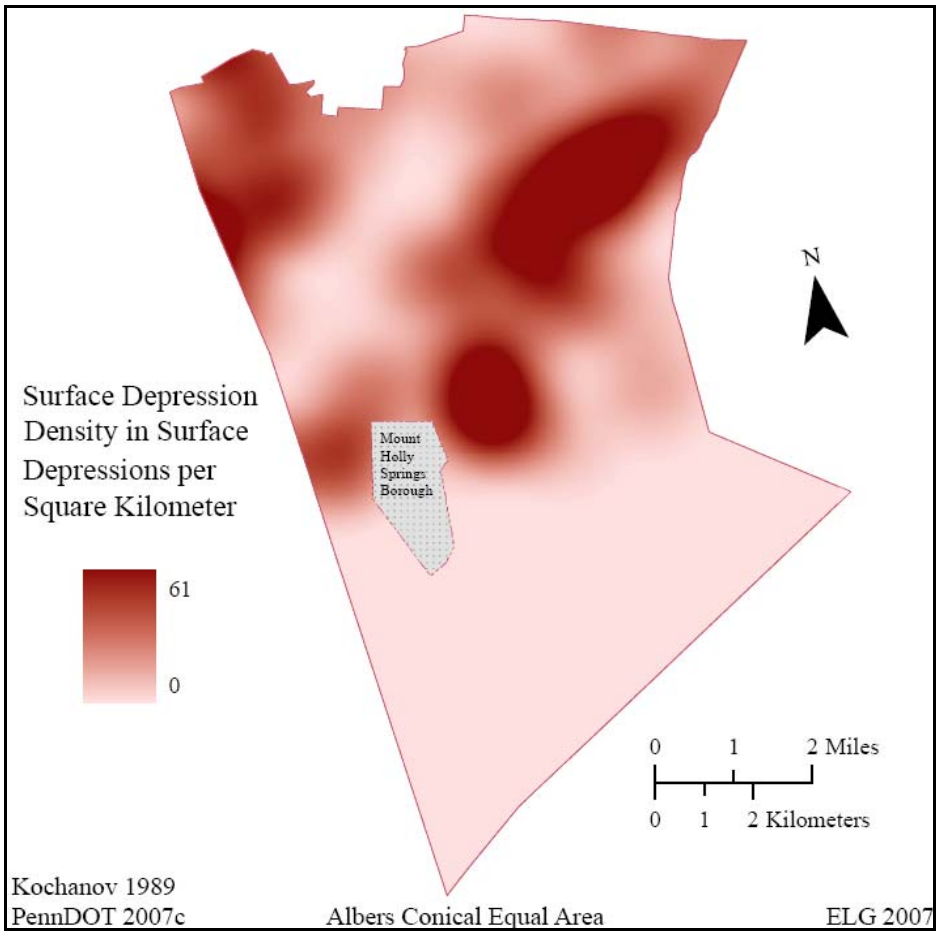


Figure 4.17. Surface depression density in South Middleton Township (Kochanov, 1989; PennDOT, 2007c).



4.2 Statistical Analysis

Logistic regression analysis was used to determine which of the explanatory variables were most responsible for nitrate concentrations exceeding 4 mg/L in ground water within South Middleton Township. Once the explanatory data were compiled, univariate analysis was performed on each dataset in order to determine which variables were statistically significant and would be utilized for the logistic regression models. All of the data were checked for non-normality, and Spearman's rank correlation coefficient measure was calculated to determine the relationship between the explanatory variables and dependent variable. Next, logistic regression analysis was performed, and test statistics were used to determine overall statistical significance of the final model.

4.2.1 Univariate Analysis

After explanatory data were compiled regarding the three different buffer sizes for each well, univariate statistical analysis was performed. Univariate analysis consisted of analyzing the relationship between the dependent variable and the independent variables. The dependent and independent variables were tested for normality utilizing the Shapiro-Wilk test, which is a nonparametric analysis. The null hypothesis for the Shapiro-Wilk test is that the data are normally distributed. Therefore, a p-value less than 0.05 indicates that the test is statistically significant at the $\alpha = 0.05$ level of significance and that the null hypothesis should be rejected. When the null hypothesis is rejected, this means that the data are not normally distributed. In addition, kurtosis values were examined for the datasets in order to determine non-normality. Kurtosis values that deviate from zero indicate datasets that are not assumed to be normally distributed (SAS Institute Inc., 1989).

Next, Spearman's rank correlation coefficient measure was utilized in order to determine the relationship between the dependent variable and each of the independent variables (Ott, 1993). For Spearman's rank correlation coefficient, the squared difference is computed from ranks that are calculated separately for each variable and averaged for tied observations (Ott, 1993). The resulting correlation coefficient shows the strength of the relationship between the two input variables assuming that the resulting coefficient was statistically significant at the $\alpha = 0.05$ level of significance, thus indicated by a p-value less than 0.05 (Lindsey *et al.*, 2006). The Spearman's rank correlation coefficient for each variable determined which variables would be utilized for the stepwise logistic regression model for each of the buffers. Additional statistics were utilized for the final

model to address multicollinearity diagnostics and how well the final model fit the data associated with the dependent variable.

4.2.2 Logistic Regression Analysis

Logistic regression was utilized for exploratory data analysis because it predicts the probability of a binary or categorical response, which in this case would be the exceedance of the 4 mg/L threshold value (Helsel & Hirsch, 1992). Therefore, based on the threshold value of 4 mg/L, a dataset containing nitrate concentrations in mg/L has all nitrate values below 4 mg/L reclassified as zeros to represent nonevents, while all concentrations equal to or exceeding 4 mg/L are reclassified as ones to represent events. Although this process simplifies the dataset, the dichotomous reclassification of nitrate concentrations according to a specific threshold value to create a variable in binary format makes this modeling approach a useful decision-making tool for officials in charge of water and land management (Greene *et al.*, 2005).

Linear regression cannot be used when the dependent variable is dichotomous, so other multivariate statistical methods must be employed (Green *et al.*, 2005). In addition, disparity among nitrate concentrations in a study can be such that predicting the probability of elevated nitrate concentrations through logistic regression will be more feasible than predicting a specific concentration with multiple linear regression (Nolan, 2001). While a logistic regression model assumes that the log-odds of the model are related to the explanatory variables, linear regression models the actual response with the explanatory variables (Greene *et al.*, 2005). The primary difference between multiple linear regression and logistic regression is that logistic regression models a change in the log-odds of the dependent variables rather than a change in the variable itself, such as with multiple linear regression (Greene *et al.*, 2005). Therefore, logistic regression better

suits the purpose of this study, which is to determine which explanatory variables are best at predicting the probability of nitrate concentrations occurring above 4 mg/L. In addition, since decision-makers are capable of drawing more conclusions from a predicted probability than from a predicted value, such as with multiple linear regression, logistic regression is better suited to increase knowledge and awareness regarding ground water issues (Focazio *et al.*, 2002). Elements regarding the risk and uncertainty issues that are associated with elevated nitrate concentrations in ground water are better interpreted by decision-makers through predicted probability maps that display the possibility of an occurrence (Focazio *et al.*, 2002).

Logistic regression analysis predicts the probability of a binary or categorical response based on explanatory variables by transforming estimated probabilities into a continuous response variable (Helsel & Hirsch, 1992; Allison, 1999). Ultimately, the odds ratio (Equation 1) is characterized as the probability of exceeding a threshold value:

$$\text{Odds ratio} = \frac{p}{1 - p} \qquad (1)$$

where p is the probability of an event and

$1 - p$ is the probability of a nonevent (Allison, 1999; Helsel & Hirsch, 1992; Gurdak & Qi, 2006).

Next, the log of the odds ratio, or logit, transforms a variable constrained between zero and one into a continuous variable that is a linear function of one or more of the explanatory variables in order to produce the logistic regression equation (Equation 2):

$$\ln\left(\frac{p}{1-p}\right) = b_o + bx \qquad (2)$$

where b_o is a logistic regression constant

bx is a vector of explanatory variables and slope coefficients (Helsel & Hirsch, 1992; Allison, 1999; Gurdak & Qi, 2006).

Subsequently, the logistic transformation (Equation 3) converts the predicted values of the response variable back into probability units:

$$P = \frac{e^{(b_o + bx)}}{1 + e^{(b_o + bx)}} \qquad (3)$$

where P is the probability of the binary response event, which is defined in this study as nitrate concentrations within ground water being equal to or exceeding the 4 mg/L threshold and

e is the base of natural logarithm (Helsel & Hirsch, 1992; Allison, 1999; Gurdak & Qi, 2006).

Therefore, the logistic regression equation with multiple explanatory variables (Equation 4) takes on the form of:

$$p = \frac{e^{(b_0 + b_1 x_1 + b_2 x_2 + \ldots b_i x_i)}}{\left[1 + \left(e^{(b_0 + b_1 x_1 + b_2 x_2 + \ldots b_i x_i)}\right)\right]} \quad (4)$$

where b_0 is the constant,

x_1 is the first explanatory variable,

b_1 is the slope coefficient of x_1,

x_2 is the second explanatory variable,

b_2 is the slope coefficient of x_2,

x_i is explanatory variable i, and

b_i is the slope coefficient of x_i (Helsel & Hirsch, 1992; Allison, 1999; Lindsey *et al.*, 2006).

When forming the logistic regression model, stepwise logistic regression was employed in order to analyze data for 500-meter, 1,000-meter, and 1,500-meter buffers surrounding each well. Stepwise logistic regression uses a statistical algorithm to add or remove variables based on each variable's statistical significance and employs methods associated with both forward selection and backward elimination techniques (Menard, 2002; Greene *et al.*, 2005). Stepwise logistic regression starts with the forward selection process (Menard, 2002; Greene *et al.*, 2005). Variables are added to the model, and if the associated variable is statistically significant at the $\alpha = 0.2$ level of significance, it is used in the model (Menard, 2002; Greene *et al.*, 2005). Next, backward elimination steps are employed, and any variables not statistically significant at the $\alpha = 0.05$ level of significance are removed from the model (Menard, 2002; Greene *et al.*, 2005). This procedure using both forward and backward selection processes continues until no more

variables can offer a change in the log-odds, which indicates that no more variables can be added to the model or removed from it (Menard, 2002; Greene *et al.*, 2005).

Results of the logistic regression for the three buffer sizes were analyzed using multicollinearity diagnostic statistics, such as the Tolerance and Variance Inflation Factor, to check for multicollinearity issues among variables. After multicollinearity diagnostics were analyzed, a final model was chosen based on the overall significance of the model, Hosmer-Lemeshow goodness-of-fit test statistic, maximum rescaled r-square values, and percent concordance. In addition, the Pearson residual statistic was employed to evaluate how well the final model fit the dependent data.

Multicollinearity diagnostics were examined in order to make sure that there was not a strong correlation among any of the explanatory variables included in the final models associated with the three buffer sizes. It is important to check for multicollinearity among variables because multicollinearity can inflate the variance of the parameter estimates, thus producing a lack of statistical significance even though the model is strongly significant (Greene *et al.*, 2005; Allison, 1999). Multicollinearity was examined using the Tolerance and Variance Inflation Factor, which are two statistics based on linear regression analysis of explanatory variables. The Tolerance is $1 - r$, where r is the coefficient of determination for the regression of one independent variable on all remaining independent variables (Allison, 1999). A tolerance value less than 0.4 is a good indicator of multicollinearity among variables (Allison, 1999). The Variance Inflation Factor is the reciprocal of the Tolerance and illustrates the inflation of the variance of coefficient compared to what it would be if there was no multicollinearity detected (Allison, 1999). A Variance Inflation Factor greater than 2.5 is an indicator of multicollinearity (Allison, 1999).

Model significance, the Hosmer-Lemeshow goodness-of-fit test statistic, maximum rescaled r-square values, and percent concordance were used to analyze logistic regression model results. A model's statistical significance is indicated by the p-value of its Wald Chi-Square statistic (Allison, 1999). If a p-value is below 0.05, then the model is statistically significant at the $\alpha = 0.05$ level of significance. Likewise, if a model's p-value is above 0.05, then the model is not statistically significant at the $\alpha = 0.05$ level of significance. If a model's p-value indicates statistical significance, then this shows that an explanatory variable improves the model's ability to predict the probability of an event occurring. The Hosmer-Lemeshow goodness-of-fit test statistic evaluates model calibration by addressing how much the outcomes from the predicted model vary from the outcomes associated with the original data (Hosmer & Lemeshow, 1989). For this test statistic, data are sorted and grouped into ten deciles of risk, and within these deciles, expected frequencies are determined then compared with the observed frequencies (Hosmer & Lemeshow, 1989). If the resulting p-values are greater than 0.05, this indicates that the model's estimates fit the original data at an acceptable level, thus a higher p-value indicates a well-calibrated model (Hosmer & Lemeshow, 1989).

Since there is no r-square value exactly like the r-square value typically utilized in linear regression, the generalized r-square and maximum rescaled r-square values are commonly used in its place in logistic regression (Allison, 1999; Lindsey *et al.*, 2006). The generalized r-square measures the predictive power of the model, and it is based on maximizing the likelihood ratio chi-square for testing the null hypothesis that all coefficients are zero (Allison, 1999). In addition, the maximum rescaled r-square value divides the generalized r-square value by its upper bound in order to account for discrete dependent variables (Allison, 1999). These values are best utilized as a comparison from

one logistic regression model to the next rather than as the percentage of variance explained by the model (Allison, 1999; Lindsey *et al.*, 2006). In addition, percent concordance is calculated by comparing every possible combination of data points with different observed responses (Lindsey *et al.*, 2006). If the lower ordered response value has a lower predicted mean score, then that pair is concordant (Lindsey *et al.*, 2006). Likewise, if the lower ordered response value has a higher predicted mean score, then that is discordant. A model with higher percent concordance will be a model with a better prediction (Lindsey *et al.*, 2006).

Subsequently, the Pearson residual statistic for the final model was calculated. The Pearson residual statistic evaluates the difference between the observed and estimated probabilities and then divides this difference by the standard deviation of the estimated probability (Menard, 2002). For this study, residual values closer to zero indicate that the probability of nitrate concentrations exceeding the 4 mg/L threshold at a specific well is what would be expected (Menard, 2002). Therefore, positive residual values indicate that the probability is greater than what would be expected, while negative residual values indicate that the probability is less than what would be expected based on the original data (Menard, 2002). Typically, Pearson residuals greater than 2 or less than -2 indicate areas where the model does not do a good job predicting the event (Menard, 2002; Gurdak & Qi, 2006). Pearson residuals associated with the final model were mapped, and individual wells were evaluated in order to determine why some areas of the model did not do a good job predicting elevated nitrate concentrations.

Although some of the previously discussed studies produced predictive maps showing the predicted probability of elevated nitrate concentrations, predictive maps were not presented for this study due to the predictive power of the results associated

with the final model. In addition, validation of the final model was not performed because of the lack of a validation dataset. The dependent dataset could have been divided into a calibration dataset, which would have been made up of 85 percent of the data, and a validation dataset, which would have included 15 percent of the data (Lindsey *et al.*, 2006). Subsequently, the dependent dataset would have consisted of 162 well samples, while the validation dataset would have consisted of 28 well samples. A validation dataset of 28 samples would not have been sufficient enough for validation, and it was not feasible to further lessen the number of wells used for model calibration. Therefore, the dependent dataset was not large enough for the extraction of a validation dataset.

Chapter 5

Results

5.1 Statistical Analysis

Methods associated with the project include univariate analysis of data and development and analysis of a final logistic regression model. Univariate analysis of the data included testing for normality and determining the relationship between the dependent variable and each of the explanatory variables. Next, logistic regression models were developed utilizing the stepwise logistic regression procedure. Multicollinearity diagnostics were performed for the final model associated with each buffer size. Next, different aspects of each model such as, overall model significance, the Hosmer-Lemeshow goodness-of-fit test statistic, maximum rescaled r-square values, and percent concordance, were analyzed in order to determine a final model. Finally, the Pearson residual statistic was calculated to establish how well the model fit the dependent dataset.

5.1.1 Univariate Analysis

Univariate analysis included testing for normality and analyzing the relationship between the dependent variable and each explanatory variable. The nonparametric Shapiro-Wilk analysis yielded p-values less than 0.05 for the dependent dataset and all of the independent variables for all buffer sizes. A p-value less than 0.05 indicates statistical significance at the $\alpha = 0.05$ level of significance and also indicates that none of the data were normally distributed, which is a common occurrence in environmental data (Shumway *et al.*, 1989). In addition, kurtosis values were analyzed, and all of the values deviated from zero, also indicating non-normality.

Since none of the data were normally distributed, the Spearman's rank correlation coefficient measure was used to determine the relationship between the dependent variable and each of the explanatory variables (Appendix B). For all of the buffers, the percent sand soil texture explanatory variable correlated most strongly with low nitrate concentrations, and this variable had rank correlation coefficients less than -0.45 for all buffers. Conversely, the percent silt soil texture explanatory variable had the strongest correlation with elevated nitrate concentrations for all buffers, and this variable's rank correlation coefficient was greater than 0.45 for all buffers.

Five explanatory variables for the 500-meter buffer data were not statistically significant at the $\alpha = 0.05$ level of significance. These variables included pit or water soil hydrologic group, onsite waste disposal on a parcel less then 0.004 km^2 in size, urban land cover, open water land cover, and population density. The p-values for these explanatory variables ranged from 0.0903 to 0.5798. Three of the variables that were not statistically significant for the 500-meter buffer data were also not statistically significant for the 1,000-meter buffer data. These variables included pit or water soil

hydrologic group, onsite waste disposal on a parcel less then 0.004 km² in size, and population density, and the p-values for these variables ranged from 0.1795 to 0.8679. In addition, the population density explanatory variable was not statistically significant at the $\alpha = 0.05$ level of significance for the 1,500-meter buffer data, and the p-value for this data was 0.7922. Furthermore, the variable including onsite waste disposal on parcels between 0.004 and 0.20 km² in size was not statistically significant at the $\alpha = 0.05$ level of significance for the 1,500-meter buffer data with a p-value of 0.3177. The Spearman's rank correlation coefficient measure enabled the identification of all data with p-values greater than 0.05 that were not considered to be statistically significant at the $\alpha = 0.05$ level of significance for each buffer size; therefore, these data were not included in logistic regression analysis.

5.1.2 Logistic Regression Analysis

Logistic regression analysis was performed using stepwise logistic regression procedures in order to create a model for each of the three different buffer sizes used for the study. A final model and corresponding buffer size were chosen based on various test statistics and model attributes. Multicollinearity diagnostic statistics were calculated for all of the final models in order to address any multicollinearity issues among explanatory datasets. When choosing a final model from the three models associated with the 500-meter, 1,000-meter, and 1,500-meter buffers, model significance, results for the Hosmer-Lemeshow goodness-of-fit test statistic, maximum rescaled r-square values, and percent concordance were determining factors for model selection. The Pearson residual statistic was then utilized in order to calculate residual values to show how well the final model fit the dependent dataset.

Once the final models associated with the three different buffer sizes were selected, multicollinearity diagnostics were run for each model in order to find out if any of the explanatory variables included in the models had multicollinearity issues (Table 9). The Tolerance and Variance Inflation Factor were examined for each model. A Tolerance greater than 0.4 and a Variance Inflation Factor less than 2.5 indicate that variables do not have multicollinearity issues (Allison, 1999). The final model for the 500-meter buffer included the total nitrogen input and percent silt soil texture explanatory variables. A tolerance value of 0.74456 and Variance Inflation Factor of 1.34308 indicated that the variables included in the model did not have any multicollinearity issues. Furthermore, the model for the 1,000-meter buffer included the percent silt soil texture and soil hydrologic group B explanatory variables, and these variables yielded a tolerance value of 0.57588 and a Variance Inflation Factor of 1.73648, thus signifying a lack of multicollinearity issues between the two variables. Additionally, the model for the 1,500-meter buffer yielded final variables of surface depression density and percent silt soil texture. These two variables had a Tolerance of 0.77227 and a Variance Inflation Factor of 1.29489, which indicates a lack of multicollinearity. Ultimately, each of the final models included percent silt soil texture as a variable, and none of the variables in the final models had multicollinearity issues.

Table 9. Multicollinearity diagnostics for the three models associated with different buffer sizes (South Middleton Township, 2001).

Model	Variables in Model	Tolerance	Variance Inflation Factor
Model for 500-Meter Buffer	Total Nitrogen Inputs	0.74456	1.34308
	Soil Texture Percent Silt		
Model for 1,000-Meter Buffer	Soil Texture Percent Silt	0.57588	1.73648
	Soil Hydrologic Group B		
Model for 1,500-Meter Buffer	Surface Depression Density	0.77227	1.29489
	Soil Texture Percent Silt		

Next, model significance, results for the Hosmer-Lemeshow goodness-of-fit test statistic, maximum rescaled r-square values, and percent concordance were determined in order to select a final model and buffer size based on these calculations (Table 10). All p-values for the Wald Chi-Square statistic for each model were statistically significant at the $\alpha = 0.05$ level of significance. The highest p-value was 0.0412 for the surface depression density variable in the model for the 1,500-meter buffer. In every model, the percent silt soil texture variable had the lowest p-value. The model for the 500-meter buffer had the lowest p-value when the p-values for the percent silt soil texture variable for all models were not taken into consideration. This p-value was 0.0051 for the total nitrogen inputs variable. In addition, p-values associated with the Hosmer-Lemeshow goodness-of-fit test statistic were greater than 0.05 for all of the models, which indicates that the estimates for all of the models fit the original data at an acceptable level. The model for the 1,000-meter buffer had the highest p-value, which was 0.6117, and the model for the 500-meter buffer had the lowest p-value of 0.0752. The maximum rescaled r-square values were very similar for each of the models. The model for the 500-meter buffer had the highest maximum rescaled r-square value of 0.3502, which means that this model had the strongest predictive power out of all of the models. Conversely, the model

for the 1,500-meter buffer had the lowest maximum rescaled r-square value of 0.3138, thus indicating that this model had the weaker predictive power out of the three models. Furthermore, the model with the highest percent concordance was the model for the 500-meter buffer with a value of 79.0, which implies that this model had the strongest prediction. On the other hand, the model for the 1,000-meter buffer had the lowest percent concordance of 77.3, thus suggesting that this model had the weakest prediction out of the three models.

Table 10. Various statistics utilized to choose a final model from the three models associated with different buffer sizes (South Middleton Township, 2001).

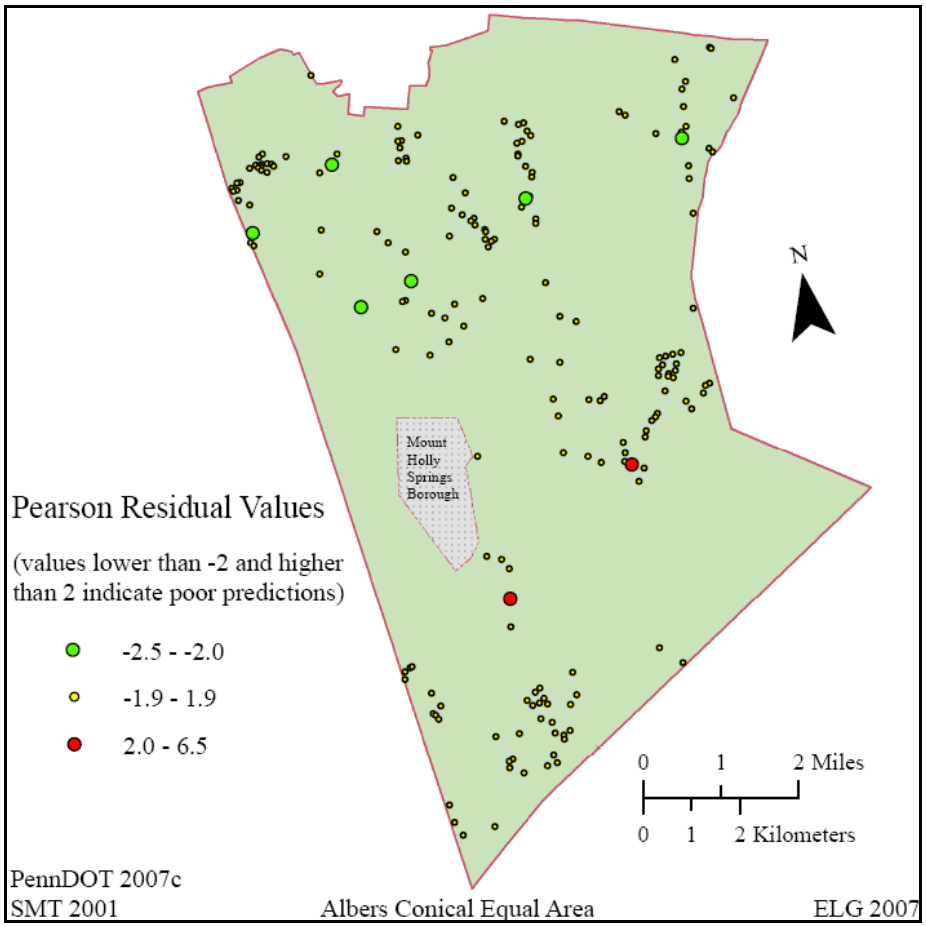
Model	Variables in Model	Model Significance		Hosmer-Lemeshow Goodness-of-Fit Test Statistic		Maximum Rescaled R-Square Values	Percent Concordance
		Wald Chi-Square	P-Value	Chi-Square	P-Value		
Model for 500-Meter Buffer	Total Nitrogen Inputs	7.8472	0.0051	14.2632	0.0752	0.3502	79.0
	Soil Texture Percent Silt	18.8845	<.0001				
Model for 1,000-Meter Buffer	Soil Texture Percent Silt	10.9709	0.0009	6.3175	0.6117	0.3270	77.3
	Soil Hydrologic Group B	5.1766	0.0229				
Model for 1,500-Meter Buffer	Surface Depression Density	4.1666	0.0412	9.1222	0.3321	0.3138	77.7
	Soil Texture Percent Silt	21.0468	<.0001				

Based on the results regarding model significance, the Hosmer-Lemeshow goodness-of-fit test statistic, maximum rescaled r-square values, and percent concordance, the model for the 500-meter buffer seemed to display the strongest predictions out of the three models. Although this model did not display the highest p-value for the Hosmer-Lemeshow goodness-of-fit test statistic, the p-value was still high enough to indicate that the model was well-calibrated. Due to the model's satisfactory calibration and strong predictions based on its maximum rescaled r-square value and percent concordance, this model was chosen as the final model for the study. Therefore,

the buffer size of 500 meters was determined to have the best-fit model that maximized the test statistics for nitrate concentrations exceeding a threshold of 4 mg/L.

Due to this finding, the Pearson residual statistic was calculated for the model associated with the 500-meter buffer in order to determine how well the model fit the dependent data, and the resulting residual values were mapped in relation to the well locations (Figure 5.1). Eight of the calculated Pearson residual values were greater than 2 or less than -2, thus indicating areas where the model either overpredicted or underpredicted nitrate concentrations. Six of the residual values overpredicted nitrate concentrations; therefore, actual nitrate concentrations at those wells were smaller than the predicted values. Conversely, two of the Pearson residual values underpredicted nitrate concentrations; thus, the actual nitrate concentrations in the dependent dataset were higher than the predicted values. Overpredictions generally occurred across the northern part of the township, while the two underpredictions arose in the south-central portion of the township.

Figure 5.1. Mapped Pearson residual values (PennDOT, 2007c; South Middleton Township, 2001).



Since the maximum rescaled r-square value associated with the final model was below 0.5, it was not feasible to create maps based on the probability of nitrate concentrations exceeding 4 mg/L (Lindsey *et al.*, 2006). Therefore, the findings regarding explanatory variables impacting nitrate concentrations within the township are presented in order to improve knowledge and awareness concerning the occurrence of nitrate in ground water. When interpreting results, it is important to keep in mind that the dependent data were collected in 2000 and 2001, and all explanatory datasets represent the landscape's condition in 2001 or as closely to this year as possible. Therefore, any results associated with the models created using these datasets most accurately represent South Middleton Township's environmental characteristics in 2001. Although these

results can be used as a reference in addition to what is currently occurring within the township, they most accurately portray the township as it was in 2001.

Chapter 6

Discussion

6.1 Statistical Analysis

Results regarding univariate and logistic regression analysis revealed important information regarding the data and the final model itself. The Shapiro-Wilk analysis showed how both the dependent and independent datasets used for the study were not normally distributed. In addition, the Spearman's rank correlation coefficient measure determined that specific variables were not statistically significant at the $\alpha = 0.05$ level of significance. Logistic regression analysis enabled a model to be chosen for each of the three buffer sizes, and multicollinearity diagnostics determined the absence of multicollinearity issues among data included in the three models. Next, the results of different test statistics determined that the model for the 500-meter buffer was the model displaying the strongest predictions out of the three final models. Therefore, the Pearson residual statistic was calculated for the model associated with the 500-meter buffer, thus revealing areas where the model best fit the independent dataset.

6.1.1 Univariate Analysis

The results from the Shapiro-Wilk analysis concluded that none of the data were normally distributed, thus the Spearman's rank correlation coefficient measure was utilized to determine the relationship between the dependent variable and all of the independent variables. For each of the three buffers, the percent silt soil texture variable correlated most strongly with high concentrations of nitrate. Likewise, it is noticeable when comparing Figures 4.1 and 4.14 that for the most part, the highest concentrations of nitrate correlate with areas containing a higher percentage of silt in soil. Although silty soils have moderate leaching potential, high percentages of silt in soil are also a good indicator of high nitrate concentrations because this variable is representative of other important variables (Smith & Cassel, 1991). Silty soils are derived from carbonate bedrock, which is responsible for karst landscape features, and silty soils are also prime agricultural soils (South Middleton Township, 1999). Therefore, this variable may be representing other factors that could be responsible for high concentrations of nitrate such as large amounts of nitrogen being applied to the agricultural landscape or karst features that allow nitrate to easily penetrate ground water supplies.

In addition, the percent sand soil texture variable correlated most strongly with low concentrations of nitrate for all buffers. Sandy soils have the highest leaching potential out of all soil texture types because of the coarse texture associated with sand, thus it would not seem logical for low concentrations of nitrate to have a high correlation with large percentages of sand in soil (Smith & Cassel, 1991). On the other hand, it must be kept in mind that the central portion of South Middleton Township possesses a thick colluvium and alluvium stratum that reaches a thickness of 61 meters in the township (Figure 3.5) (Root, 1968; Sevon, 2001). As surface runoff percolates through this thick

stratum, nitrate is delayed from leaching into ground water, and there is a greater chance that denitrification occurs before surface water runoff leaches to ground water (Knox & Moody, 1991). Furthermore, sandy soils are derived from crystalline bedrock, which does not contain karst landscape features because it is not as vulnerable to the dissolution process that creates those features (Winter *et al.*, 1998). Therefore, crystalline bedrock is also not as flat as carbonate bedrock, which means that areas of crystalline bedrock are not as prone to development and contain more forested land than carbonate bedrock.

The variable that was not statistically significant at the $\alpha = 0.05$ level of significance for all buffer sizes was population density (Appendix B). Population density was statistically insignificant for all buffer sizes because the dependent dataset includes data for domestic wells. According to Figure 3.10, densely populated areas within the township, such as Boiling Springs and areas surrounding Mt. Holly Springs and Carlisle Boroughs, were serviced in 2001 by public water suppliers, thus eliminating the need for domestic wells within these areas (Cumberland County Planning Commission, 2001). Therefore, most of the well data, regardless of the associated nitrate concentration, are only representative of areas with the lowest population range, since densely populated places in the township utilized public water in 2001 instead of domestic wells.

In addition, the variable including onsite waste disposal on a parcel between 0.004 and 0.20 km^2 in size was not statistically significant at the $\alpha = 0.05$ level of significance for the 500-meter and 1,000-meter buffer sizes. Figure 4.9 shows that there are very few parcels within the township less then 0.004 km^2 in size that utilize onsite waste disposal methods. Almost all of the parcels in South Middleton Township are greater than 0.004 km^2 in size, so this variable poorly represents any of the land area within the township.

Furthermore, the pit or water soil hydrologic group was found to be statistically insignificant at the $\alpha = 0.05$ level of significance for both 500-meter and 1,000-meter buffers (Appendix B). Additionally, open water land cover was statistically insignificant for the 500-meter buffer (Appendix B). The reasoning for these variables to be statistically insignificant for these buffers is very similar to the reasoning behind the insignificance of the variable including parcels less then 0.004 km^2 in size that utilize onsite waste disposal methods for all buffers. To begin with, there are already few areas within the township with pit or water hydrologic soil groups (Figure 4.15). When a 500-meter or 1,000-meter buffer around the dependent variable further limits that data, they will not be represented well enough to determine statistical significance. The same is true for open water land cover within a 500-meter buffer of the dependent data, especially since open water only accounts for 0.4 percent of the land area within South Middleton Township to begin with (Figure 3.8) (USGS, 2001). Subsequently, these variables were found to be statistically insignificant according to the Spearman's rank correlation coefficient measure and were not utilized for logistic regression analysis.

6.1.2 Logistic Regression Analysis

After logistic regression analysis was performed, none of the variables in the final models had multicollinearity issues, but each of the models for the three buffer sizes included the percent silt soil texture variable. Since this variable correlated most strongly with high concentrations of nitrate according to the Spearman's rank correlation coefficient measure for each of the three buffers, the fact that it was also included in all of the models was not surprising. The model for the 500-meter buffer also yielded the total nitrogen inputs as another variable in the final model, and this is a logical variable to be included since nitrate is one of the four primary forms of nitrogen (Canter *et al.*, 1987).

On the other hand, the final model for the 1,000-meter buffer also included hydrologic soil group B as a final variable. Since this hydrologic soil group characteristically includes silt loam soils, it is surprising that this variable did not show any multicollinearity issues with the percent silt soil texture variable. In addition, the final model for the 1,500-meter buffer included the surface depression density variable, and this is a reasonable variable to include since surface depressions are indicators of areas where there are unstable areas in the bedrock that are penetrable by surface water runoff, thus causing nitrate to easily reach ground water supplies.

Next, model significance, results for the Hosmer-Lemeshow goodness-of-fit test statistic, maximum rescaled r-square values, and percent concordance were determined, and the model associated with the 500-meter buffer was selected as the final model for the study based on these criteria. The p-values for this model were very low and indicated that the model was significant, while the p-value for the Hosmer-Lemeshow goodness-of-fit test statistic was high enough to show that the estimates for the model fit the original data at an acceptable level. Additionally, this model had the highest maximum rescaled r-square value and percent concordance out of all of the models. These statistics indicated that the variables chosen for this model were better predictors for this buffer size than the associated variables included in any of the models for the other buffer sizes. The model for the 500-meter buffer may have been more predictive than the models associated with the larger buffer sizes because the 500-meter buffer was able to address characteristics and occurrences located within proximity of each well, while the larger buffers enveloped too large of a land area.

A significant finding was that differences among the final variables for each buffer size seemed to be a result of scale differences. Although the percentage of silt in

soils was a final variable for each of the three models, the remaining variable for the 500-meter buffer was total nitrogen inputs, while the remaining variable for the 1,500-meter buffer was surface depression density. Interestingly, it seems that a larger buffer may have been necessary to detect the processes associated with surface depressions that impact ground water quality. For example, although a surface depression may be located hundreds of meters from a well, it is still capable of impacting ground water quality at that well. Surface depressions can provide a direct path for contaminants to enter ground water supplies, and since they are associated with carbonate bedrock and karst features, this means that ground water can move very quickly in these areas, thus impacting ground water quality at a well that is hundreds of feet away (Winter *et al.*, 1998). If there are no surface depressions within 500 meters of a well, then a 500-meter buffer will not detect ground water quality impacts caused by surface depressions.

Conversely, ground water quality at a well can be directly impacted if the total nitrogen inputs within proximity of that well are substantial. Although other factors such as bedrock and soil types come into play, there is still a potential for surface runoff to quickly leach into ground water, thus causing elevated nitrate concentrations in areas where total nitrogen inputs are high. Therefore, elevated nitrate inputs seem to be more detectable within a 500-meter buffer of a well, which was the smallest buffer size utilized for the study. Since different variables were found to be more significant when they were associated with different buffer sizes, this suggests that the final model may not contain the most statistically significant variables in this study. For example, if the total nitrogen input variable associated with the 500-meter buffer and the surface depression density variable associated with the 1,500-meter buffer were included in the same model, the

model's predictive power could be much stronger than the predictive power of the final model associated with this study.

The results of the final model for the 500-meter buffer suggest that elevated nitrate concentrations within South Middleton Township are not a result of one variable but of a combination of two different variables. The percentage of silt within soils and nitrogen inputs to the landscape are shown to correlate strongly with elevated nitrate concentrations. Although a higher percentage of silt within soils corresponds with a carbonate lithologic unit, the same is true of higher percentages of clays within soils (Knox & Moody, 1991). Clays within soils are known to delay nitrate from reaching ground water supplies, thus allowing additional time for denitrification to occur, which is capable of decreasing the amount of nitrate reaching ground water (Canter *et al.*, 1987; Knox & Moody, 1991). Figures 4.12 and 4.14 illustrate the fact that where there are larger percentages of silt in the soils in the township, there are also larger percentages of clay. Conversely, wherever high silt and clay percentages within soils occur simultaneously, it appears that the percentage of silt is consistently double the percentage of clay. A majority of soils spread across the township contain 41 to 60 percent silt, while these same areas contain only 11 to 30 percent clay. Ultimately, silt is the dominant soil texture type across a majority of the landscape. Since this soil texture type displays a moderate leaching potential, it is understandable that when silty soils are coupled with substantial nitrogen inputs across the landscape, this occurrence is very likely to result in elevated nitrate concentrations in ground water.

After the final model was chosen, the Pearson residual statistic was calculated for the model, and the resulting values were mapped (Figure 5.1). The mapped values showed residual values less than -2 and greater than 2 in the northern and south-central

portions of the township, thus revealing where the model made poor predictions. Six of the residual values showed overpredictions in the northern part of the township, while two of the values displayed underpredictions in the south-central portion of South Middleton Township. Two of the values overpredicting nitrate concentrations had original nitrate concentrations of 3.7 and 3.8 mg/L, which are both very close to the 4 mg/L threshold. Since these two values were very close to the threshold, the overpredictions associated with these two wells were not substantial.

In addition, three more of the values that made overpredictions were located on parcels that appeared to have a large amount of forested land according to 2003 aerial photographs (USGS, 2004). As indicated by the aerial photographs, these three overpredictions were located on residential parcels where homeowners chose to leave large fragments of forested areas intact on their properties (USGS, 2004). Forested areas can assist with ground water quality because once nitrogen reaches the landscape and has undergone nitrification, the resulting nitrate can be readily used by plants since it is water soluble, thus causing it to be absorbed easily by plant roots (Makuch and Ward, n.d.). This process is important because nitrate that has been absorbed by plant roots is no longer capable of leaching into ground water supplies (Makuch and Ward, n.d.).

Conversely, the last value displaying an overprediction had an original nitrate concentration of 1.8 mg/L. This well was located in an area containing carbonate bedrock on a parcel utilizing an onsite waste disposal method that was completely surrounded by agricultural lands. In addition, the parcel contained almost no forested land cover according to 2003 USGS (2004) aerial photographs. Each of these indicators suggests that the nitrate concentration at this well should have been higher than the 4 mg/L threshold. Since the nitrate concentration at this well did not exceed the threshold

and none of the remaining explanatory variables justify the occurrence, this finding suggests that there may be a significant factor, such as ground water flows in karst terrain or a data quality issue, which was not able to be addressed in this study.

Both of the residual values indicating underpredictions were located on parcels utilizing septic tanks in 2001, thus suggesting that there could have been issues with the septic tanks on these parcels during this time period (Cumberland County Planning Commission, 2001). When septic tank systems are designed, built, maintained, or situated inadequately, they are more susceptible to leaching excessive nitrate, thus threatening ground water quality (Canter *et al.*, 1987; Makuch and Ward, n.d.). When these instances occur, the effluent from septic tanks is not exposed to the removal mechanisms associated with soils because the soil is overloaded, the effluent is percolating too quickly through the soil, or the effluent is being discharged below the soil profile (Canter *et al.*, 1987). Therefore, septic tanks experiencing problems such as these are capable of causing elevated nitrate concentrations in ground water.

6.2 Challenges

The results of the final model for the 500-meter buffer were statistically significant, but the predictive power of the model was not strong enough to predict the occurrence of nitrate concentrations exceeding 4 mg/L throughout South Middleton Township. The literature states that variables such as land cover, nitrogen inputs, presence of onsite waste disposal, population density, bedrock type, soil characteristics, and presence of sinkholes or surface depressions are capable of impacting nitrate concentrations in ground water (Canter *et al.*, 1987; Canter, 1997; Knox & Moody, 1991; Smith & Cassel, 1991). In addition, other studies have been performed in the past regarding nitrate concentrations in ground water that yielded models with a strong

predictive power (Eckhardt & Stackelberg, 1995; Tesoriero & Voss, 1997; Nolan *et al.*, 2002; Hitt & Nolan, 2005; Rupert, 2003; Greene *et al.*, 2005; Gurdak & Qi, 2006; LaMotte & Greene, 2007). It is difficult to determine why the predictive power of this model was not strong enough, but these factors may include the challenges associated with karst terrain, scale, or spatial autocorrelation.

One difference between this study and other studies that have been performed is that a majority of the study area was located on karst terrain. Ground water interactions in areas with karst terrain are difficult to address because ground water is capable of traveling a few miles in a day through underground networks of conduits; therefore, nitrate concentrations in ground water may not reflect the anthropogenic or hydrogeologic variables that are present on the land surface surrounding domestic wells (Winter *et al.*, 1998). In order to find out more about elevated nitrate concentrations in South Middleton Township, ground water flow paths need to be delineated in the township through dye tracing techniques, which involve injecting dyes into the ground water at a specific point and performing tests in order to determine if they are detected in other areas (Winter *et al.*, 1998).

In addition, studies of this nature are not typically performed at local levels. Of the previous studies examined, only one of them had a study area similar in size to that of South Middleton Township. The study was performed by Gardner and Vogel (2005) for Nantucket Island, Massachusetts, which has an area of 124 km^2, while South Middleton Township has an area of 127 km^2. Predictive maps regarding elevated nitrate concentrations were not created for this study, although the predictive power of the final model for this study was very high. Instead of predicting the probability of elevated nitrate concentrations occurring throughout the study area, this study sought to determine

significant relationships between land use and ground water nitrate concentrations. The study claimed to prove its usefulness to planning and land managers through the usage of publicly available data and relatively simple regression methods.

It seems that studies involving logistic regression analysis regarding ground water quality at local levels are not commonly utilized to predict the probability of elevated nitrate concentrations in ground water. When studies are performed at the local level, local factors such as the karst features previously discussed, have much more weight on the outcomes of the study. For instance, the study performed by Greene *et al.* (2005) looked at the Mid-Atlantic region where a small portion of the study area contained carbonate bedrock relative to the study area size. The karst features associated with the carbonate bedrock did not have the opportunity to impact the study outcomes because the area containing this bedrock type was so small when taking the whole study area into account.

Also, studies such as this may not commonly be performed at the local level because data for local areas are not always readily available. For example, total nitrogen input data at the municipality level were not available for this study. Therefore, it was determined that the available county level data would be applied to the municipality. Utilizing county level data to represent a municipality presents issues because the municipality may not be representative of the rest of the county. Consequently, many assumptions must be made regarding conditions within the municipality when municipality level data are unavailable or are not collected firsthand in the field.

Another challenge associated with this study involves spatial autocorrelation. Spatial autocorrelation regards the spatial distribution of data and assumes that data located closer together will display similar attributes, while those data dispersed further

apart will have more diverse characteristics. When examining Figure 4.1, there appear to be many areas where there are clusters of wells situated closely together, which means that positive spatial autocorrelation may be impacting the results of the study. Since the data are clustered together in several areas, this means that the 500-meter, 1,000-meter, and 1,500-meter buffers associated with the wells in these clusters are all overlapping. Since the buffers for these wells overlap so much, this means that the data extracted according to these buffers are relatively homogeneous, which can lead to the same data being included in the final analysis for multiple wells that may have been representing very similar sample points.

When similar data are accounted for too many times, this data will impact the outcomes of the study because certain characteristics will become overrepresented in the statistical analysis while other data simultaneously become underrepresented and will not seem as significant as the clustered data. For example, a majority of the wells in the dependent dataset were located in areas with carbonate bedrock, silty soils, and agricultural land cover with the highest total nitrogen input values, and this means that these areas may have been overrepresented while those areas with crystalline bedrock, sandy soils, and forested land cover were not represented so well. Therefore, there is a possibility that some of the variables associated with the final models for each buffer were overrepresented in the final dataset used for statistical analysis, thus causing them to appear in the final models.

6.3 Future Studies

If a study similar to this one would be performed in the future, it would be beneficial to divide the current study area, address a larger study area, or utilize different buffer sizes for different explanatory variables. Dividing the current study area according

to bedrock type and physiographic province in order to separate the unique carbonate areas containing karst features from areas with a differing lithology would be an interesting way to approach the current study. For example, the study area could be divided according to those areas that predominantly have carbonate bedrock in the Ridge and Valley physiographic province and those areas that predominantly contain crystalline bedrock in the Blue Ridge physiographic province (Figures 3.2 and 4.11). This would divide the study area into a carbonate Ridge and Valley study area containing 139 wells and a crystalline Blue Ridge study area containing 51 wells. Although the dependent data would not be divided equally, the outcomes associated with a study involving this divided study area may still reveal new conclusions to be drawn about the uniqueness of South Middleton Township's geology.

Furthermore, along with the dependent dataset, the explanatory data associated with the two new study areas would be divided accordingly, and logistic regression analysis could be used in order to create new models for the divided study area. The resulting variables, test statistics, and Pearson residual values associated with each of the models could then be compared in order to determine how the differing bedrock types influence the results of the study. The results associated with such a comparison would have the potential to show how carbonate areas containing karst features are capable of impacting ground water quality in relation to anthropogenic and hydrogeologic explanatory variables.

There is also the possibility that a township with varied geology is too small a land area for this study method, and dividing the study area would further limit the results of the study. Therefore, it might be interesting to base future studies on a larger geologic unit-based region. For example, instead of just looking at the crystalline Blue Ridge

portion of South Middleton Township, a similar study could incorporate the entire crystalline Blue Ridge physiographic province into the study area. Choosing a larger study area based on a specific geologic unit would provide the broader land area needed to apply county level data to the study area and the geologic uniformity to ensure that one type of geology is not represented more in the dependent dataset than another.

Another interesting way to look at this study in the future would be to use different buffer sizes for different explanatory variables instead of using the same buffer size for all explanatory variables. For example, in the study performed by Rupert (2003) for the state of Colorado, univariate logistic regression was utilized in order to determine which buffer size had the most significant relation with ground water quality and land cover classifications. Rupert (2003) determined an optimum buffer size of 2,000 meters for agricultural land cover and an optimum buffer size of 500 meters for urban land cover. A similar approach could be used for South Middleton Township by selecting buffers of various radii to evaluate in order to determine the most significant relation between a buffer size and elevated nitrate concentrations in ground water and various explanatory variables. Ultimately, the different buffer sizes associated with different explanatory variables would be used to extract data for statistical analysis.

For example, as it was discussed earlier, it would be interesting to see if the 500-meter buffer associated with the total nitrogen input variable or the 1,500-meter buffer associated with the surface depression density variable were the optimal buffer sizes to be used for these variables. If these buffers were the optimal sizes to utilize with those variables, then these two variables with their different buffer sizes would be included in the explanatory data for the model, and it would be interesting to see if they would both be included in the final model resulting from the logistic regression analysis. When using

the techniques associated with different buffer sizes, it would also be interesting to see how much stronger the predictive power of the final model associated these techniques might be than the predictive power of the final model associated with this study. Using distinct buffer sizes for each different variable would enable the demonstration of how different variables interact on different scales and are statistically significant at the scales that represent them most appropriately.

Chapter 7

Conclusion

In this study, ground water data describing nitrate concentrations at 190 privately owned domestic drinking water wells in South Middleton Township were correlated to the percentage of silt in soils and total nitrogen inputs resulting from atmospheric deposition and non-farm and farm fertilizer and manure applications. Logistic regression analysis indicates that nitrate concentrations exceeding 4 mg/L in the township in 2001 are directly correlated with these two explanatory variables. These results illustrate that the combination of the increased amounts of nitrogen inputs across the landscape from atmospheric deposition and manure and fertilizer applications and high percentages of silt within soils within a 500-meter radius around domestic wells can be predictors of the presence of elevated nitrate concentrations in ground water in South Middleton Township.

The strength of these correlations supports the premise that large amounts of nitrogen inputs and increased percentages of silt within soils are capable of having an impact on ground water quality. These impacts can be attributed to the substantial

application of nitrogen across agricultural areas typically containing silty soils, which have moderate leaching potential and are characteristic of areas with carbonate bedrock. The predictive power of the correlations was statistically significant but not strong enough to predict nitrate concentrations exceeding 4 mg/L throughout the township, thus determining a need for future research within the township involving the division of the current study area by geology and physiographic province, the incorporation of a larger study area that does not contain a varied geology, or a similar study involving buffers of varying sizes for different explanatory variables. Conversely, the statistical significance of the correlations indicates that total nitrogen inputs and percentage of silt in soils are predictors of ground water quality within the township.

7.1 Findings

Important findings associated with this study are those involving scale. Ultimately, it was discovered that different variables become statistically significant when they are addressed at different scales. For example, the variable including the percentage of silt in soil was statistically significant across all scales, but the total nitrogen inputs variable was only significant for the model associated with the 500-meter buffer, while the surface depression density variable was only significant for the model associated with the 1,500-meter buffer. These findings show that different variables operate at different scales and that they will become statistically significant factors regarding elevated nitrate concentrations in ground water depending on the scale at which they are examined.

In addition, the findings of this study show that those studies involving logistic regression that are meant to predict the probability of elevated nitrate concentrations in ground water may be less applicable at the local level, and this may be especially true for

those areas with varying geology. Not many studies have been performed at the local level regarding elevated nitrate concentrations in ground water. According to this study, it seems that performing smaller scale studies of this nature may improve issues involving explanatory data accuracy and spatial autocorrelation issues associated with dependent datasets. Furthermore, performing studies for areas with a uniform geology eliminates issues involving the different physical properties and factors associated with varying geologies that are capable of impacting study results.

7.2 Recommendations

The meaning of these results is evidently useful to decision-makers and officials in charge of water and land management and enables the improvement of knowledge and awareness concerning the occurrence of elevated nitrate concentrations. A study such as this may be more useful to county or regional land managers and planners, rather than those who perform the same duties at the municipality scale. On the other hand, all land managers can be made more aware of the extent and associated impacts that can be caused by various explanatory variables addressed in this study. For example, it is helpful for all land managers to be conscious of the impacts that nitrogen applied to the landscape can have within the immediate area of application as opposed to the impacts that surface depressions can have on ground water quality across a much broader area. Therefore, it is important for those in charge of water and land management to keep in mind the scales that different variables will operate at and how they interact.

In relation to South Middleton Township, the results of the study determined significant relationships among elevated nitrate concentrations in ground water and total nitrogen inputs and percentage of silt in soil within 500 meters of wells. According to these results, nitrogen inputs should be more closely managed for agricultural areas

across the township, since fertilizer and manure inputs on agricultural lands are the values contributing the most to total nitrogen inputs and since agricultural areas typically have high percentages of silt. Also, according to these results, residential areas located close to agricultural fields are those that should be most concerned about testing their private wells in order to be proactive and recognize ground water quality issues that could potentially become a larger problem in the future.

Appendices

Appendix A

Nitrate concentration data for wells in South Middleton Township (South Middleton Township, 2001).

Well Identification Number	Nitrate Concentrations in mg/L	Categorical Variable Based on the 4 mg/L Threshold	Well Identification Number	Nitrate Concentrations in mg/L	Categorical Variable Based on the 4 mg/L Threshold
1	3.8	0	51	8.56	1
2	3.79	0	52	10.7	1
3	6.96	1	53	4.06	1
4	6.26	1	54	6.52	1
5	4.11	1	55	6.71	1
6	6.68	1	56	12.4	1
7	8.91	1	57	14.8	1
8	6.39	1	58	17.9	1
9	8.22	1	59	7	1
10	5.78	1	60	9.48	1
11	3.23	0	61	12.4	1
12	4.24	1	62	7.13	1
13	5.68	1	63	2.35	0
14	1.93	0	64	6.86	1
15	7.56	1	65	4.1	1
16	4.31	1	66	6.72	1
17	5.84	1	67	5.28	1
18	7.83	1	68	7.09	1
19	5.88	1	69	9.88	1
20	4.33	1	70	8.43	1
21	3.25	0	71	5.98	1
22	3.79	0	72	7.21	1
23	4.65	1	73	8.16	1
24	4.38	1	74	7.5	1
25	0.25	0	75	8.22	1
26	18.4	1	76	4.46	1
27	4.94	1	77	4.13	1
28	5.06	1	78	7.72	1
29	6.24	1	79	3.67	0
30	3.4	0	80	6.23	1
31	4.21	1	81	7.54	1
32	6.24	1	82	4.38	1
33	6.45	1	83	3.96	0
34	4.78	1	84	6.8	1
35	3.83	0	85	5.38	1
36	3.45	0	86	6.73	1
37	8.32	1	87	7.45	1
38	4.23	1	88	2.47	0
39	2.09	0	89	6.15	1
40	6.48	1	90	1.66	0
41	7.36	1	91	1.75	0
42	9.43	1	92	5.61	1
43	6.55	1	93	5.34	1
44	1.1	0	94	0.99	0
45	5.34	1	95	2.67	0
46	4.02	1	96	0.5	0
47	4.51	1	97	0.25	0
48	5.99	1	98	6.45	1
49	7.86	1	99	8.55	1
50	8.15	1	100	1.94	0

Well Identification Number	Nitrate Concentrations in mg/L	Categorical Variable Based on the 4 mg/L Threshold	Well Identification Number	Nitrate Concentrations in mg/L	Categorical Variable Based on the 4 mg/L Threshold
101	7.94	1	151	0.78	0
102	6.24	1	152	0.25	0
103	3.63	0	153	1.73	0
104	10.4	1	154	2.38	0
105	6.95	1	155	2.18	0
106	3.63	0	156	0.82	0
107	6.04	1	157	1.58	0
108	9.88	1	158	1.04	0
109	6.32	1	159	6.26	1
110	6.77	1	160	5.36	1
111	2.95	0	161	7.64	1
112	4.64	1	162	5.48	1
113	2.87	0	163	0.25	0
114	3.52	0	164	5.5	1
115	1.71	0	165	0.77	0
116	1.59	0	166	3.62	0
117	1.35	0	167	4.83	1
118	0.8	0	168	2.26	0
119	1.41	0	169	4.99	1
120	0.25	0	170	9.13	1
121	0.5	0	171	1.89	0
122	1.19	0	172	4.29	1
123	0.5	0	173	4.67	1
124	3.57	0	174	7.12	1
125	8.65	1	175	3.03	0
126	7.64	1	176	1.4	0
127	0.25	0	177	7.12	1
128	9.92	1	178	0.5	0
129	0.5	0	179	4.91	1
130	2.4	0	180	7.09	1
131	9.78	1	181	7	1
132	3.3	0	182	2.78	0
133	3.58	0	183	5.32	1
134	2.32	0	184	3.01	0
135	0.5	0	185	7.48	1
136	2.2	0	186	3.54	0
137	0.25	0	187	4.16	1
138	0.51	0	188	5.25	1
139	1.97	0	189	1.3	0
140	3.37	0	190	4.91	1
141	0.25	0			
142	0.25	0			
143	0.5	0			
144	5.06	1			
145	8.01	1			
146	0.25	0			
147	0.25	0			
148	0.5	0			
149	6.24	1			
150	0.25	0			

Appendix B

Spearman's rank correlation coefficient statistical data (South Middleton Township, 2001).

Data	Explanatory Variables	Spearman's Rank Correlation Coefficient	
		Coefficient	P-Value
500-Meter Buffer Data	Crystalline Bedrock	-0.2778	0.0001
	Soil Hydrologic Group D	-0.26133	0.0003
	No Onsite Waste Disposal	-0.25119	0.0005
	Onsite Waste Disposal on a Parcel Greater Than 0.020 km² in Size	0.24639	0.0006
	Onsite Waste Disposal on a Parcel between 0.004 and 0.20 km² in size	-0.17073	0.0185
	Wetlands Land Cover	-0.17028	0.0188
	Siliciclastic Bedrock	-0.16833	0.0203
	Soil Hydrologic Group Pit or Water	0.12322	**0.0903**
	Onsite Waste Disposal on a Parcel Less Than 0.004 km² in Size	0.08569	**0.2398**
	Urban Land Cover	0.05527	**0.4488**
	Open Water Land Cover	-0.05214	**0.475**
	Population Density	-0.04042	**0.5798**
	Soil Texture Percent Sand	-0.48006	<.0001
	Forested Land Cover	-0.3892	<.0001
	Soil Hydrologic Group C	-0.30402	<.0001
	Carbonate Bedrock	0.31113	<.0001
	Surface Depression Density	0.31353	<.0001
	Soil Hydrologic Group B	0.32578	<.0001
	Sinkhole Density	0.33174	<.0001
	Soil Texture Percent Clay	0.36679	<.0001
	Total Nitrogen Inputs	0.43291	<.0001
	Agricultural Land Cover	0.4333	<.0001
	Soil Texture Percent Silt	0.45853	<.0001
1,000-Meter Buffer Data	Onsite Waste Disposal on a Parcel between 0.004 and 0.20 km² in Size	-0.24655	0.0006
	Wetlands Land Cover	-0.19998	0.0057
	Open Water Land Cover	-0.17571	0.0153
	Siliciclastic Bedrock	-0.16359	0.0241
	Urban Land Cover	0.15404	0.0338
	Soil Hydrologic Group Pit or Water	-0.0978	**0.1795**
	Onsite Waste Disposal on a Parcel Less Than 0.004 km² in Size	0.06144	**0.3997**
	Population Density	-0.01215	**0.8679**
	Soil Texture Percent Sand	-0.45309	<.0001
	Forested Land Cover	-0.44428	<.0001
	Soil Hydrologic Group C	-0.40014	<.0001
	Soil Hydrologic Group D	-0.36899	<.0001
	Crystalline Bedrock	-0.34332	<.0001
	No Onsite Waste Disposal	-0.32394	<.0001
	Sinkhole Density	0.33309	<.0001
	Onsite Waste Disposal on a Parcel Greater Than 0.020 km² in Size	0.37562	<.0001
	Surface Depression Density	0.37954	<.0001
	Carbonate Bedrock	0.37998	<.0001
	Soil Texture Percent Clay	0.39115	<.0001
	Soil Hydrologic Group B	0.4241	<.0001
	Agricultural Land Cover	0.44977	<.0001
	Total Nitrogen Inputs	0.45547	<.0001
	Soil Texture Percent Silt	0.46095	<.0001

Bold indicates that p-values are not statistically significant at the $\alpha = 0.05$ level of significance

Buffer Size	Explanatory Variables	Spearman's Rank Correlation Coefficient	
		Coefficient	P-Value
1,500-Meter Buffer Data	Wetlands Land Cover	-0.22845	0.0015
	Onsite Waste Disposal on a Parcel Less Than 0.004 km² in Size	0.21906	0.0024
	Siliciclastic Bedrock	-0.15946	0.028
	Onsite Waste Disposal on a Parcel between 0.004 and 0.20 km² in Size	-0.07287	**0.3177**
	Population Density	0.01924	**0.7922**
	Soil Texture Percent Sand	-0.4596	<.0001
	Forested Land Cover	-0.43048	<.0001
	Soil Hydrologic Group C	-0.41363	<.0001
	Soil Hydrologic Group D	-0.40213	<.0001
	Crystalline Bedrock	-0.3908	<.0001
	Open Water Land Cover	-0.33482	<.0001
	No Onsite Waste Disposal	-0.32581	<.0001
	Soil Hydrologic Group Pit or Water	-0.28228	<.0001
	Sinkhole Density	0.33238	<.0001
	Onsite Waste Disposal on a Parcel Greater Than 0.020 km² in Size	0.34772	<.0001
	Urban Land Cover	0.36498	<.0001
	Soil Texture Percent Clay	0.4127	<.0001
	Carbonate Bedrock	0.41938	<.0001
	Surface Depression Density	0.42205	<.0001
	Agricultural Land Cover	0.43399	<.0001
	Total Nitrogen Inputs	0.43611	<.0001
	Soil Hydrologic Group B	0.44744	<.0001
	Soil Texture Percent Silt	0.47153	<.0001
Bold indicates that p-values are not statistically significant at the $\alpha = 0.05$ level of significance			

References

Adams County Office of Planning and Development. 1991. Comprehensive Plan: Adams County, Pennsylvania, accessed August 7, 2007 at URL *http://www.adamscounty.us/adams/cwp/view.asp?A=1642&Q=472580*.

Adams County Office of Planning and Development. 2001. Adams County Pennsylvania: Water Supply and Wellhead Protection Plan, accessed August 7, 2007 at URL *http://www.adamscounty.us/adams/cwp/view.asp?A=1642&Q =472580*.

Allison, P.D. 1999. Logistic regression using the SAS system: Theory and application. SAS Institute Inc.: Cary, NC.

Becher, A.E. and Root, S.I. 1981a. Bedrock Geologic Map Showing the Hydrology of the Northern Part of the Cumberland Valley, Cumberland County, Pennsylvania, digital map, accessed October 16, 2007 at URL *http://www.libraries.psu.edu/ emsl/guides/pageomaps.html*.

Becher, A.E. and Root, S.I. 1981b. Groundwater and geology of the Cumberland Valley, Cumberland County, Pennsylvania. Harrisburg: Commonwealth of Pennsylvania, Department of Environmental Resources, Office of Resources Management, Bureau of Topographic and Geologic Survey.

Canter, L.W. 1997. Nitrates in groundwater. Lewis Publishers, Inc.: Boca Raton, FL.

Canter, L.W., Knox, R.C., and Fairchild, D.M. 1987. Ground water quality protection. Lewis Publishers, Inc: Chelsea, MI.

Cumberland County Planning Commission. 2001. Tax Parcels: CCPC, digital data, obtained July 13, 2007 from Cumberland County Planning Commission, Cumberland County, Pennsylvania.

Driscoll, C. and Lambert, K.F. 2003. Nitrogen pollution: from the sources to the sea. Hubbard Brook Research Foundation. Science Links Publication: Hanover, NH.

Eckhardt, D.A.V. and Stackelberg, P.E. 1995. Relation of ground-water quality to land use on Long Island, New York. *Ground Water*, 33 (6), 1019-1033.

Environmental Systems Research Institute. 2000a. Canada Provinces, digital data, obtained September 3, 2007 from Shippensburg University, Shippensburg, Pennsylvania.

Environmental Systems Research Institute. 2000b. Mexico States, digital data, obtained September 3, 2007 from Shippensburg University, Shippensburg, Pennsylvania.

Environmental Systems Research Institute. 2000c. United States States, digital data, obtained September 3, 2007 from Shippensburg University, Shippensburg, Pennsylvania.

Focazio, M.J., Reilly, T.E., Rupert, M.G., and Helsel, D.R. 2002. Assessing ground-water vulnerability to contamination: Providing scientifically defensible information for decision makers. US Geological Survey Circular 1224.

Focazio, M.J., Tipton, D., Shapiro, S.D., and Geiger, L.H. 2006. The chemical quality of self-supplied domestic well water in the United States. *Ground Water Monitoring & Remediation*, 26 (3): 92-104.

Gardner, K.K. and Vogel, R.M. 2005. Predicting ground water nitrate concentration from land use. *Ground Water*, 43 (3), 343-352.

Greene, E.A., LaMotte, A.E., and Cullinan, Kerri-Ann. 2005. Ground-water vulnerability to nitrate contamination at multiple thresholds in the Mid-Atlantic region using spatial probability models. US Geological Survey Scientific Investigations Report 2004-5118.

Gurdak, J.J. and Qi, S.L. 2006. Vulnerability of recently recharged ground water in the High Plains Aquifer to nitrate contamination. US Geological Survey Scientific Investigations Report 2006-5050.

Helsel, D.R. & Hirsch, R.M. 1992. Statistical methods in water resources, Studies in environmental science-49. Elsevier Science Publishing Company Inc.: New York, NY.

Hitt, K.J. and Nolan, B.T. 2005. Nitrate in ground water: Using a model to simulate the probability of nitrate contamination of shallow ground water in the conterminous United States. US Geological Survey Scientific Investigations Map 2881.

Hosmer, D.W. and Lemeshow, S. 1989. Applied logistic regression. Wiley & Sons Inc: New York, NY.

Killingstad, M.W., Widdowson, M.A., and Smith, R.L. 2002. Modeling enhanced in situ denitrification in groundwater. *Journal of Environmental Engineering*, 128 (6): 491-504.

Knox, E. and Moody, D.W. 1991. Influence of hydrology, soil properties, and agricultural land use on nitrogen in groundwater. 19-57. In: Follett, R., Keeney, D., and Cruse, R. (eds.), Managing Nitrogen for Groundwater Quality and Farm Profitability. American Society of Agronomy, Madison, WI.

Kochanov, W.E. 1989. Sinkholes and karst-related features of Cumberland County, Pennsylvania: Pennsylvania Geological Survey, 4th ser., Open-File Report 89-02, scale 1:24,000, 20 maps plus 16-page text.

LaMotte, A.E. and Greene, E.A. 2007. Spatial analysis of land use and shallow groundwater vulnerability in the watershed adjacent to Assateague Island National Seashore, Maryland and Virginia, USA. *Environmental Geology*, 52, 1413-1421.

Lindsey, B.D., Falls, W.F., Ferrari, M.J., Zimmerman, T.M., Harned, D.A., Sadorf, E.M., and Chapman, M.J. 2006. Factors affecting occurrence and distribution of selected contaminants in ground water from selected areas in the Piedmont Aquifer System, eastern United States, 1993-2003. US Geological Survey Scientific Investigations Report 2006-5104.

Makuch, J. and Ward, J.R. n.d. Groundwater and agriculture in Pennsylvania. Penn State College of Agriculture, Cooperative Extension, Circular 341.

Menard, S.W. 2002. Applied Logistic Regression Analysis. Sage Publications Inc: Thousand Oaks, CA.

Natural Resources Conservation Service. 1986. Urban hydrology for small watersheds. US Department of Agriculture, Natural Resources Conservation Service Technical Release 55.

Natural Resources Conservation Service. 2004a. Soil Survey Geographic (SSURGO) database for Adams County, Pennsylvania: US Department of Agriculture, Natural Resources Conservation Service, digital data, accessed May 20, 2007 at URL *http://soildatamart.mrcs.usda.gov*.

Natural Resources Conservation Service. 2004b. Soil Survey Geographic (SSURGO) database for Cumberland County, Pennsylvania: US Department of Agriculture, Natural Resources Conservation Service, digital data, accessed May 20, 2007 at URL *http://soildatamart.mrcs.usda.gov*.

Natural Resources Conservation Service. 2004c. Soil Survey Geographic (SSURGO) database for York County, Pennsylvania: US Department of Agriculture, Natural Resources Conservation Service, digital data, accessed May 20, 2007 at URL *http://soildatamart.mrcs.usda.gov*.

Nolan, B.T. 2001. Relating nitrogen sources and aquifer susceptibility to nitrate in shallow ground water of the United States. *Ground Water*, 39 (2), 290-299.

Nolan, B.T., Hitt, K.J., and Ruddy, B.C. 2002. Probability of nitrate contamination of recently recharged groundwaters in the conterminous United States. *Environmental Science & Technology*, 36 (10), 2138-2145.

Ott, R.L. 1993. An introduction to statistical methods and data analysis. Fourth Edition. Duxbury Press: Belmont, CA.

Pennsylvania Department of Environmental Protection. 1999. Ambient and Fixed Station Network (FSN) Groundwater Monitoring Point Data (1985 – 1998). PADEP, digital data, accessed August 7, 2007 at URL *ftp://www.pasda.psu.edu/pub/pasda/drinkwater/gwndep99.zip*.

Pennsylvania Department of Environmental Protection. 2000a. Drought Information Center: Drought Status Map History. PADEP Website, accessed August 7, 2007 at URL *http://www.depweb.state.pa.us/watershedmgmt/cwp/view.asp?a=1435&q=527747*.

Pennsylvania Department of Environmental Protection. 2000b. Drought Information Center: Precipitation Historical Tables and Monthly Graphs. PADEP Website, accessed August 7, 2007 at URL *http://www.depweb.state.pa.us/watershedmgmt/cwp/view.asp?a=1435&q=530876*.

Pennsylvania Department of Environmental Protection. 2001a. Drought Information Center: 2001 Drought Archive. PADEP Website, accessed August 7, 2007 at URL *http://www.depweb.state.pa.us/watershedmgmt/cwp/view.asp?a=1435&q=525122*.

Pennsylvania Department of Environmental Protection. 2001b. Drought Information Center: Precipitation Historical Tables and Monthly Graphs. PADEP Website, accessed August 7, 2007 at URL *http://www.depweb.state.pa.us/watershedmgmt/cwp/view.asp?a=1435&q=530876*.

Pennsylvania Department of Environmental Protection. 2006. Act 537: Pennsylvania Sewage Facilities Act, With Index. PADEP Document 3800-BK-DEP1416, accessed August 7, 2007 at URL *http://www.depweb.state.pa.us/watersupply/cwp/view.asp?a=1260&Q=449298&watersupplyNav=|30160|#537*.

Pennsylvania Department of Transportation. 2007a. Pennsylvania County Boundaries. PennDOT, digital data, accessed May 20, 2007 at URL *ttp://www.dot.state.pa.us/*.

Pennsylvania Department of Transportation. 2007b. Pennsylvania State Roads: PennDOT, digital data, accessed May 20, 2007 at URL *ttp://www.dot.state.pa.us/*.

Pennsylvania Department of Transportation. 2007c. PennDOT – Pennsylvania Municipality Boundaries 2007: PennDOT, digital data, accessed May 20, 2007 at URL *http://cegis2.cas.psu.edu/uci/MetadataDisplay.aspx?entry=PASDA&file=pamunicipal2007.xml&dataset=41*.

Pennsylvania Geological Survey. 1998. Pennsylvania Physiographic Provinces, digital data, accessed May 20, 2007 at URL *http://www. dcnr.state.pa.us/topogeo/*.

Pennsylvania Geological Survey. 2001. Bedrock Geology of Pennsylvania: Pennsylvania Bureau of Topographic and Geologic Survey, Department of Conservation and Natural Resources, digital data, accessed May 20, 2007 at URL *http://www.dcnr.state.pa.us/topogeo/map1/bedmap.aspx*.

Reese, S.O. and Lee, J.J. 1998. Summary of Groundwater Quality Monitoring Data (1985 – 1997) from Pennsylvania's Ambient and Fixed Station Network (FSN) Monitoring Program. PADEP Bureau of Water Supply Management, accessed August 7, 2007 at URL *www.dep.state.pa.us/dep/deputate/watermgt/wc/subjects/srceprot/ground/sympos/ground_paper.doc*.

Risser, D.W. and Siwiec, S.F. 1996. Water-Quality assessment of the Lower Susquehanna River Basin, Pennsylvania and Maryland: Environmental setting. US Geological Survey Water-Resources Investigations Report 94-4245.

Root, S.I. 1968. Geology and mineral resources of southeastern Franklin County, Pennsylvania. Harrisburg: Commonwealth of Pennsylvania, Department of Internal Affairs, Topographic and Geologic Survey.

Ruddy, B.C., Lorenz, D.L., and Mueller, D.K. 2006. County-Level estimates of nutrient inputs to the land surface of the conterminous United States, 1982-2001. US Geological Survey Scientific Investigations Report 2006-5012, data table, accessed May 20, 2007 at *http://pubs.usgs.gov/sir/2006/5012/*.

Rupert, M.G. 2003. Probability of detecting atrazine/desethyl-atrazine and elevated concentrations of nitrate in ground water in Colorado. US Geological Survey Water-Resources Investigations Report 02-4269.

SAS Institute Inc. 1989. SAS/STAT user's guide. Version 6, Fourth Edition, Volume 1. SAS Institute Inc.: Cary, NC.

Sevon, W.D. 2001. Landscape evolution in the Cumberland Valley, Southeastern Pennsylvania, in Potter, Noel, ed., The Geomorphic Evolution of the Great Valley near Carlisle, Pennsylvania, Southeast Friends of the Pleistocene 2001 Annual Meeting Guidebook.

Shirk, W.R. 1980. A guide to the geology of southcentral Pennsylvania. Chambersburg: Robson & Kaye, Inc.

Shumway, R.H., Azari, A.S., and Johnson, P. 1989. Estimating mean concentrations under transformation for environmental data with detection limits. *Technometrics*, 31 (3), 347-356.

Smith, S.J. and Cassel, D.K. 1991. Estimating nitrate leaching in soil materials. 165-188. In: Follett, R., Keeney, D., and Cruse, R. (eds.), Managing Nitrogen for Groundwater Quality and Farm Profitability. American Society of Agronomy, Madison, WI.

South Middleton Township. n.d. Township information. South Middleton Township, accessed April 30, 2007 at URL *http://www.census.gov/population/www/cen2000/90vs00.html*.

South Middleton Township. 1999. Comprehensive plan. South Middleton Township, Cumberland County, Pennsylvania.

South Middleton Township. 2000. On-lot Septic Ordinance. South Middleton Township Ordinance No. 00-02, accessed August 7, 2007 at URL *http://www.smiddleton.com/ordinances/other.htm*.

South Middleton Township. 2001. SMT-1-ACT537 Well Totals: South Middleton Township, data table, obtained February 22, 2007 from South Middleton Township, Cumberland County, Pennsylvania.

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*, 35 (6), 1029-1039.

Thornbury, W.D. 1965. Regional geomorphology of the United States. John Wiley & Sons, Inc: New York, New York.

US Census Bureau. 2000a. American Factfinder Block-Level Population Data for South Middleton Township, Cumberland County, Pennsylvania: US Census Bureau, data tables, accessed May 20, 2007 at URL *http://factfinder.census.gov/servlet/DTGeoSearchByListServlet?ds_name=DEC_2000_SF1_U&_lang=en&_ts=208623945514*.

US Census Bureau. 2000b. Census 2000 TIGER/Line Data: US Census Bureau, digital data, accessed May 20, 2007 at URL *http://arcdata.esri.com:80/data/tiger2000/tiger_download.cfm*.

US Geological Survey. 1999a. National Elevation Dataset: US Geological Survey, digital data, accessed May 20, 2007 at URL *http://ned.usgs.gov/*.

US Geological Survey. 1999b. National Hydrography Dataset. US Geological Survey, digital data, accessed December 16, 2007 at URL *http://nhd.usgs.gov/*.

US Geological Survey. 2001. National Land Cover Database Zone 61 Land Cover Layer: US Geological Survey, digital data, accessed May 20, 2007 at URL *http://www.mrlc.gov/mrlc2k_nlcd.asp*.

US Geological Survey. 2004. USGS High Resolution Orthoimages. US Geological Survey, digital data, accessed May 20, 2007 at URL *http://www.pasda.psu.edu/*.

Ward, M.H., Mark, S.D., Cantor, K.P., Weisenburger, D.D., Correa-Villasenor, A., and Zahm, S.H., 1996, Drinking water nitrate and the risk of non-Hodgkins lymphoma: *Epidemiology*, 7, 465-471.

Way, J.H. 1986. Your guide to the geology of the Kings Gap area, Cumberland County, Pennsylvania. Harrisburg: Commonwealth of Pennsylvania, Department of Environmental Resources, Office of Resource Management, Bureau of Topographic and Geologic Survey.

Winter, T.C., Harvey, J.W., Franke, O.L., and Alley, W.M. 1998. Ground water and surface water: A single resource. US Geological Survey Circular 1139.

York County Planning Commission. 2000. Franklin Township Land Use: YCPC, digital data, obtained July 13, 2007 from York County Planning Commission, York County, Pennsylvania.

York County Planning Commission. 2003. York County Water Management: York County Comprehensive Plan Resource Report, accessed August 7, 2007 at URL *http://www.ycpc.org/Comprehensive_plan.htm*.