EXPLORING THE ENVIRONMENTAL COVARIATES OF *E. COLI* AND TOTAL COLIFORMS IN ONTARIO'S GROUNDWATER

by

TENNY RISETTE BACHE

B.A., The University of British Columbia, 2002.

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

In

The Faculty of Graduate Studies

(Geography)

THE UNIVERSITY OF BRITISH COLUMBIA

October, 2006

ABSTRACT

Microbial contamination of drinking water poses a danger to human health, and although the risk of exposure to waterborne microbes is highest in developing nations, outbreaks continue to affect populations in developed countries. Complacency toward groundwater protection can have tragic consequences – an example of which is the May 2000 outbreak of waterborne disease in Walkerton, Ontario, where seven people died and over 2100 became ill. In a response to such incidents, our research investigates the potential mechanisms of microbial contamination of Ontario's groundwater sources, which supply over 25% of the Province's twelve million residents with drinking water.

In this project, we identify environmental risk factors for private well contamination by coliform bacteria, specifically *Escherichia coli* (*E. coli*), in Southern Ontario. Inspired by concepts of landscape epidemiology, multiple methodologies were employed to assess the impact of local environmental characteristics on groundwater quality. A Geographic Information System (GIS) was used to integrate and analyze several datasets, including: land use, agricultural animal densities and farming practices, private well locations and corresponding water quality, human population densities, and geology. Through spatial and statistical analyses, we found that areas of agricultural land, low infiltration rate soil and surficial geology, and carbonate bedrock are significantly more prevalent near contaminated wells; whereas areas of developed land, soils with high infiltration rates, and non-carbonate bedrock are more prevalent near clean wells. The outcomes and methodologies identified in this project help further our understanding of the potential processes responsible for effective transfer of microbes from the environment and animals to humans.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AR	Antimicrobial Resistance
ARO	Antimicrobial Resistance in Organisms Research Project
CANSIS	Canadian Soil Information System
CAR	Census Agricultural Region
CCS	Census Consolidated Subdivision (Municipality)
CFU	Colony-forming Units
CIHR	Canadian Institute of Health Research
CART	Classification and Regression Trees
CLI	Canadian Land Inventory
DA	Dissemination Area
DEM	Digital Elevation Model
DFA	Discriminant Function Analysis
E. coli	*Escherichia coli*
ESRI	Environmental Systems Research Institute
GIS	Geographic Information Systems
GLM	Generalized Linear Models
ha	Hectares
Ma	Million years ago
MNR	Ministry of Natural Resources (Ontario)
MOE	Ministry of Environment (Ontario)
MOHLTC	Ministry of Health and Long Term Care in Ontario
MRA-GIS	Microbial Risk-Assessing Geo-ecological Information System
MAUP	Modifiable Areal Unit Problem
NO^3-N	Nitrogen (Nitrate)
OMAF	Ontario Ministry of Agriculture and Food
SPSS	Statistics Package for the Social Sciences
TC	Total Coliforms
UCLA	University of California
UNEP	United Nations Environment Program
US EPA	United States Environmental Protection Agency
VIF	Variance Inflation Factor
WHO	World Health Organization
WHPA	Wellhead Protection Area
WWIS	Water Well Information System

ACKNOWLEDGEMENTS

I would like to thank the following people for their support:

Dr. Michael Buzzelli, my supervisor, for providing invaluable direction and support of my learning process. Thank you making this possible.

Dr. Marie Louie and the ARO research team for providing me with the opportunity to participate in their research work; and Caroline Guenette from the Public Health Agency of Canada for geocoding the well database.

Dr. Christopher Smart from the University of Western Ontario, for insight into the hydrogeology of Ontario, and for directing me to a study by Mary Jane Conboy and Michael Goss that became influential to this thesis.

Dr. Brian Klinkenberg, my second reader, for providing excellent feedback and a welcoming work environment in my first year. Thanks also to my third committee member, Dr. Ray Copes.

Dr. David Ley, whose comments on an undergraduate paper led me to pursue graduate school in the first place; the forever-encouraging Dr. Graeme Wynn; and my mentor Sally Hermansen.

I want to thank my friends:

Ness and Kristin, for me your friendship is the most important result of the last two years. Unexpected and wonderful. Thank you for your relentless support and all of the *hooning* around.

Zoe and Steve, my beach house family, who always give me space to grow and the closeness to thrive. Te amo!

Cohen, our afternoon chats made me smile even when I didn't feel like it. Thank you for all of the encouragement, distractions, and friendship.

Many thanks to Jason and Mike (whose humour helped me get through school the first two times around); and Katie, Charis, and Jaz - my sisters.

And to my fabulous colleagues at "Geog. High," you made it a great place to be.

Last never least, my family. Thank you:

Mom, for your faith in me, for teaching me as a child, and for the treats and friendship. Thank you to all of the vivacious Tchakedjian women for the laughs and your love.

Dad and Sue, your warmth and support continuously carry me through challenges. Your encouragement and Sunday dinners have been so much appreciated.
Thank you from the bottom of my heart.

This thesis is dedicated to my inspiration,
My Grandma,

Risette Etmekdjian Tchakedjian

1. INTRODUCTION

1.1 Background

Historically, groundwater has been considered a more reliable source of clean water than surface water due to the attenuating effects of protective aquifers against microbial pollution (United Nations Environment Programme [UNEP], 2003). It is argued that efforts to protect groundwater sources have been halfhearted for this reason (De Loe and Kreutziser, 2005), and as human development and agricultural production intensifies worldwide, shortcomings in knowledge and groundwater policy management pose a significant risk to population health. Complacency toward the protection of groundwater sources has ended in tragic results – a prime example being a recent outbreak of waterborne disease in Walkerton, Ontario. In May 2000, seven people died and more than 2100 people became ill with gastroenteritis when the municipal water supply in Walkerton was contaminated by *Escherichia coli (E. coli)* O157:H7 and *Campylobacter jejuni*. In the Inquiry report to the Ontario Attorney General that ensued (O'Connor, 2002a, 2002b), investigators attributed operator error as just one cause of many that contributed to the outbreak, citing lack of knowledge of the processes by which pathogens contaminate groundwater, and gaps in water protection policy as the foundations for the system failure (De Loe et al., 2005; Holme, 2003; Hrudey, Payment, Huck, et al., 2003). As a result, many research projects were launched to investigate the mechanisms of microbial contamination of Ontario's groundwater, which supplies over 25% of the Province's 11.5 million residents with drinking water.

The Canadian Institute of Health Research's (CIHR) Safe Food and Water Initiative has funded a research project to examine patterns of Antimicrobial Resistance (AR) in waterborne pathogens found in Alberta, Quebec, and Ontario. Led by Dr. Marie Louie of the University of Calgary, the project is entitled 'Prospective multi-Province surveillance for antimicrobial-resistant *E. coli* in drinking and recreational source waters: Impact on humans and the environment.' There are four research components of the project, detailed in Table 1.1.

Table 1.1: ARO Study Group Overview

Group 1	**Antimicrobial Surveillance**
Aim:	To determine the prevalence of AR in *E. coli* isolated from private drinking water sources in Alberta and Ontario, and recreational/beach waters in Alberta, Ontario, and Quebec
Group 2	**Case-control Study**
Aim:	To determine risk factors for well-water contamination with AR *E. coli*
Group 3	**Molecular Characterization**
Aim:	To characterize the resistance determinants in selected strains of AR *E. coli,* and to compare their resistance genotype(s)
Group 4	**Spatial Analyses**
Aim:	To characterize the spatial distribution of *E. coli* and AR *E. coli,* and derive models that predict the presence of AR *E. coli* strains based on land use and population attributes.

Dr. Michael Buzzelli of Queens University, previously of the University of British Columbia (UBC), is an investigator in the Spatial Analyses group. The author of this thesis is a graduate student at UBC working with Dr. Buzzelli on the research questions posed to Group 4, specifically investigating the impact of local environmental conditions on a well's microbial water quality. The second team in Group 4, based out of the Laboratory for Foodborne Zoonoses in St-Hyacinthe, Quebec, focuses on the effects of regional (rather than local) processes on AR in *E. coli.*

Research on local and regional effects of non-point pollution on private drinking water sources is especially important today as agricultural intensification coincides with the growth of rural communities in Southern Ontario. Although agricultural production is increasing in Ontario, the total number of farms in the Province decreased by 13% from 1996 to 2001. In the same time period, the total acreage of farms decreased 2.8% - however the average size of a farm increased 8.8% (see Appendix 1 for detailed totals). These trends signal a general shift from small family-run farms to high production operations, with large numbers of livestock[1] (Miller, 2000). Animal densification and

[1] While not all farms have large numbers of livestock, there has been a general trend to livestock intensification. Although numbers of cattle have decreased 6.8% from 1991-2001, there has been an 18.1% increase in the number of pigs in the Province, an 18.4% increase in sheep and lambs, and a 31.6% increase in hens and chickens.

intensified agricultural production leads to increases in nutrient (manure) storage and application. Juxtaposed against rural population growth and increased pressure on shared groundwater resources, the potential mismanagement of agricultural byproducts poses a significant risk to public health.

In Ontario, "dependence on groundwater is highest in rural areas, where the population is served predominantly by private wells" (De Loe et al., 2005, p. 245), and unlike government-maintained municipal water supplies, the water quality of a private well is the sole responsibility of the well owner. However, a well owner does not have the right or capability to monitor agricultural activities on a nearby property that may affect the microbial quality of their well – these responsibilities lie with various levels of government. Questions regarding groundwater protection have been raised by concerned citizens, environmental groups, and experts alike, calling for research and environmental policies to be updated to match new intensified forms of agricultural production (O'Conner, 2002b; Miller, 2000; Bocking, 2002). Fueling these worries is the fact that waterborne disease continues to place a burden on population health and healthcare systems across the nation. The Canadian Water Research Institute reported in 2003 that exposure to waterborne infections causes 90,000 cases of illness (approximately 1400 of which are due to verotoxic *E. coli* (Charron, Thomas, Waltner-Toews, et al., 2004)), and 90 fatalities across Canada every year (Edge, Byrne, Johnson, et al., 2003).

1.2 Objectives

The purpose of this research is to contribute to the emerging (and entwined) fields of spatial epidemiology, GIS, and groundwater studies in a Canadian context. There exists gaps in our knowledge concerning the environmental and anthropogenic processes responsible for the fecal contamination of groundwater, and we are interested in the potential of spatial analysis to address these types of public-health related questions. There is minimal research into the effect of local landscape characteristics on the microbial water quality of private wells in Southern Ontario, and we aim to expand this

knowledge base through the exploration of multiple spatial and statistical methodologies. Our specific objectives are to:

1) Compile a spatial database of environmental data pertaining to microbial water quality in Southern Ontario
2) Explore methods for measuring land characteristics surrounding private wells using GIS
3) Identify environmental risk factors for private well contamination on the basis of these local land characteristics
4) Derive inferential models to predict the presence of *E. coli* and Total Coliforms (TC) in private wells based on land use and population attributes
5) Further our understanding of the potential processes responsible for effective transfer of microbes from the environment and animals to humans

The outcomes measured from the models and related statistical analyses will elucidate the environmental characteristics most closely associated with *E. coli*, which will be an important contribution for guiding the selection of case-wells and control-wells in subsequent analyses in the larger ARO project.

This thesis is comprised of five chapters. Following the introduction, in chapter two I review current research in spatial epidemiology and waterborne disease. The microbiology of *E. coli* is discussed, in particular the environmental conditions required for bacterial survival and transport (pertaining to the study area). The chapter closes with an outline of the hypotheses, where we argue that certain environmental characteristics are associated with the presence of *E. coli* and TC in the private wells sampled in Southern Ontario during the 2003 and 2004 summer seasons. On the basis of the literature review and research questions, data are singled out for inclusion to the GIS for further analyses.

In chapter three, I describe the origin and quality of all spatial data entered into the GIS, and the methodologies employed to clean and resolve the information. These layers include well water samples, environmental data, and census data. The last section

describes the GIS-specific buffering and overlay techniques applied to measure the landscape characteristics adjacent to sampled wells.

In chapter four, I report the various statistical methods applied to these data, and the results. First, a comparison of multiple circular buffer sizes reveals that one radius is of greater importance for measuring the impact of land local to private wells on water quality. Secondly, the total areas of land selected by the well buffer zones are analyzed with both descriptive and inferential statistics, to determine if certain land characteristics are significantly more prevalent near contaminated wells versus clean wells.

In chapter five I discuss the results reported in chapter four, and the implications of these findings to the hypotheses, and other current studies. A section of this chapter includes ideas for possible improvements of the analyses for future research. These five chapters come together to form an investigation into geo-statistical methods in health research, and as a whole work to contribute to the emerging science of landscape epidemiology as applied to groundwater quality.

2. LITERATURE REVIEW

"The landscape that distinguishes a place is a complex expression of physical, biotic, and cultural process. When one knows how to analyze its elements and patterns, one can usually determine what diseases can occur" (Meade, Florin, and Gesler, 1988, p. 59).

2.1 Health Geography and Spatial Epidemiology

Health Geography refers to the study of the spatial dimensions of health and disease (Andrews, 2002). The discipline emerged over three decades ago as a quantitative scientific analysis of the spatial distributions of disease and health care provision, and was more commonly referred to as 'medical geography.' In the 1980s a theoretical shift away from quantitative methods and toward critiques of structuralism and humanist thought occurred in the discipline of human geography. These new 'health geographies' sought to integrate the study of health within broader social, cultural, political concerns (Jones and Moon, 1987; Kearns and Gesler, 1998). Most recently, there has been a convergence of the twin streams of 'medical' and 'health' geographies (Andrews, 2002), to form a more comprehensive analysis of the geography of health and wellbeing (Rosenberg, 1998). It has been established that multilevel perspectives, facilitated by technologies such as GIS, give new recognition to the complexity of hierarchy and variation in data and space (Kearns and Moon, 2002).

This is a study in health geography, as we apply methodologies of spatial analysis to examine the covariation in space of disease occurrence and related environmental factors (Meade and Earickson, 2001). This is also an epidemiological inquiry, as we are investigating patterns in disease presence that may give important clues about the behaviour of *E. coli* in the natural environment. Pathogens are living organisms that require a certain set of environmental parameters for survival and for their contamination of subsurface aquifers. By identifying the intersection of these parameters on a map (in a GIS), researchers can predict geographic areas where the risk of microbial contamination of groundwater might potentially occur. This practice is known as spatial epidemiology – a field "concerned with describing and understanding the small-area variations in disease

risk as inferred from investigations of geographically referenced health and population data" (Raper, 2004, p. 627). Spatial epidemiology can be broken down into three main areas:

1. **Disease mapping**: Individuals with a health outcome (disease) are represented as points on a map, or cases are aggregated to a spatial unit to form a chloropleth map. Visualizing disease outcomes juxtaposed with other data can guide map readers to identify disease 'hot spots' or associations with other data layers. Dr. John Snow applied this method to study an 1849 cholera outbreak in London. By mapping disease outcomes in relation to well locations, Snow was able to identify the cause of the outbreak: contaminated water from a specific well (Bingham, 2004).

2. **Disease clustering** (outcome surveillance): Spatio-temporal patterns in disease events identified by geo-statistical methods (i.e. cluster analysis) can give important clues about the etiology (cause and behaviour) of the studied disease. Most recently, cluster analysis has been put to use to analyze trends in cases of illness due to *E. coli* O157:H7 by postal code district in Scotland. The study found a significant trend increasing from west to east, as well as seasonal trends in outbreak patterns (Innocent, Mellor, McEwen, et al., 2005).

3. **Geographic correlation studies** (exposure surveillance): This newest field in spatial epidemiology aims to measure and determine the factors (environmental, social, political, etc.) that put certain parts of a population at risk for exposure to a disease, or disease-causing substances. Nygard, Andersson, Rottengen, et al. (2004) investigated associations between *Camploylobacter* incidence and municipal environmental characteristics (land use, livestock densities, water supply infrastructure), and found that municipalities with higher animal densities and longer water pipe lengths were at an increased risk of exposure to *Campylobacter* than other regions.

There is consensus among spatial epidemiologists that geographic correlation studies are rare and understudied compared to the first two study designs (Graham, Atkinson, and Danson, 2004; Elliott and Wartenberg, 2004; Jacquez, 2000). Although studies in landscape epidemiology first emerged over 50 years ago (Pavlovsky, 1966), research has been slow to advance. In their review of the field, Ostfeld, Glass, and Keeling (2005) argue "although the spatial dynamics of infectious diseases are the subject of intensive study, the impacts of landscape structure on epidemiological processes have so far been neglected" (p. 328).

The use of GIS to investigate the presence of waterborne microbes in relation to land use, human patterns, and environmental factors is also an emerging field. Prior to 2000, "surprisingly, GIS have only sparsely been applied to describe and analyze health risks due to microbial hazards" (Kistemann, Dagendorf, and Exner, 2001, p. 226). The epigraph by Meade et al. (1988) raises an interesting point – the authors mention that 'when one knows *how* to analyze' the complex elements and patterns of the environment, one can predict where a specific disease might occur. However, there is no methodological standard for *how* these investigations should be conducted, especially pertaining to environmental processes and waterborne disease. Every study varies widely in data content, scale, and geo-statistical methodologies. To my knowledge, no literature has been published on the effects of land characteristics local to private wells (measured by GIS methods) on the presence of *E. coli* or TC in private well water. A comprehensive body of knowledge does, however, exist on the local environmental effects on nitrate (NO^3-N) concentrations in both ground and surface waters. Synonymous to fecal coliforms, the major source of nitrates in groundwater is agriculture (manure is rich in nitrates).

In the extant literature, GIS buffering and overlay techniques have been widely used for over a decade to measure land characteristics surrounding water sources for comparison to nitrate levels (Barringer, Dunn, Battaglin, et al., 1990; Eckhardt and Stackelburg, 1995; Kolpin, 1997; Sliva and Williams, 2001; Lee, Min, Woo, et al., 2003; Wang, Liu, Wu, et al., 2006). Using similar techniques to investigate fecal coliform levels is not as widespread, despite the fact that "contamination of rural drinking water by

bacteria is commonly more prevalent than contamination due to excessive nitrate or pesticides" (Conboy and Goss, 2001, p. 101). Kistemann et al. (2001) used buffer zones to measure land uses near surface water tributaries to compare to the bacterial loading of streams and rivers – see Table 2.1 for details on these and other relevant studies. However, the majority of other research employing GIS techniques to study fecal coliforms examine patterns of disease outcomes (Dagendorf, Herbst, Reintjes, et al., 2002; Nygard et al., 2004; Innocent et al., 2005), instead of the presence of environmental *E. coli* in source waters. Further, most studies focus on smaller scale regions (drainage basins, municipalities, postal code districts) rather than land characteristics immediate to the source.

One paper of particular relevance to our work is entitled 'Statistical models for the assessment of NO^3-N contamination in urban groundwater using GIS' (Lee et al., 2003). In this study, environmental characteristics surrounding sampled wells are measured using circular buffers with radii ranging from 50-400m. These quantities are compared with nitrate concentrations using various descriptive and inferential statistical techniques (Mann-Whitney tests, correlation analyses, linear regression). The study found that 200m and 250m buffers are the most effective radii for measuring land characteristics. The results also show that nitrate concentrations in groundwater are associated with cropped land and mixed residential and business areas. Although their study focuses on chemical concentrations in groundwater, it serves as a useful guide for our investigation.

A 2001 paper by Thomas Kistemann et al. provides the most pertinent information on using GIS as a tool to study the microbial contamination of drinking water sources. The researchers recommend a microbial risk-assessing geo-ecological information system (MRA-GIS). This GIS includes a database built to incorporate a host of data relevant to microbial water quality: geology, soil, topography, vegetation, precipitation, human land use patterns (settlement, traffic, agriculture, forest, and industry), watercourses, farming practices, and unique features (drainage systems, cattle tracks, sewage plants, and private sewer systems). Kistemann et al. use the MRA-GIS for spatial and statistical analyses to provide a geo-ecological portrait of each watershed by

Table 2.1: Literature Review Summary

Authors/ Location	Metric of interest	Study Design	Sample Size, Time frame	Geo-statistical Methods	Landscape characteristics considered	Scale	Major findings
Wang et al., 2006. North China	NO^3, N concentration of groundwater	Geographic correlation	616 samples taken from shallow irrigation wells	Buffer zones (200m-2000m), Back propagation neural networks.	crop yields, nitrogen inputs, groundwater depth, soil organic matter content, soil sand content.	sample sites (buffer zones)	Vegetable cropping systems held high nitrogen surpluses. Nitrogen budget combined with GIS-based neural networks are effective in predicting nitrogen.
Innocent et al., 2005. Scotland	outbreaks of illness caused by *E. coli* O157:H7	Disease clustering, geographic correlation	All incidences of illness caused by *E. coli* (4-10 cases per 100,000 per year 1996-1999).	Choropleth mapping, Poisson model, Moran's I, spatial scan statistic.	cattle density, human density, number of cattle per person, lat/long, urban/rural landuse	Postal code districts	Case rate increases West to East, and South to North. Cattle density, human pop density, and # of cattle per person were significant.
Nygard et al., 2004. Sweden	*Campylobacter jejuni* infections among human populations	Geographic correlation	7007 cases of *Campylobacter* infections, 1998-2000	Data overlay/joining in GIS, Multivariate Poisson regression	cattle, swine, and poultry densities; population receiving water from public water supply, water pipe length	Munici-palities	Positive associations with water-pipe length, ruminant density, and a negative association with % of the population receiving water from a public supply.
Lee et al., 2003. Korea	*NO3, N concentration of groundwater*	Geographic correlation	1988 Well samples	Buffer zones (50-400m) surrounding wells, overlay, regression	Well depth, geology, precipitation, land uses, surface waters	sample sites (well buffer zones)	NO is associated with crop lands, mixed residential & business areas. Favourable results returned with 200m and 250m buffers.
Kistemann et al., 2001. Germany.	*Cryptospridium* and *Giardia lamblia* loads in surface water	Geographic correlation	70 surface water samples	Areas measured by overlay, Landuse % in basins vs. % in 50m river buffers	landuse, population density, cattle breeding, settlement areas, sewerage systems, surface water hydrology, weather patterns, soil	watershed, and local effects on tributaries (50m buffer)	*Cryptospridium* is associated with agriculture, and *Giardia lamblia* is related to discharge of human wastewater.
Sliva and Williams, 2001. Southern Ontario	Chemicals in surface waters in 3 Toronto watersheds, 1990-1993.	Geographic correlation	Averaged results from 12 surface water sampling stations across the 3 watersheds.	catchement measured by overlay in GIS vs. areas measured by 100m buffer of sampling station.	Surface elevation, slope, geology, landuse.	100m buffer zone, drainage basins.	Trend in increased chemical fluxes with increasing urban land use intensity. Catchment processes reflect water quality to a greater degree than 100m buffer zones.

buffering surface water bodies, calculating land classification areas within the buffer, and using descriptive statistics to investigate possible associations. We borrow from these concepts in our research and create a similar spatial database as a platform for studying private wells in Southern Ontario.

Although the above-mentioned studies provide guidance on general study design issues, it is clear that we are exploring new research territory. Our sample sizes (in the hundreds of thousands) are considerably larger than any other study published on the topic. Further, none of the examples found in current literature are specific to our study region, or use GIS methodologies to explore the effects of the local environment on *E. coli* in private wells. Despite the lack of a strong guiding tradition, an exciting opportunity exists to explore new methodologies and ideas. As the literature reveals, when working with extensive datasets in a GIS environment the possibilities for analyses are endless, and the potential for knowledge contribution is great. The first challenge comes in selecting data layers to collect and analyze in the GIS. To address this question, we consider the bio-ecology of *E. coli*: the relation between cell biology and the environmental processes that together are responsible for the presence of fecal coliforms in groundwater aquifers.

2.2 Waterborne E. coli and Total Coliforms

E. coli are single-celled organisms found in the intestinal systems of most mammals, and are a normal component of the human digestive system. Most *E. coli* strains are not harmful to our heath; in fact, their competitive presence prevents mammalian gastrointestinal tracts from being overrun by harmful bacteria and fungi. *E. coli* also plays a large role in synthesizing K and B-complex vitamins for absorption by the body. Despite the active role of *E. coli* in maintaining gastrointestinal health, the bacterium is probably best-known for its pathenogenic properties.

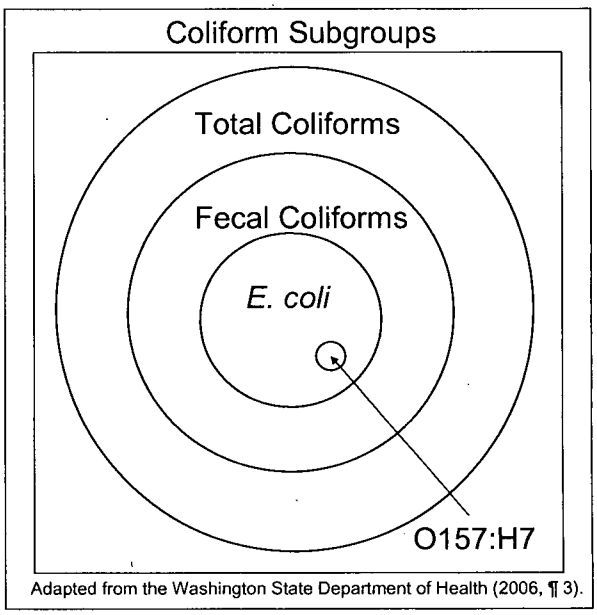
Five classes of *E. coli* are not normally found in the digestive system and cause intestinal disease in humans: enterotoxigenic, enteroinvasive, enterohemorrhagic,

enteropathogenic, and enteroaggregative *E. coli*. Each of these classes displays distinct features in pathogenesis, ranging from urinary tract infections, cramps, nausea, diarrhea, and fever, to more severe and chronic outcomes. One serotype of enterohemorrhagic *E. coli* (O157:H7) carries a toxin that when ingested by humans can cause bloody diarrhea, infant mortality, severe renal or neurological complications, long-term kidney failure, and death (Berg, 2003; Postgate, 2000). It is estimated that *E. coli* O157:H7 has an infectious dose of less than 100 organisms (Griffin and Tauxe, 1991), therefore, even a small presence of the coliform in drinking water is a serious concern. Heath Canada drinking water guidelines state that

> "no sample should contain *Escherichia coli*. *E. coli* indicates recent faecal contamination and the possible presence of enteric pathogens that may adversely affect human health. If *E. coli* is confirmed, the appropriate agencies should be notified, a boil water advisory should be issued, and corrective actions taken" (Health Canada, 2004, p. 3).

In a laboratory, it is costly and time-consuming to differentiate *E. coli* from other coliforms, let alone identify specific strains of *E. coli* that may or may not be harmful to human health. Therefore, any strain of *E. coli* found in drinking water is considered dangerous, and for ease of testing most private and governmental drinking water testing agencies do not isolate *E. coli*; rather they identify whether or not TC bacteria is present in a sample. TCs refer to the larger family of rod-shaped bacteria of which Fecal Coliforms (coliforms originating from the gastrointestinal tracts of mammals) and *E. coli* are subgroups (see Fig. 2.1). TCs are pervasive in the natural environment, and although not usually harmful to human health (Raina, Pollari, Teare, et al. 1999), TCs are used as an indicator to signal the potential presence of *E. coli* in a water sample.

Figure 2.1: Coliform Subgroups



Adapted from the Washington State Department of Health (2006, ¶ 3).

As part of the fecal coliform subgroup, *E. coli* thrive in the warm,

nutrient-rich atmosphere of a mammal's intestines and are not as pervasive in the natural environment as TCs. However, *E. coli* have been recorded as surviving for up to 110 days in the open environment (Na, Miyanaga, Unno, et al., 2006; Chalmers, Aird, and Bolton, 2000). A major source of environmental *E. coli* is the fecal waste of livestock (Altherholt, Feerst, Hovendon, et al., 2003), deposited by grazing animals or as a nutrient to crop beds (manure). When microbes come into contact with water through precipitation, or by being deposited into surface waters, water acts as a transport medium and can percolate through the subsurface carrying pathogens, eventually contaminating drinking water aquifers. As Schaffter, Zumstein, and Parriaux (2004) describe, once *E. coli* has been released into the environment,

> "the survival capacity of enteric bacteria depends on their physiological characteristics and on the properties of their *immediate* [emphasis added] environment (availability of nutrients and energy, light, temperature, pH, the presence of other organisms: predators and/or antagonists). In groundwater, the natural elimination of microorganisms is accentuated by other processes like adsorption, dispersion, and filtration" (Schaffter et al., 2004, p. 226).

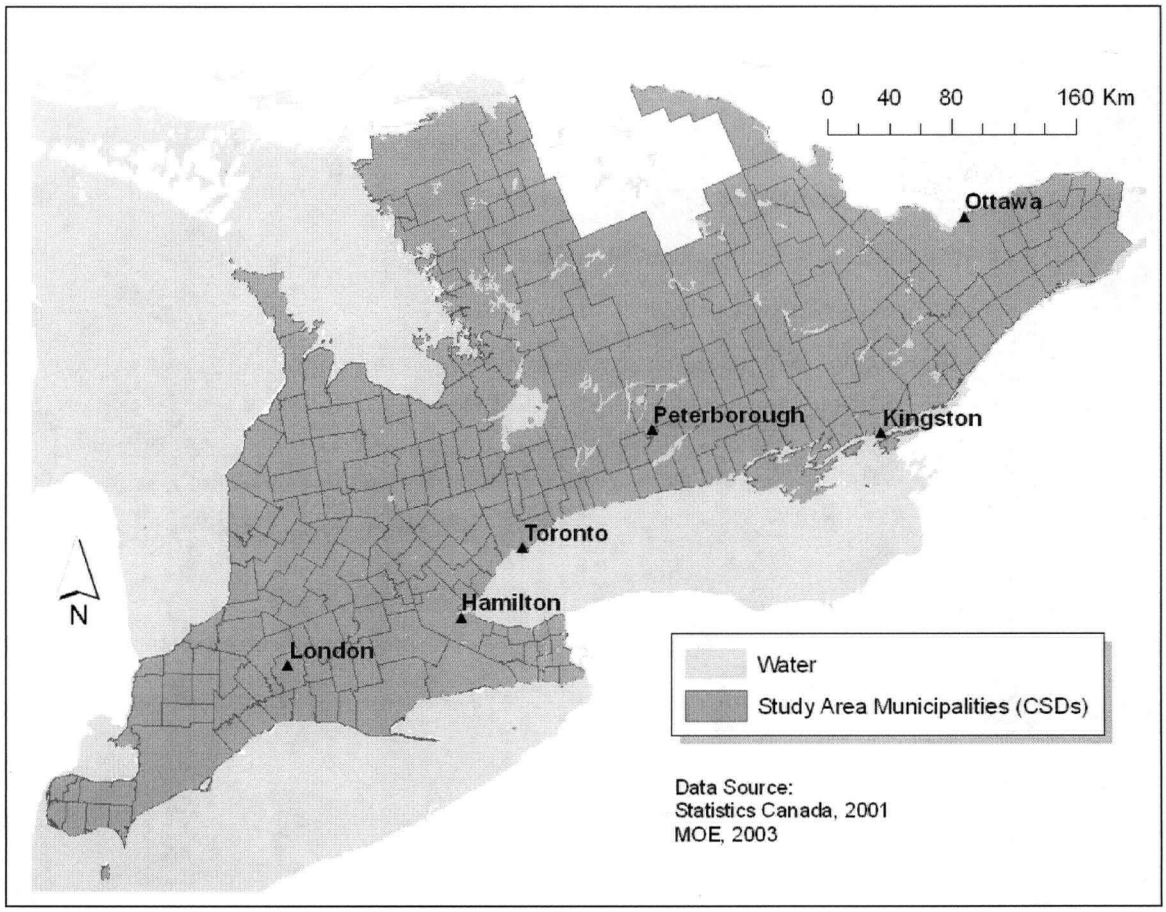
Naturally, then, to investigate the prevalence of *E. coli* in well water, it is essential to understand the environmental characteristics *local* to groundwater sources. In the following section, I provide an overview of the hydrogeology, land characteristics, and agricultural activities of Southern Ontario. Each of these elements is important to the presence and survival of environmental *E. coli*, and will be added to our MRA-GIS for closer inspection.

2.3 Study Area

The study area of the ARO project is located in the southern part of the Province of Ontario, Canada. The boundaries of the study area coincide with those of four Ontario Regional Health Authority Districts: Southern, Western, Central, and Eastern Ontario, comprising 15% of the Province's total landmass. Within these four districts lie 224 municipalities (depicted in Fig. 2.2), home to over eleven million people and 94% of the total provincial population. Over time, settlement patterns have been influenced by the

availability of groundwater supplies, and today the rural population of Southern Ontario relies almost entirely on groundwater resources (Conboy and Goss, 2000).

Figure 2.2: Study Area



Southern Ontario has a humid continental climate, characterized by cold winters and hot humid summers. The region receives approximately one meter of precipitation distributed evenly throughout the year (Government of Ontario, 2006), a sufficient amount to recharge groundwater aquifers for private use. Because the climate is moist, the presence of groundwater in Southern Ontario depends primarily on hydrogeologic characteristics. In the study area two principal geologic materials can be tapped for water supply: fractured bedrock, and sand/gravel overburden deposits (Novokowski, Beatty, Conboy, et al., 2006). Because water movement in bedrock is restricted to fractures, fissures, folds, joints, and eroded channels that may disrupt the formation, the porous overburden deposits generally provide higher-yielding wells.

Three major bedrock formations exist in the study region: the Canadian Shield in the north, the Great Lakes Lowlands in the southwest, and the Ottawa-St. Lawrence Lowlands in the east. The Canadian Shield is the oldest formation, consisting of billion-year old (Precambrian) granite, gneiss, metasedimentary and metavolcanic rocks. This region is characterized by exposed bedrock and thin soils. As a result, hydrological drainage is poor, and few high-capacity wells exist on the Shield. The most densely populated regions in the Province are located in the Great Lake Lowlands, which consist largely of Paleozoic sedimentary rock. The Niagara Escarpment is a prominent geologic feature of the region, dividing the lowlands into two physiographic parts. East of the escarpment, the rocks are composed of shale and limestone, and west of the escarpment, some of Ontario's most productive bedrock aquifers lie in numerous limestone and dolostone formations. Lastly, the Ottawa-St. Lawrence Lowlands are characterized by newer Paleozoic sedimentary rocks, mainly sandstone, limestone, and dolostone.

These three bedrock formations are overlain by highly variable sediment deposits, characterized by the glacial formations of the Pleistocene Epoch (Novokowski et al., 2006). During this Ice Age, ice sheets deposited a wide variety of water-bearing sediments (called overburden), across the study area. Glacial landforms such as drumlins, eskers, and moraines are often composed of sands and gravels. These surficial materials are permeable in nature, have complex soil structures, and typically yield productive aquifers (Regional Municipality of Waterloo, 2003). Groundwater availability combined with deep fertile soils render overburden deposits attractive locations for human settlement and agriculture. The majority of high-quality agricultural land in the Province is located in Southern Ontario, and as a result the study area has an active and diverse agricultural sector, generating over seven billion dollars annually.

A wide range of fruit, vegetables, and grains are grown; and the region supports a thriving livestock industry, especially in dairy, cattle, and hog farming. Multiple hydrological drainage basins in Southern Ontario have consistently been placed in the top 5 basins for manure production in Canada, with farms located in these basins producing between 2,000 and 5,000 kilograms per hectare of land yearly, when the national average

is 775 kg/ha (Statistics Canada, 1996, ¶2; Statistics Canada, 2001, Livestock Manure Production, ¶2). In rural regions, intensification of agricultural production is occurring alongside population growth and an increased dependence on groundwater, signaling that an investigation into processes responsible for microbial contamination of aquifers is relevant, especially in Southern Ontario.

2.4 Hypotheses

A goal of this research is to identify local environmental risk factors for private well contamination by waterborne fecal coliforms[2]. Based on a survey of current literature, we assume that the prevalence of *E. coli* in private well water is associated with environmental characteristics local to wells, specifically:

- Agricultural land use
- Agricultural animal densities (Cattle, sheep, poultry, hens)
- Low human population densities (rural areas)
- Soils with low water infiltration rates
- Surficial geology with low permeability
- Bedrock composed of carbonate materials
- Ancient (mesoproterozoic) bedrock
- Non-moraine material
- Selected agricultural practices (manure application, lack of tilled soil, irrigation)

The rationale behind these assumptions will be discussed at length in the remainder of this section. To investigate the specific relations between the above-listed environmental characteristics and the presence of *E. coli* and TC in well water, we will:

- Compile relevant environmental and population data in a spatial database
- Use a GIS to measure land characteristics local to wells with circular buffers

[2] The results are also applicable to public wells, but because our samples have all been obtained from private wells (which are generally smaller in capacity, and therefore have a smaller zone of influence footprint on the surface than public or commercial wells), in this thesis I will exclusively discuss the relevance of the results to private wells.

- Apply statistical tests to assess the degree of association between land characteristics and well contamination

Specifically, we expect to see a higher prevalence of *E. coli* in wells located near agricultural areas, as activities typical to agriculture (animal husbandry, manure storage, nutrient application to crops) are major sources of environmental fecal coliforms. Extensive research shows that agricultural land use and livestock density are associated with groundwater coliform concentrations (Goss, Barry, and Rudolph, 1998; Crowther, Kay, and Wyer, 2002; Kistemann et al., 2001; Nygard et al., 2004; George, Anzil, and Servais, 2004). Agriculture typically exists in rural regions with lower population densities; therefore we hypothesize that a higher prevalence of contaminated wells exists in areas with low human population densities. Previous studies have shown a clear rural/urban gradient in well contamination (Innocent et al., 2005; Haack, Jelacic, Besser, et al., 2003).

Once *E. coli* has been deposited on the surface, microbes must travel through and survive in the subsurface in order to contaminate an aquifer, and hydrogeologic characteristics have a significant impact on these processes. We hypothesize that wells in Southern Ontario will be at increased risk for contamination by *E. coli* if they are surrounded by soils with low water infiltration rates; and surficial geology[3] with low water permeability. The reason lies in filtration: soils act as a natural filter for bacteria. As water percolates down through pores in soil, bacteria can adsorb to sediment surfaces and may be removed from the infiltrating water. The larger sediment surface area that the bacteria come into contact with, the more likely it is that bacteria will be removed from the percolating water. Natural filtration can be extremely effective at protecting groundwater sources. Crane, Westerman, and Overcash (1980) irrigated bacteria onto a soil surface, and found that 92% to 97% of bacteria were removed by the top 1cm of soil. Well-drained, sandy soils allow for the unhindered infiltration of water through the subsurface, and have been shown to restrict bacteria and provide a protective

[3] The term 'surficial geology' refers to all geologic materials from the surface to the bedrock layer; and the term 'soils' refers to only the top 1 meter of geologic materials. This difference will be discussed to greater extent in the data section 3.2.2 and 3.2.3.

environment for wells (Crane, Moore, Grismer, et al., 1983). Contrary to this, shallow soils or exposed bedrock offer little or no filtration. Non-permeable materials such as clay have finer grain sizes than sandy soils, are less porous, and have lower infiltration rates. Clay is a hydrophilic material, meaning that when it comes into contact with water it absorbs water and expands. When the water evaporates, the clay particles shrink and fractures can form in drying soil formations. In summer months, soils with low infiltration rates are often inherently fractured, and water is able to travel quickly through preferential flow paths into the subsurface without filtration. Further, clay soils are often high in organic content, providing bacteria with a favourable nutrient-rich environment.

The works of Ontario-based researchers Conboy and Goss (2000, 2001) focus on the impact of Ontario-specific hydrogeology on microbial water quality. They describe a survey of rural groundwater quality in Ontario (2000, 2001), the results of which are instrumental to our hypotheses. Among other findings, they established that

> "high risk wells in Ontario were located most often at sites with older limestone or dolostone bedrock, and in clay or clay loam soil. The presence of a sandy soil may offer some protection to groundwater resources in very vulnerable dug or bored wells" (2000, p. 1).

We hypothesize that wells surrounded by ancient (Mesoproterozoic) bedrock are at higher risk for contamination by microbes. Over time, this bedrock is subject to greater erosion than the newer Neoproterozoic and Paleozoic materials, and therefore preferential flow paths are more likely to exist in the formations.

Similarly, we expect to find a high prevalence of carbonate bedrock surrounding contaminated wells. Carbonate materials such as limestone, calcite, and dolomite are partially soluble; therefore areas of carbonate bedrock weather readily, especially in acidic conditions (i.e. acid rain). Erosion and weathering leads to the development of preferential flow paths through the subsurface, rendering aquifers in carbonate bedrock more vulnerable to contamination from surficial sources of pollution (Allen and Morrison, 1973). Conboy and Goss note that in Ontario, the "fracturing, dissolution and especially karstification of limestone appeared to result in higher potential movement of bacterial contaminants through limestone rock than in any other geological formation" (2000, p.3). To summarize, we hypothesize that wells are at *higher* risk of contamination

if they lie in areas of permeable bedrock, but at a *lower* risk of contamination if they lie in areas of permeable soils and surficial geologic materials.

As described in section 2.3, a major source of groundwater in Ontario is found in the permeable surficial aquifers formed by glacial deposits, such as moraines. Moraines are mounds of unconsolidated sands and gravels, composed of a wide variety of sediment sizes once carried and deposited by retreating glaciers. Some of Southern Ontario's most productive sand and gravel aquifers are located on moraines (Novokowski, Beatty, Conboy, et al., 2006). The permeable sand and gravels of the deposits

> "act like a sponge, and absorb precipitation. This precipitation is then stored in layers of aquifers, filtered, and slowly released as cool fresh water. Hence, a moraine can provide drinking water and act as a recharge/discharge area - sustaining the health of many watersheds and communities" (Regional Municipality of Waterloo, 2003, p. 6).

In accordance, we postulate that wells located on moraine deposits are less susceptible to contamination by *E. coli* than wells not sited on moraines.

Lastly, we hypothesize that a number of anthropogenic agricultural processes contribute to the presence of *E. coli* in groundwater sources, although the data associated with obtaining information on farming practices is on a municipal scale, not a local one. We expect to see an association between *E. coli* presence and the proportion of land in a municipality that manure nutrients have been applied to. This is not only because manure is a primary source of environmental *E. coli*, but also because the presence of nitrogen facilitates the survival of *E. coli* in competition with indigenous bacterial flora in an aqueous environment (Lim and Flint, 1989; Rosen, 2000). We also expect to see an association between *E. coli* presence and the proportion of untilled cropland in a municipality. The practice of leaving soils untilled is becoming more frequent as farmers make efforts to reduce the dilapidating effects of erosion. Tillage disrupts the surface and causes water to percolate more evenly through soils, breaking down preferential flow paths that may have developed over time. The result is increased filtration of water through the surface layers, and a more protected aquifer (Gaglardi and Karns, 2000; Conboy et al., 2000). Finally, because *E. coli* survives in aqueous environments and is

transported through the subsurface by water, we propose that there is an association between *E. coli* presence the total area of irrigated land in a municipality.

A number of natural and human-induced processes work together in order for a well to become contaminated with fecal coliforms. Compiling a GIS database of the geologic, landuse, and agricultural information pertinent to the above-stated hypotheses will provide a solid platform for measuring and analyzing environmental characteristics in relation to water quality. In the following chapter, I outline the data layers sought and entered into to the MRA-GIS, and the spatial methods used to measure land characteristics local to private groundwater sources.

3. DATA AND METHODS

The chief aim of this research is to identify environmental characteristics that contribute to the prevalence of waterborne fecal coliforms in Southern Ontario's private groundwater sources. As discussed in section 2.1, this question lends itself to the practice of landscape epidemiology: the study of associations between geographic location, environment, and disease (Clark, McLafferty, and Tempalski, 1996). In order to facilitate this approach, a spatial database of multiple layers of environmental data was compiled using ArcGIS 9.0 software. The most important data for this research are private well water samples provided to the ARO research team by the Ontario Ministry of Health and Long Term Care (MOHLTC). This was the first and only data provided to our research team and served as the foundation of the spatial database, however, we required a host of other environmental data to compare the test results to.

In May 2005, I began a search for online data through spatial data clearinghouses, and contacted data custodians in both government and industry via telephone, email, and letter mail. Based on a survey of the current literature outlined in sections 2.3 and 2.4, the goal was to obtain (from reliable sources) continuous digital data layers at the finest resolution possible; describing land use patterns, soils, surficial geology, bedrock geology, aquifer boundaries, precipitation, temperature, and topography. Data on anthropogenic influences such as the location of septic fields, human population densities, and agricultural practices (animal densities, manure application, tillage, etc.) were also sought after. The data acquisition process continued until August 2005, at which point I had collected (at no cost) most of the data required[4]. However, certain data simply does not exist in digital format, or is unavailable for release due to data sharing constraints and confidentiality purposes. Data in this category include septic field sites, feedlot locations, manure storage sites, outhouse locations, aquifer boundaries, and subsurface hydraulic gradients. There will be further discussion of the cause and effects of missing data in section 5.2.1. In this chapter, I will provide a detailed account of the data collection, cleaning, and manipulation processes undertaken for each layer stored in

[4] See Appendix 3 for a comprehensive list of data attained, and sources.

the spatial database. In sections 3.1 through to 3.3, data structure and quality will be discussed, and in the final section of the chapter, I will describe the methods applied in the GIS for measuring environmental characteristics local to tested wells in the study area.

3.1 Well Water Data

The backbone of this project is a Province-wide database of water test results, compiled by the Safe Drinking Water Unit of the MOHLTC. These data were provided to the ARO research team under a Memorandum of Understanding, signed by Dr. Michael Buzzelli and myself. The data arrived in the format of two Access databases, one containing tests from the 2003 sampling season (May 1 – Oct 1 2003), and one from the 2004 season (May 1 – Oct 1 2004). In total, there are 181558 test results for the 2003 season; and 280139 test results for the 2004 season. The 2003 season is considered a pilot study for the ARO project, thus the 2004 season contains a greater number of records. No sub-sampling of the water test records occurred prior to our receipt of the data.

Each record in the database refers to a sample collected by a private well owner, sent to the MOHLTC laboratories for complimentary testing. As part of this public health initiative, well custodians pick up water testing kits from their Health Region's departmental offices, and conduct their own sampling from raw well water. The kit includes detailed instructions on appropriate sampling methods[5]; however there is no guarantee that the samples have been acquired accurately, and no indication of the amount of error introduced during the sampling process. Well owners are required to fill out a form to accompany the sample, providing information on the date the sample was collected and the mailing address of the property the well is located on. Once the sample reaches the laboratory, it is tested for the presence of *E. coli* and TCs. The results of each test are disclosed to the well owner, and recorded into a digital database along with the information copied from the hand-written forms completed by well owners. The data

[5] Find the sampling directions provided to well custodians in Appendix 4: Water Sampling Guidelines

provided to the ARO study team by MOHLTC are simplified versions of the original test records, and exclude the owner's name and other potentially sensitive or redundant information. Data attributes provided for each water test (where available) include: the date of sample collection, the date the sample was processed by the laboratory, *E. coli* count [6], TC count, and a description of the outcome of the water test (i.e. 'no significant evidence of contamination' or 'unsafe for consumption').

Geographic coordinates for the wells were not provided, preventing the immediate input of the well data to the GIS. Caroline Guenette, the ARO Project coordinator from the Public Health Agency of Canada, used a software package called GeoPinPoint to assign spatial coordinates to each test result based on the address fields in the database[7] (this process is called *geocoding*). Geocoding relies on the matching of two databases: a spatial database of digital road networks (an Ontario road network compiled by DMTI), and the attribute data (MOHLTC records). The address fields used in the geocoding process (in order from most to least accurate) include: the civic (street) address of the water source, lot number, six, five, and three-digit postal codes, and municipality. The address field in the water test database is matched to the attribute data of the road network, starting with the most accurate data (civic address). When records match, x,y coordinates are assigned to a well based on the corresponding address location in the street network. The remaining records are then re-matched to the next most accurate address field, until all records have been matched. In this case, only a small percentage of records were not assigned coordinates due to incomplete or inconsistent address information. For the 2003 season, 172572 entries (95%) were geocoded, while 8986 (5%) were not. In the 2004 dataset, 267174 (95%) samples were geocoded, while 12965 (5%) were not. Please refer to Appendix 2 for a detailed record of geocoding results for both seasons. Approximately 25% of the records in both databases were geocoded according to their civic address, and a further 5% by their lot and concession numbers, which also identify individual properties. The remaining records were

[6] The count is the total amount of Colony-Forming Units[6] (CFU) per 100ml sample (for both *E. coli* and TC bacteria), to a maximum of 81 CFU per sample. Any value over this arbitrarily-chosen number was considered completely overgrown and not worth additional laboratory time to count.
[7] For technical details, refer to Caroline Guenette's geocoding report, entitled: Performance Report: Geocoding a database related to water testing: ARO Study (2003 v.2.1, 2004 v.1)

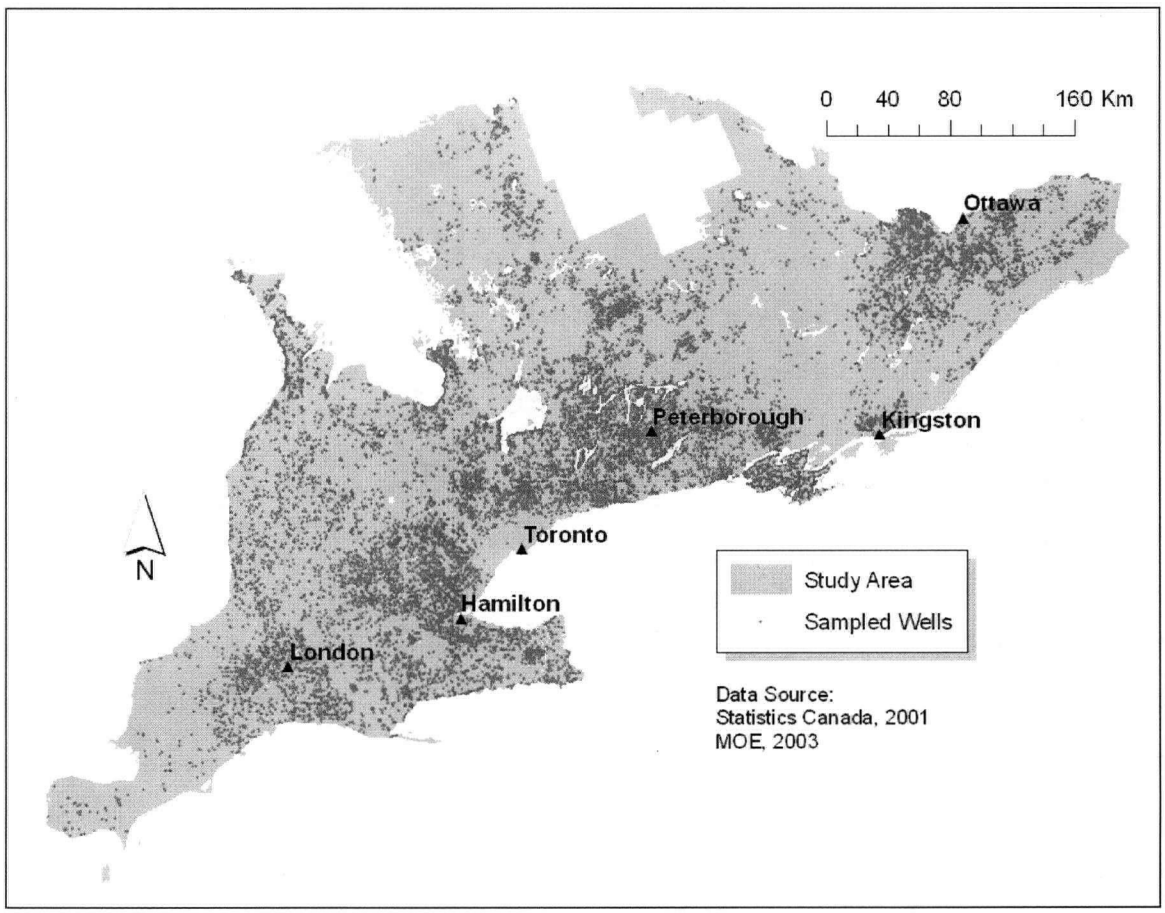
geocoded using more relaxed parameters such as the centroids of postal code and municipality areas.

Even if a record is geocoded with civic address information, there is always some degree of inaccuracy incurred as part of the geocoding process. The GeoPinPoint application assigns geographic coordinates (used to represent a well location) based on linear interpolation along a road segment of arbitrary length. In urban areas where road networks and house addresses are dense, research has shown that there can be as little as an average of 100m difference between the actual and geocoded locations (Bonner, Han, Nie, et al., 2003). However, in rural areas the discrepancies can be much larger, depending on the road network available. Numerous studies have found that positional errors are greater when rural addresses are geocoded, in comparison to urban addresses (Ward, Nuckols, Giglierano, et al., 2005; Vine and Degnan, 1997). Even in urban areas, addresses entered in the MOHLTC database do not represent authentic well locations. The geocoding process automatically places the x,y coordinate directly on a road segment, whereas in reality a well could be located anywhere on a property of unspecified size at the address. We also know that well locations cannot possibly be located at Post Office Box addresses, although thousands of records in the database were addressed as such by well owners. Unfortunately, obtaining additional address information to improve geocoding results is impossible; therefore we are forced to work within the constraints of these data, acknowledging their limitations.

Once x,y coordinates were assigned to each sample record, Caroline Guenette released the updated database to our research team. At our lab, I used ArcMap 9.0 software to plot the coordinates, creating a digital shapefile composed of 29134 points and 144501 attribute records for the 2003 season. The shapefile of the 2004 data is composed of 37359 points and 225052 records (see Fig. 3.1). The first reason for the significant discrepancy between well points and well records is because multiple samples are often sent in from a single well over the course of a season, so in some cases many test records are associated with a single point. Secondly, if a record does not contain specific address information, many sample records can be assigned to a single point (such as a municipality centre point) even if the samples originate from different wells. For

example, if two neighbors test their wells and do not provide a street address or lot number that can be geocoded, but do provide a postal code, then both wells (regardless of

Fig. 3.1: Sampled Well Locations, 2003



the lot they are located on) are assigned to the same point: the centroid of their 6-digit postal code region. Each record contains an *E. coli* count (number of colonies per 100ml sample), therefore one x,y coordinate could potentially have up to 957 different *E. coli* counts linked to it[8]. This case is an exception, as a majority of the unique wells in the database are associated with only one or two test results (79.6% in 2003, 76.2% in 2004), and the mean value of tests linked to each well is 4.96 in 2003 and 5.73 in 2004. Among wells geocoded to their civic address or lot number, the mean number of tests per well is only 1.24 in 2003, and 1.26 in 2004.

[8] 957 records assigned to one point (well) is the maximum number of records linked to a point in both 2003 and 2004. This number is high because all 957 of these records lacked detailed address information, and were geocoded to the same municipality centroid.

After a preliminary exploration of these data, it was apparent that for the purposes of our analyses, the database required refining. Selected samples were removed due to geocoding inaccuracies, incomplete laboratory testing status, and location outside of the study area. In addition, a smaller sub-sample was selected to comply with software constraints and data processing time. The following section describes the data cleaning progression that led to the four final sub-samples used in the analysis.

3.1.1 Cleaning the Well Database

We aim to use samples that provide the most accurate representation of microbial water quality in each well, and derive a measure of the prevalence of *E. coli* and TCs across numerous samples. This was accomplished through a five-step process of record elimination (refer to Table 3.1 for details). First, records not geocoded due to lack of sufficient address information were deleted from the database. Second, well points

Table 3.1: Cleaning the Well Database

Steps Taken to Clean Well Test Database	Number of Records Remaining	
	2003 Season	*2004 Season*
Total Records	181,558	280,139
1. Geocoded (assigned x,y coordinates)	172,572	267,174
2. Within Study Area	160,423	248,514
3. Tested for E. coli and Coliforms	144,501	225,052
4. Unique Wells (aggregated to x,y coordinates)	29,134	37,359
5. Accurate Location (geocoded to civic address, lot/con#)	27,096	35,296

located outside of the study area boundary were removed, as were wells that were not laboratory tested for *E. coli* or TCs[9]. Due to time constraints (and the lack of resolved climatic data), we chose not to pursue temporal or climate-related cross-sectional analyses. Instead, we focus on the impacts of local environmental characteristics on well water quality in a cross-sectional analysis of each sampling season. Therefore, for the purpose of using statistical methods to compare a well's microbial water quality with the

[9]If samples were received frozen, in a broken or leaking container, or if they were sampled from tap water instead of the raw well source, they were not eligible to be tested for the presence of *E. coli* or TCs.

surrounding landscape over the course of the entire season, we required a single *E. coli* outcome value, and a single TC outcome value for each point. This was achieved in the fourth step - records were not *removed* from the database, rather, many records were aggregated in into a single record (based on common x,y coordinates); therefore reducing the total number of records, but not well points. This method condensed the number of test records associated with each well to one, regardless of how many tests had been performed on the well. This aggregation prompted the question of how best to quantify *E. coli* and TC measurements in order to provide a suitable and standardized outcome for analysis.

Two approaches were considered: finding a coliform 'rate' for each well, or simply coding a well 'positive' or 'negative' according to coliform presence. The first option was to assign a *rate* to each well to represent the sampled *E. coli* and TC concentration over the course of an entire 5 month season. Applying this method raises the question of whether a static continuous value the most accurate way to represent these data. Assigning a continuous value to a well implies precision that does not exist. *E. coli* and TC counts in wells are not static. Rates vary over time and over climatic conditions (i.e. heavy rainfall events), and it is natural that tests performed on different days may return varying results. The presence of coliforms in groundwater rises over the course of a May-Oct sampling season, as temperature increases provide a more favorable environment for bacterial survival (Miller, Beasley, Yanke, et al., 2003). Coliform counts may also be variable throughout a single well at the time of sampling - a 250ml sample scooped from a well may not present a representative cross-section of the well's microbial climate. Atherholt et al. (2003) suggest that a groundwater sample should be analyzed ten times or more to confidently determine its sanitary status[10]. Further, the Provincial water testing laboratories do not count higher than 81 coliforms per 100ml sample. Any value over this arbitrarily-chosen number is considered overgrown and not worth additional laboratory time to count. Therefore, an absolute number of coliforms

[10] The methods used for the analysis of bacteriological water quality are standardized in each Province, and the MOE laboratories employ these established processes for sample submission, testing, and reporting. For the complete list of standards, refer to the MOE Protocol of Accepted Drinking Water Testing Methods (MOE, 2006).

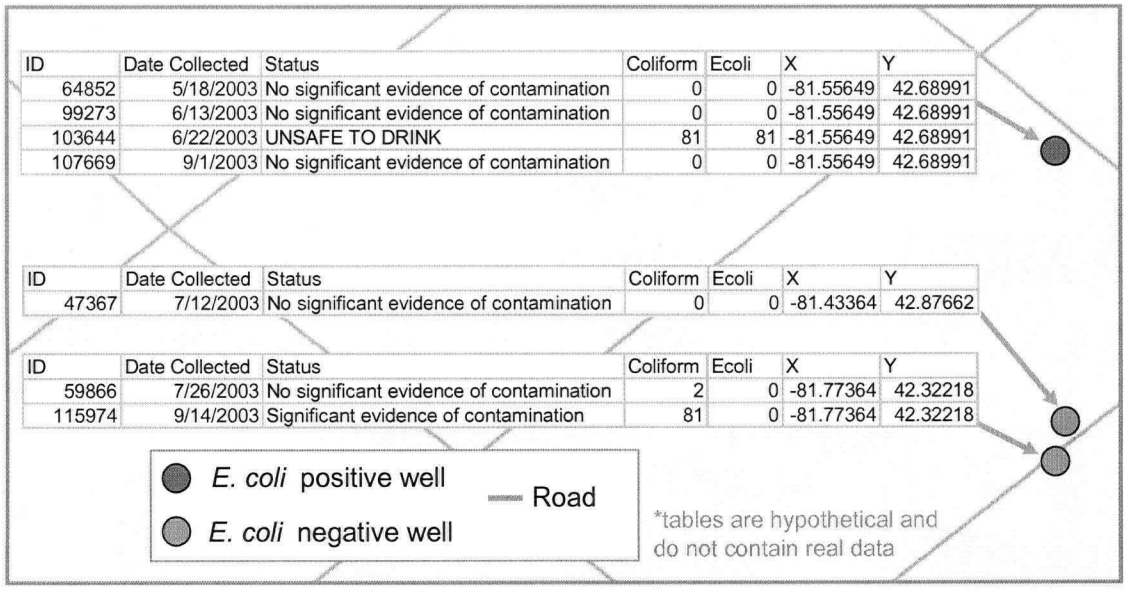
per 100ml sample can never be provided, and any rates created with the 81 CFU cut-off value are not accurate.

An additional problem with assigning a rate to each well lies in the sampling bias caused by self-reporting, as it is the well owner's responsibility to initiate the sampling process. Well custodians whose water tests positive for *E. coli* are more likely to send in further samples for testing than owners whose water tests clear. For example, in 2003, wells positive for *E. coli* were tested 2.97 times on average; whereas wells that never tested positive for *E. coli* were tested 1.66 times on average. In 2004, wells positive for *E. coli* were tested 3.49 times on average; whereas wells never tested positive for *E. coli* were sampled 1.83 times on average. This indicates a potential for error, as spurious estimates could be produced based on the autocorrelation of multiple samples. The rates also have the potential to be skewed by the repeated measures of wells that had tested positive earlier in the sampling season. If all wells were sampled the same number of times at standardized temporal intervals, then a rate would be a more reliable way to compare water quality across wells. However, due to the sampling and laboratory processes, and the dynamic nature of the living organisms whose patterns we are trying to describe, we rejected the idea of using a continuous rate as an outcome variable.

An alternative to using a continuous rate is to simplify the outcome variable to ordinal/discrete categories by assigning a binary code (1 or 0) to each well, indicating its potential for contamination by *E. coli* and TC. This method does not imply precision – it simply reports whether *E. coli* or TC have been found in any of a well's tests over the course of the season. I recoded each well either 'positive' for *E. coli* (if one or more *E. coli* coliforms had ever been detected in the well, no matter how many samples taken), or 'negative' if no sample had ever detected the presence of an *E. coli* coliform in that well at any point from May 1- Oct 1 each year. To achieve the aggregation, I imported the Access database file into SPSS 12.0, and employed the 'Data → Aggregate' function, using common x,y coordinates as the grouping variables for attribute aggregation (simply, if two records have the same coordinates, the records are combined). In the grouping parameters, I specified for the *E. coli* and TC counts to be *summed* during the aggregation. Once aggregation was complete, all *E. coli* and TC values >0 were recoded

to the value of 1 to indicate that the potential for contamination exists at that well site. Conversely, all wells for which *E. coli* or TC had never been found remained coded as '0' (see Fig. 3.2 for a diagram of this process).

Figure 3.2: Aggregating Well Records to Unique Well Locations



ID	Date Collected	Status	Coliform	Ecoli	X	Y
64852	5/18/2003	No significant evidence of contamination	0	0	-81.55649	42.68991
99273	6/13/2003	No significant evidence of contamination	0	0	-81.55649	42.68991
103644	6/22/2003	UNSAFE TO DRINK	81	81	-81.55649	42.68991
107669	9/1/2003	No significant evidence of contamination	0	0	-81.55649	42.68991

ID	Date Collected	Status	Coliform	Ecoli	X	Y
47367	7/12/2003	No significant evidence of contamination	0	0	-81.43364	42.87662

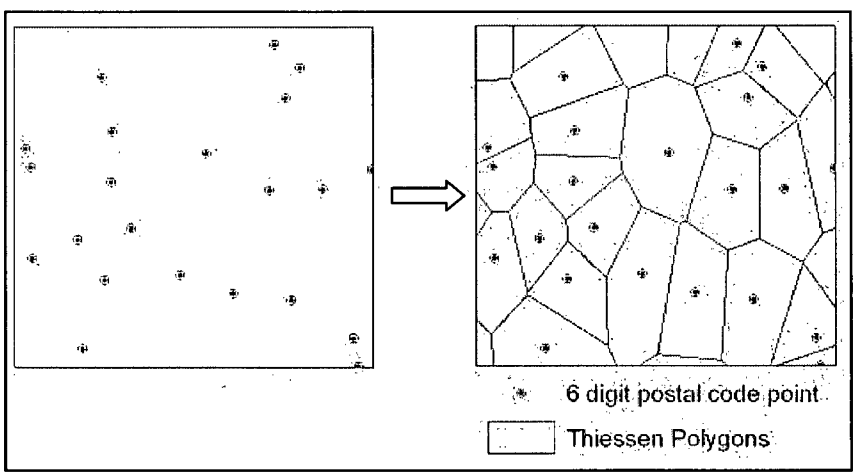
ID	Date Collected	Status	Coliform	Ecoli	X	Y
59866	7/26/2003	No significant evidence of contamination	2	0	-81.77364	42.32218
115974	9/14/2003	Significant evidence of contamination	81	0	-81.77364	42.32218

The result of the aggregation is a table that contains one record per well; with the following variables: X and Y coordinates, *E. coli* status (0 or 1), TC status (0 or 1), and a code conveying the level of geocoding precision. Although no data from any well sample was lost (all wells were considered), the aggregations significantly reduced the number of records for both sampling seasons: from 144501 to 29134 in 2003 (a 79.8% reduction), and from 225052 to 37359 in 2004 (an 83.3% reduction). These high reduction levels are due to the fact that wells geocoded to inaccurate locations have not yet been excluded from the pool; therefore the example of the 'well point' with 957 observations described earlier (which in actuality is a municipal centroid) is still provisionally included in the sample.

This leads us to the fifth and final step in database cleaning: excluding wells geocoded to imprecise locations. We are interested in examining the *local* environmental characteristics that affect well water quality; therefore it is logical to only include wells

that are accurately geocoded (i.e. to a specific property) in the analysis. Current literature on well protection area zoning reports that regions of influence for wells vary widely, depending on environmental and well characteristics (to be discussed further in section 3.4.1). However, estimates for small-capacity private wells place the radius of microbial influence anywhere from 50-300m (MOE, 2001; Lin, 2001; Horsley and Witten, 1995), thereafter effects drop off. Consequently, including a well that has been geocoded to the centroid of an area covering more than the area of a circle with a 300m radius ($282660m^2$, or 28.2 Ha) would introduce an increasing magnitude of error into an analysis of local effects. In order to determine the level of geocoding to be used as the threshold for discarding records, I examined a shapefile of the locations of 6-digit postal codes, published by Statistics Canada. In our study area, this shapefile contains 253070 points, each point representing the centre point of a six-digit postal code region. I created Thiessen polygons[11] around each point in an attempt to estimate the postal code regions using the point shapefile (refer to Fig. 3.3 for a diagram on the creation of Thiessen polygons).

Figure 3.3: Thiessen Polygons



I calculated the area for each of the 253070 polygons, and found the mean area of all polygons in the study region to be $523329m^2$ (52.3ha). Because the mean area of the 6-digit postal code regions is almost twice as large as the well buffer zones (28.2ha),

[11] Also known as Voronoi polygons, Thiessen polygons bound the region that lies closer to a point than any other point. The boundaries lie exactly between two adjacent points.

using wells geocoded to the 6-digit postal codes would introduce uncertainty to the analysis. Thus, the only two levels of geocoding acceptable to use for the analysis are those records that are geocoded to the civic address, and lot/concession number. By restricting our sample to a lot-specific level of geocoding, we also eliminate the possibility that multiple samples aggregated to one point might have originated from separate properties[12]. To complete the final step of data reduction, I selected the records geocoded to their civic addresses or lot/concession numbers, and discarded the remaining records. For the 2003 season, this resulted in a 7% reduction in sample size (n = 27096), and in 2004 only a 5.5% reduction (n = 35296).

3.1.2 Selecting Wells for Analyses

Large samples provide a more accurate, true-to-life representation of a phenomenon than do small sample sizes; yet complications can arise when dealing with very large samples. Because our sample sizes are in the tens of thousands, computing ability is an issue to contend with. Our spatial database is stored in a powerful desktop PC, with ArcGIS (v. 9.0) software. When the first stages of analysis were carried out, essential GIS processes (buffering and overlay, to be further discussed in section 3.4) could not be completed without the system failing. The largest sample size that the computer managed to process successfully was limited to 2000 wells, so for efficiency's sake the sample size required reduction.

The sub-sampling design was influenced by our objective to compare landscape characteristics surrounding clean wells to landscape characteristics surrounding contaminated wells. Logically, each sample must contain both positive and negative wells (as a control). Four samples were required to analyze *E. coli* and TC patterns separately (two samples per season). In the 2003 season, 6.9% of wells tested positive for *E. coli*, and 43.3% tested positive for TC (see Table 3.2), with similar proportions for the 2004 season. The reason for the high percentage of TC positive results is that TCs are

[12] However, there is no indication whether multiple samples might originate from more than one well at the same address.

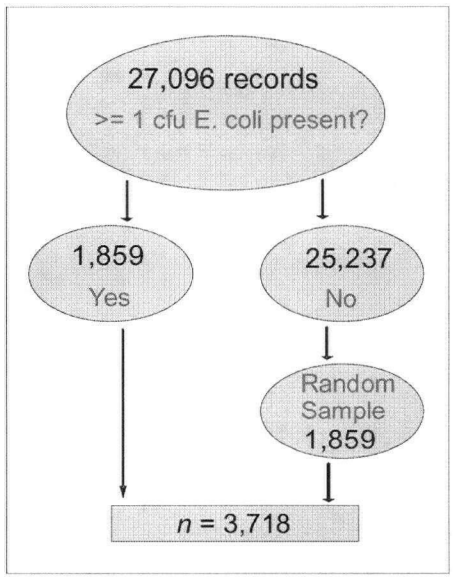
naturally more prevalent in the environment than *E. coli* (An et al. 2005, Health Canada 2006). Further, each of the *E. coli* positive results also tests positive for TCs, as *E. coli* are part of the TC family of fecal coliform bacteria.

Table 3.2: Positive and Negative Well Samples

	2003		2004	
	Count	**% of Total**	**Count**	**% of Total**
Total samples	**27096**	**100.0%**	**35296**	**100.0%**
E. coli positive	1859	6.9%	2779	7.9%
E. coli negative	25237	93.1%	32517	92.1%
TC positive	11752	43.4%	15925	45.1%
TC negative	15344	56.6%	19371	54.9%

To reduce the sample size without losing the limited number of positive contamination results, every record that had tested positive for *E. coli* or TC was retained. Then, as a control, an equal number of negative results were selected using a random number selection process in SPSS 12.0. The issues surrounding the use of an aspatial sampling scheme will be discussed further in section 5.2.2. The diagram in Figure 3.4 depicts a flow chart outlining the sampling process for *E. coli* in 2003. Selecting a 1:1 ratio of positive to negative results ensures that the statistical power of positive and negative results are the same, and reduces the sample sizes sufficiently. The final samples are comprised of 50% positive and 50% negative results, and are still large in size:

Figure 3.4: Sample Selection, *E. coli*. 2003



2003: *E. coli* n = 3718 TC n = 23504
2004: *E. coli* n = 5558 TC n = 31850

3.2 Local Environmental Data

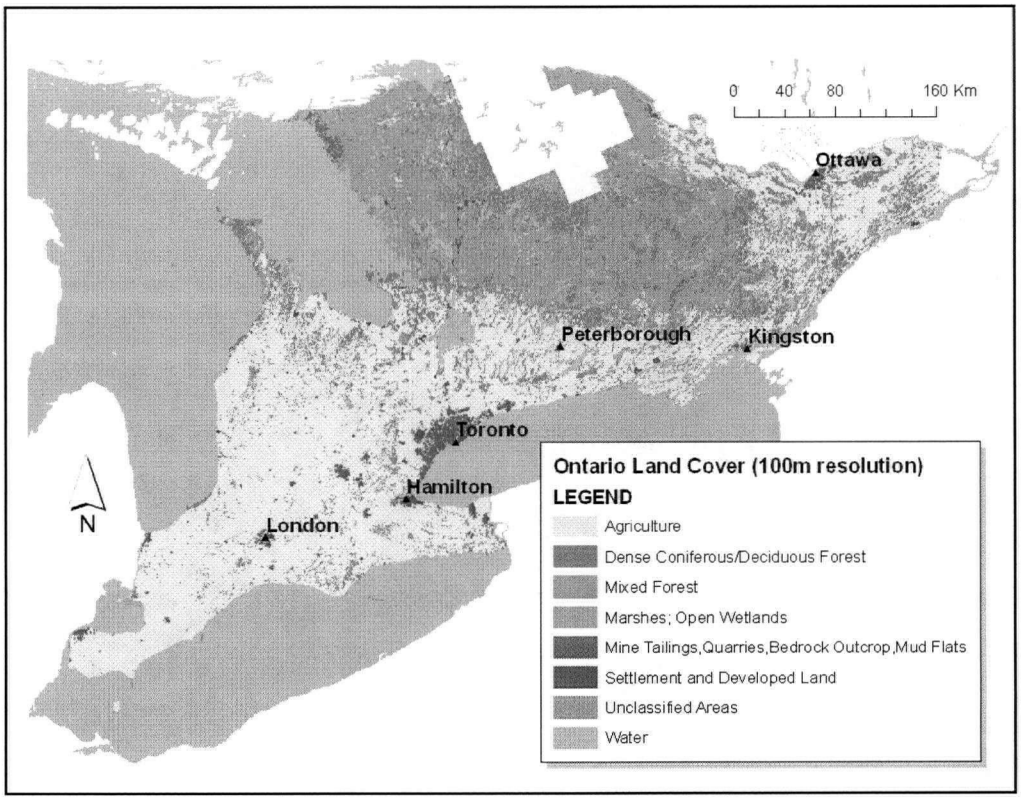
The locations of the water samples extend across the entire study area, therefore continuous data of fine resolution are required to examine land characteristics surrounding each of the wells. All of the layers listed in section 3.2 are 100m or less in resolution (cell size), meaning that variability in each data layer will be captured with well buffer zones with radii 100m and over.

3.2.1 Land Cover

It is widely acknowledged that landuse in a water catchment has a profound impact on water quality (Bolstad et al., 1997; Sekhar et al., 1995; George et al., 2004; Nygard et al., 2004). Thus, data describing Ontario's land characteristics are important components of this project. Two possible sources of data were considered: a landuse layer produced by the Canadian Land Inventory (CLI), and a landcover layer produced by the Ontario Ministry of Natural Resources (MNR). There is no significant difference between the land classification groups of the two layers (agriculture, mixed forest, developed land, marsh, etc), and both have similar patterns when visualized. The majority of the CLI landuse data was collected in the 1970s, with some updates performed in urban areas in the 1990s. These data were primarily collected via the interpretation of air photographs, and the resolution is approximately 25m. However, data are only available in Southern Ontario for 'populated' areas, which accounts for approximately 75% of the study area. In comparison, the MNR landcover data were collected more recently (1986-1997) from remote sensing data, but the layer has a coarser resolution at 100m. Although the resolution is not as fine as the CLI landuse data, it is more recent and is available for the entire Province. For this reason, the MNR landcover layer was chosen for analysis.

The MNR data is available online, free of charge through GeoGratis[13]. The Ontario layer is available in approximately 80 zipped files, which were downloaded from the website, unzipped, and converted from .e00 (interchange) format to ESRI 'coverages' for input to the GIS. I projected each tile into UTM coordinates separately, and then used the *union* tool in ArcToolbox to merge the tiles together in groups of 5, eventually forming one cohesive layer (see Fig. 3.5).

Figure 3.5: Land Cover Map



3.2.2 Soils

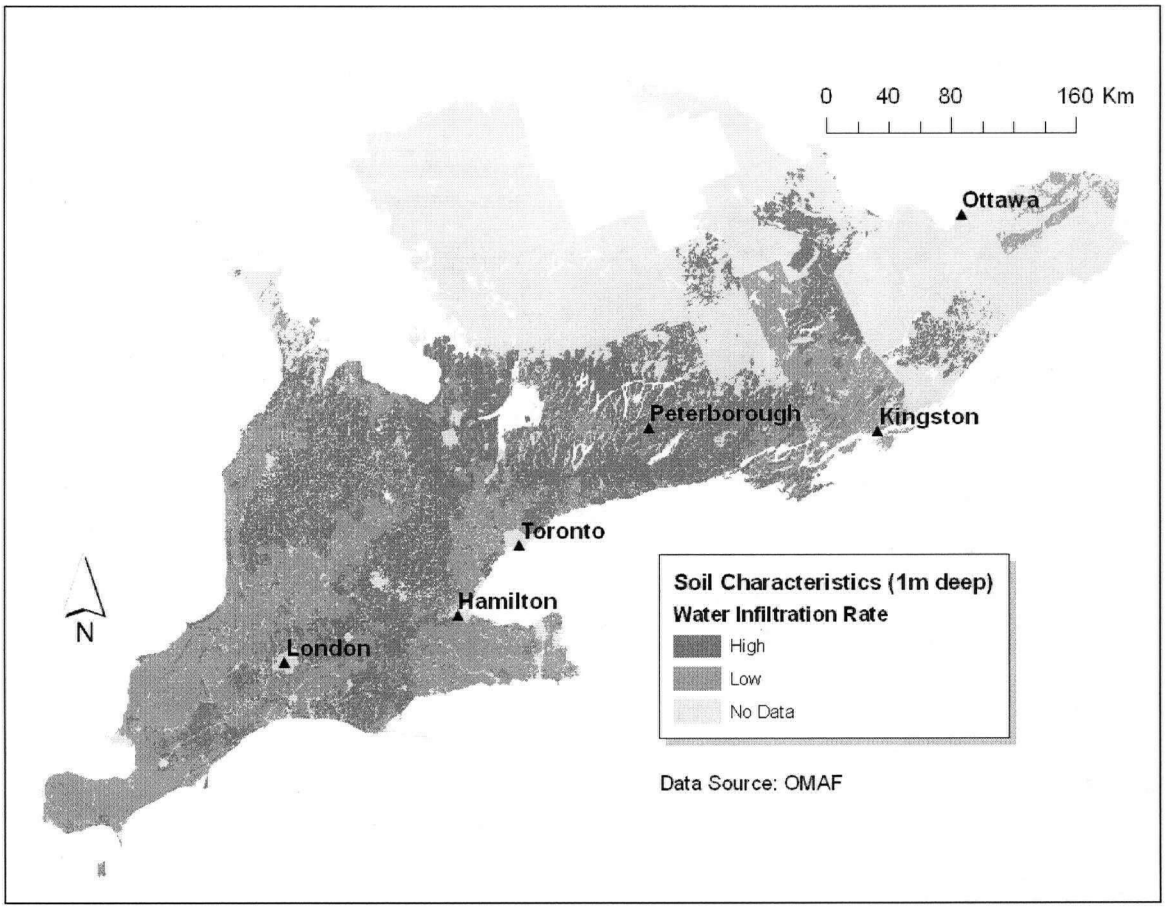
While the landcover layer provides information on surficial processes, the composition of the soil structure can give insight into the movement of microbes from the surface to a groundwater source. The soil layer was provided on CD by the Ontario Ministry of Agriculture and Food (OMAF). The data was collected on a municipal basis

[13] GeoGratis is a website for the dissemination of Canadian geospatial data, hosted by Natural Resources Canada, and can be accessed at the following URL: http://geogratis.cgdi.gc.ca/clf/en

by the Canadian Soil Information System (CanSIS) from 1970 – 1990, and describe soil characteristics from the surface to 1m in depth. At a 50m resolution, these soil surveys are the most detailed data available, however data is missing for large parts of the study region (see Fig. 3.6). OMAF data custodians provided two reasons for this: the first being a lack of soil in an area (i.e. the exposed bedrock of the Canadian Shield, or paved-over urban centres), and the second reason being that data was not recorded in areas where the land is considered unsuitable for farming (i.e. wetlands and exposed rock).

As discussed in the hypotheses section (2.4), we are most interested in the water infiltration rate of a soil: the velocity at which water enters into the subsurface. OMAF scientists created a derived variable to classify soils into four hydrological soil types (A, B, C, and D) according to water run-off and infiltration rates, based on soil texture, type, porosity, and drainage. Soils in group A are typically sandy and gravel soils, with low runoff and a high infiltration rate, whereas group D have a high runoff and low infiltration rate, and include clay soils with a high swelling potential or shallow soils over impervious material. For ease of analysis, I combined soils in groups A and B together, and soils in groups C and D together to form two categories: high to moderate infiltration rates (A, B), and slow to very slow infiltration rates (C, D). See Fig. 3.6 for a map of the soil characteristics in the study region.

Figure 3.6: Soils Map

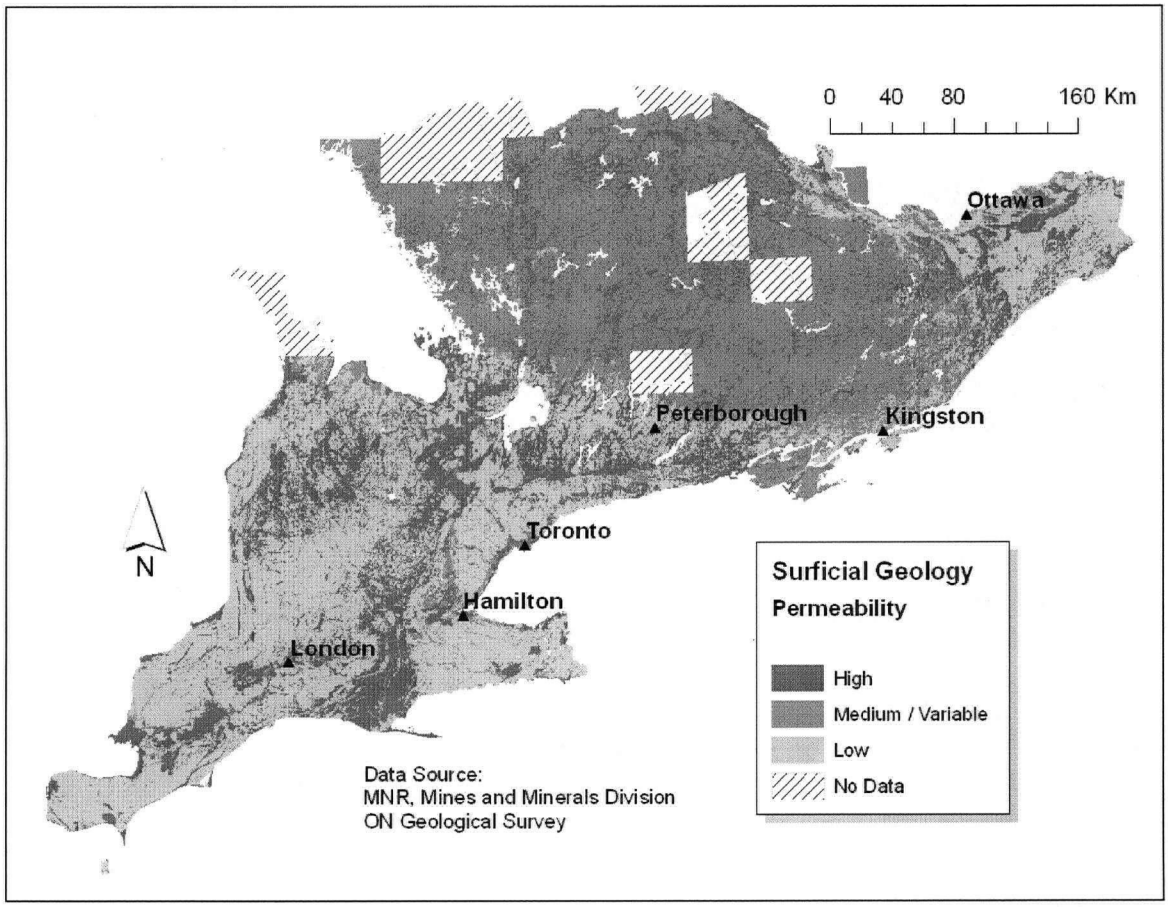


3.2.3 Geology

The soil layer describes geology from the surface down to one meter in depth, however most wells collect water from a source deeper in the subsurface, as private wells in the region typically range from 10m to 100m in depth. Data on surficial and bedrock geology provides insight into the hydrogeological processes occurring below the uppermost soil cover. This data was provided to us by the MNR Mines and Minerals Division, and was collected by the Ontario Geological Survey and its predecessor organizations over the course of 100+ years, at a resolution of 100m. The three data layers chosen for input to the GIS are surficial geology (permeability), bedrock geology (material and age), and areas of moraine deposits – all of which offer nearly complete coverages of the study area.

Surficial geology refers to all materials overlying the bedrock layer, including the top 1m of the surface; therefore we expect these data to be similar to the soils layer[14]. Included in the attribute table of is a derived variable calculated by MNR, classifying the permeability of the surficial geology into three classes: high, medium, and low (See Fig. 3.7). Permeability refers to the rate at which water infiltrates into the subsurface, and there is no apparent difference between this metric and the infiltration rate metric in the OMAF soil layer, except for terminology.

Figure 3.7: Map of Surficial Geology

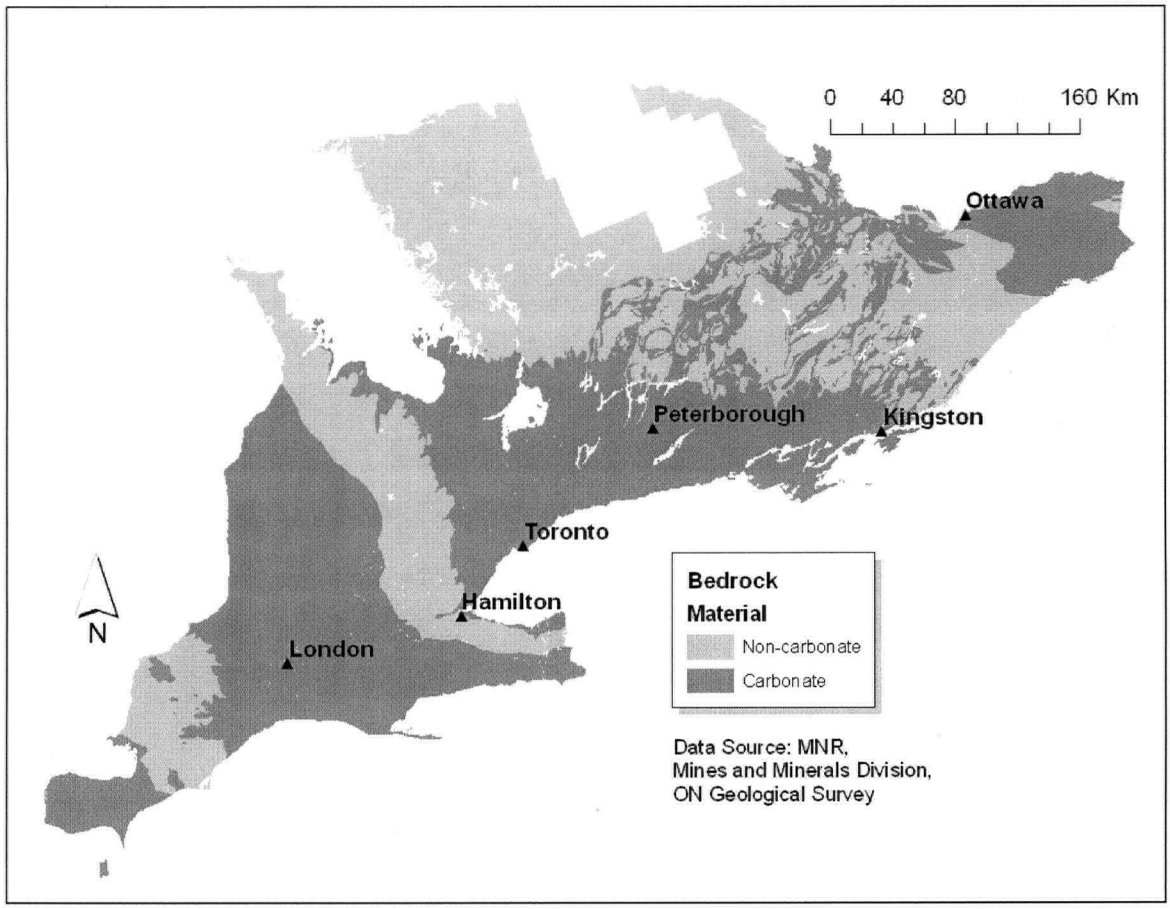


Bedrock Geology refers to the solid rock underlying all loose geological materials. As described in the hypotheses (section 2.4), we are interested in whether the

[14] Visually, the surficial geology and soils layers have similar patterns. The dangers associated with using two similar datasets in statistical analyses (i.e. misspecification bias caused by multicollinearity) will be further discussed in section 4.2.3.

bedrock is composed of carbonate or non-carbonate materials. In the attribute table of the bedrock layer, a field lists the primary rock types of each map area. By querying this field, I reclassified each polygon into one of two categories: carbonate bedrock (limestone, calcite, or dolomite), or non-carbonate bedrock (all remaining records). See Fig. 3.8 for a map of bedrock materials.

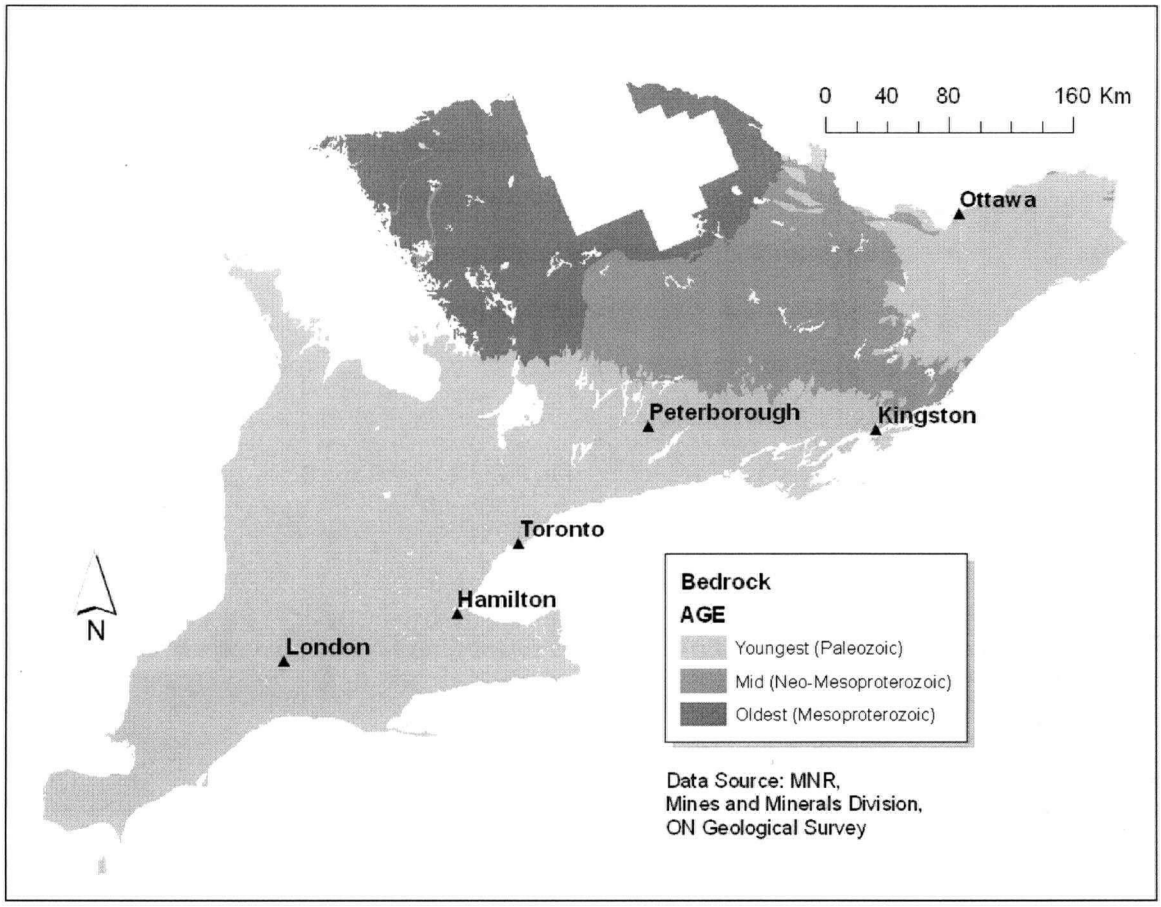
Figure 3.8: Map of Bedrock Material



The third variable of interest is the age (era) of the bedrock. Using the bedrock age field in the attribute data, I reclassified the bedrock layer into three categories listed below. For details see Fig. 3.9.

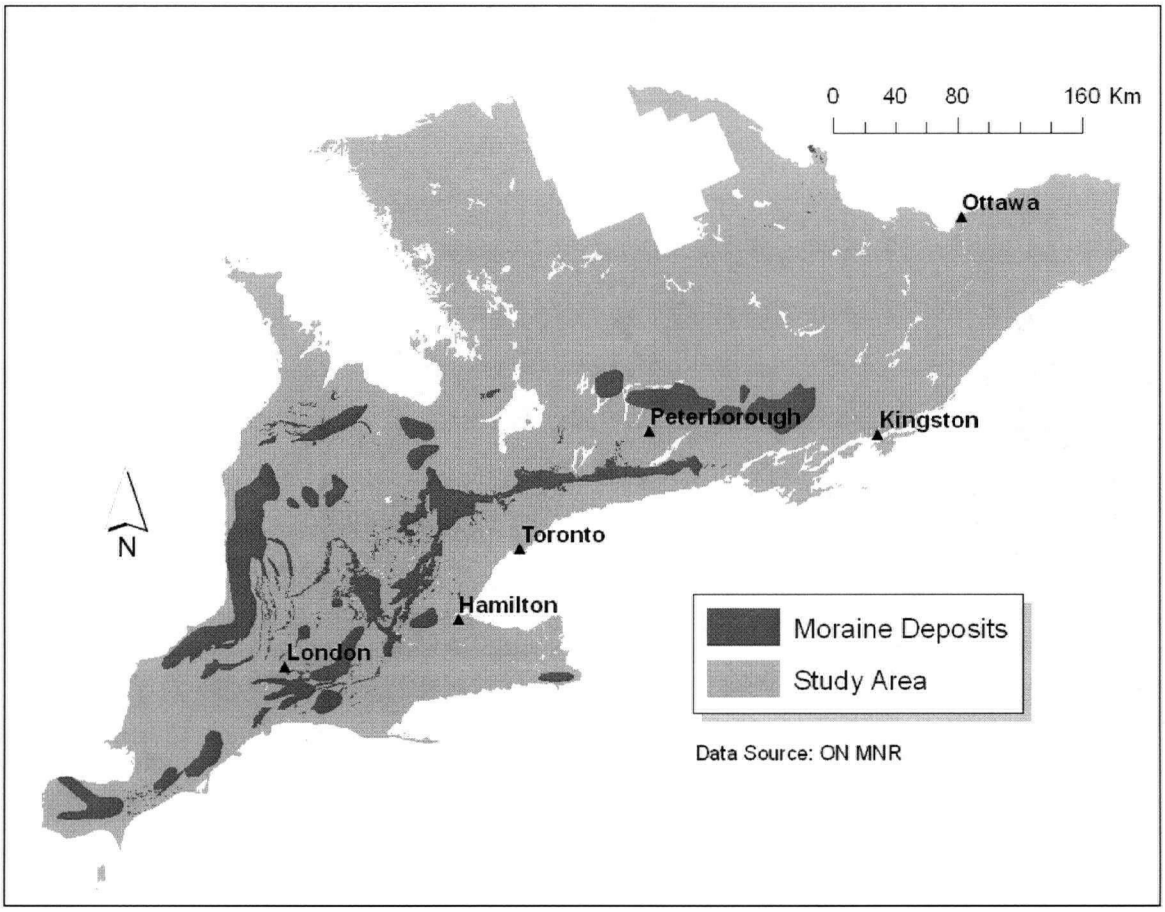
1. Paleozoic 542 Million years ago (Ma) to 251 Ma
2. Neoproterozoic 1000 Ma to 542 Ma
3. Mesoproterozoic 1600 Ma to 1000 Ma

Figure 3.9: Map of Bedrock Age



Due to the important hydrogeologic properties of moraine deposits and their abundance in the study area, the fourth geologic dataset added to the GIS indicates areas of glacial moraine deposits, also obtained from MNR (see Fig. 3.10 for details). These data are the last of the local environmental layers to be included in the analyses.

Figure 3.10: Map of Moraine Deposits



3.3 Aggregated Census Data

The environmental data described above are provided at fine resolutions, the detail of which is usually dictated by the data collection process rather than confidentiality issues. Data regarding human population and agricultural demographics are more sensitive in nature. Statistics Canada collects human and agricultural census information at the individual level, and for confidentiality reasons these individual data are aggregated to large, arbitrarily-chosen administrative boundaries for public release. Dissemination Areas (DAs) are the smallest units for which data are available for the human census, with a mean area (in the study area) of 720ha, and Census Consolidated

Subdivisions[15] (CCS) are the smallest units for which agricultural data are available in the study area, with a mean area of 54123ha.

3.3.1 Agricultural Census

Every five years, coinciding with the Canadian population census, Statistics Canada performs a census of the Agricultural industry. The most recent census for which data is available was conducted in 2001. Thousands of different variables are provided, ranging from farm revenue, size, crops, and animal counts to agricultural practices. Available for download through University of British Columbia (UBC) Data Services[16], these data tables can be linked to large geographic areas called Census Agricultural Regions (CAR), or to CCS regions. CCS data was used exclusively for this analysis as it is at a higher resolution (smaller geographic areas) than CAR data. The spatial data are also provided by Statistics Canada, in digital shapefile format. Only variables pertaining to the hypotheses were selected for analysis, listed below:

- Total area of land fertilized by manure (ha)
- Total area of irrigated land (ha)
- Total area of tilled soil (ha)
- Total area of untilled soil (ha)
- Total cattle and calves
- Total pigs
- Total sheep and lambs
- Total horses and ponies
- Total hens and chickens

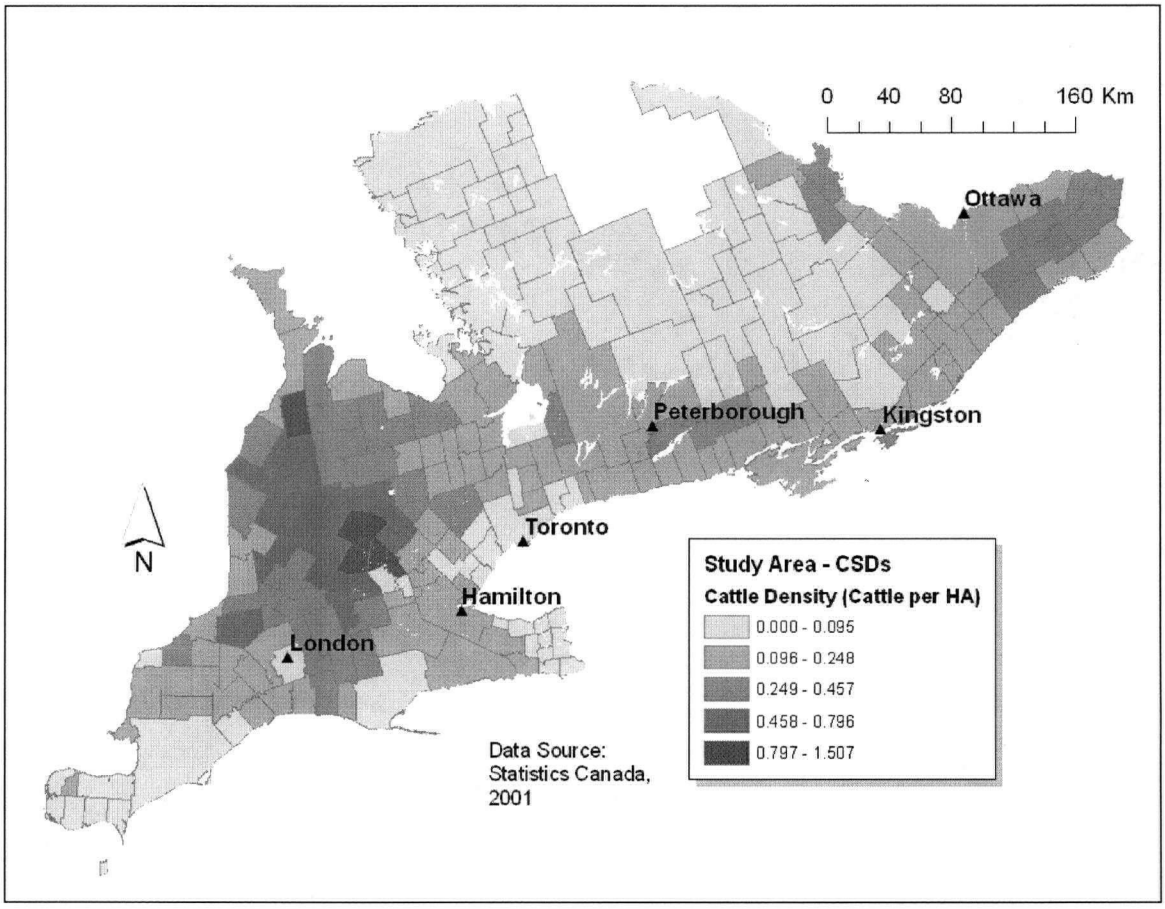
Linkage of the data tables to the CCS boundary file was performed in ArcMap, using the *join attribute* function. Both the tabular and spatial data have a field containing

[15] CCS areas are the same size and shape as municipal boundaries in Ontario.

[16] The UBC data services website is available to UBC students, faculty, and staff. The website is available at: http://data.library.ubc.ca/

CCS identification codes, used as the connection key between the two data sets. The initial *join* was successful, however upon inspection, agricultural census data were missing for 45 CCSs. There is no documentation of this in the metadata, so I contacted a census official at Statistics Canada who informed me that the 45 CCS areas in question contained fewer than 20 farms. For confidentiality purposes, Statistics Canada aggregates data for all CCS regions with <20 farms to an adjacent CCS tract, leaving null fields for 45 CCS regions. Statistics Canada publishes a correspondence file that documents these census tract aggregations, and I used this file to create a similar correspondence file in the GIS. I then used the 'dissolve' function in ArcView 3.2 to join the adjacent CCS units together where data aggregations had occurred, in order for the spatial data to 'match' the aggregated census data. Fig. 3.11 below is a chloropleth map of the final CCS regions, symbolized according to cattle density.

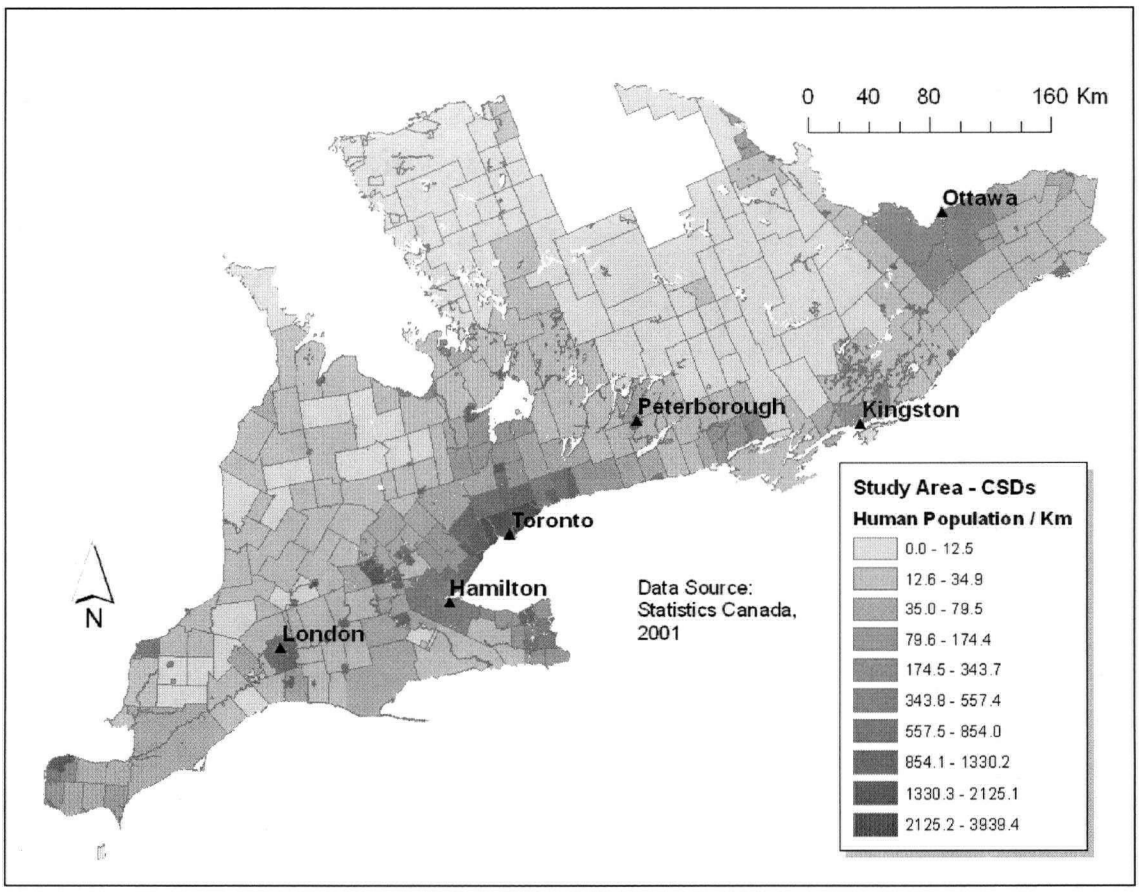
Figure 3.11: Map of Cattle Density



3.3.2 Human Population Census

Data from the 2001 human population census was used to examine the association between population density and water quality in a watershed. The methods for preparing these data are identical to the agricultural data: spatial and tabular population data were downloaded from E-Stat online (Statistics Canada's data site[17]), and joined in the GIS. The area of each DA was found in ArcView 3.2, and then the total population of each DA divided by the area to find the population density (number of persons per square kilometer). The mean population density for the study region in 2001 is 4084 people/km^2 – see figure 3.12 for a map of population by CSD. The contrasts in population density between rural and urbanized regions are clearly visible on this chloropleth map.

Figure 3.12: Map of Human Population Density



[17] The E-stat website can be accessed at:
http://estat.statcan.ca/cgiwin/CNSMCGI.EXE?ESTATFILE=EStat\English\E-Main.htm

3.4 GIS Methods

The data described above are compiled within a GIS to form a set of layers that represent many natural and anthropogenic processes affecting groundwater quality in the study region. Not only does the GIS allow for the storage of spatial data, it also serves as a platform for data querying, manipulation, and analysis. This research examines the impact of local land characteristics on a well's microbial water quality; therefore rather than using the environmental data for the entire study area, only spatial data in close proximity to well locations were selected for examination. I used the *buffer feature* tool in ArcMap's *proximity toolbox* to create digital circles around well locations, representing a zone surrounding a well where land characteristics might potentially have an impact on the source's microbial water quality. Then, using these circular buffer zones as 'cookie cutters,' I used the *overlay* feature to select only the environmental data in close proximity to a well, and discarded the irrelevant remainder of the data. By performing database queries and area calculations on the remaining dataset, the final tables used for analyses were constructed. These tables consist of one record per well (coded either positive or negative for microbial contamination), with many fields, each containing a quantitative value listing the total area of an environmental characteristic found within the buffer zone of each well. This format allows for the statistical analysis of patterns in land characteristics (based on the area measurements) surrounding clean wells versus contaminated wells. In this section I will describe the methodologies employed to derive the tables used for the analyses described in chapter 4 (results).

3.4.1 Well Buffer Zones

In order to select land characteristics for analysis, a circular buffer was drawn around each well to represent a zone of influence. The purpose of the uniform buffers is to *efficiently* capture land characteristics that have an impact on the source's microbial water quality. Two major questions arose, first: a well's zone of influence is never perfectly circular, nor static – how adequately does a circular buffer describe complex environmental processes? Secondly, what size should a buffer radius be to capture a

zone of influence over such a large sample size? As described in the literature review, research on the use of GIS in analyzing the impact of local land characteristics on microbial well water quality has not been widely attempted; therefore there is no substantial body of literature to draw from to develop methodologies.

A well capture zone is the land area surrounding a groundwater source where there is a potential for microbiological contaminants to be deposited on the surface, enter the water table and travel through the subsurface, eventually contaminating the well. This process is mediated by specific environmental characteristics in that region such as landuse type, soil porosity, and direction of groundwater flow. Well characteristics (depth, casing material, and pump rate) also affect the extent of the zone of influence. Due to the variability of the natural environment and the wells themselves, the fingerprint of a capture zone differs for every well. No fixed guidelines appear in academic literature or in technical documents, probably for fear of risks associated with underestimation. Many techniques are used by scientists to delineate well capture zones, ranging from complex computer modeling to assigning an arbitrarily-fixed radius to a source. In 2001, the Ontario Ministry of Environment (MOE) published a protocol for delineating wellhead protection areas for municipal wells (MOE, 2001). In the ministry's eyes, the preferred methods for Wellhead Protection Area (WHPA) delineation are three-dimensional steady-state computer modeling programs, like MODFLOW. These models, along with other analytical methods of delineating WHPAs[18] require a host of input data that are unavailable to us at this time, such as hydraulic flow gradients, well pump rate, and aquifer thickness. For these reasons, using analytical methods to determine WHPAs is beyond the scope of the project[19].

A simpler approach to zone delineation known as the arbitrary fixed radius method is supported by the World Health Organization (WHO, 1993), ON MOE (2001), US EPA (1991), and many local governments and technical institutions (NJDEP 2003, BCWWA, 2006) as a cost-effective and reliable method for WHPA delineation. This

[18] For example, the Uniform Flow and Calculated Fixed Radius Methods.
[19] Graham McIntyre, another UBC geography graduate student working with Dr. Buzzelli on the ARO project, is conducting research on using computer modeling, to analyze the same ARO water data used in this research.

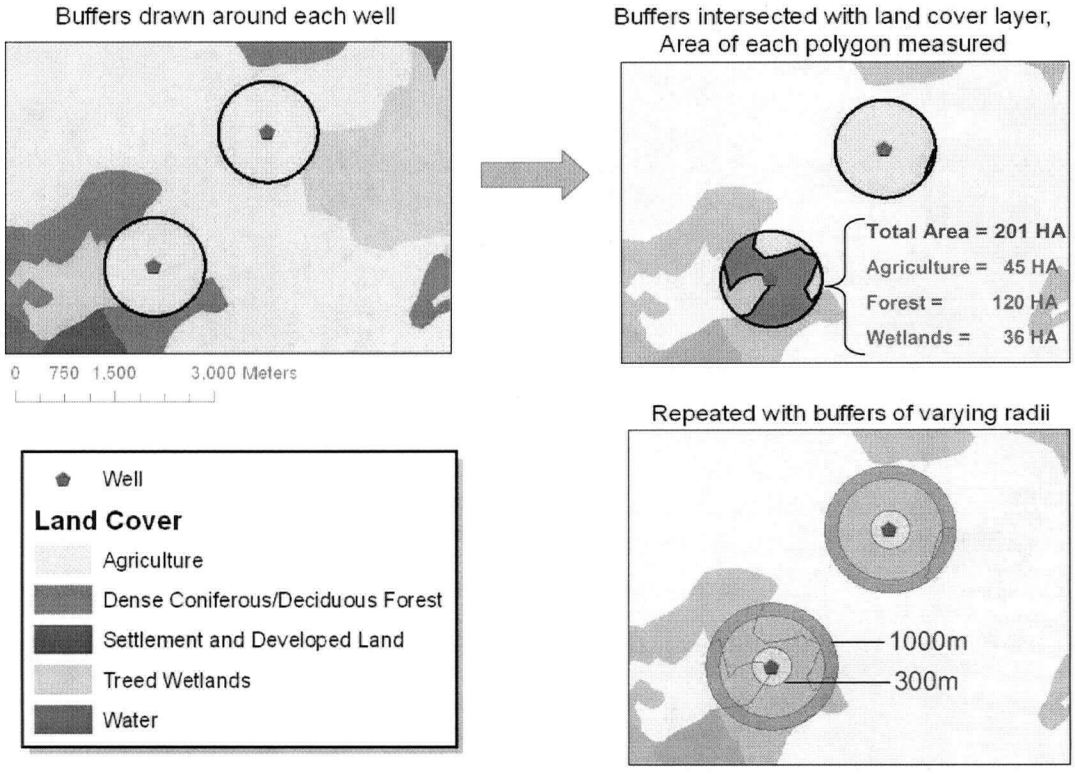
method involves simply drawing a circle around a well to estimate the travel time of organisms to a well source. Although the subsurface water gradient near each well is different, it has been suggested that "buffer areas are best calculated as circular entities, allowing ease of generation in a geographic information system (GIS), and avoiding problems of orientation when groundwater flow direction is unknown" (McLay and Dragten, 2001, p. 193). The circular zones represent isochrones: contour lines that denote the time of travel. For medium capacity municipal wells, the WHO suggests a 50-day isochron (parallel with the assumption that most microbes die after 50 days in groundwater) which can range from fifteen meters in a confined aquifer, to 300m in an unconfined aquifer (WHO, 1993). The Ontario MOE suggests that a two-year Time of Travel (TOT) zone be drawn around a municipal wellhead, starting from 100m (MOE, 2004). The New Jersey Department of Environmental Protection (NJDEP) suggests a two-year TOT range of 200-900m, and a US Standard handbook on water calculation methods suggests a 300m radius (Lin, 2001). A workshop on Source Water Assessment and Protection published online by Groundwater.org (a US-based groundwater research organization) also recommends 200m-300m radii values for a two-year TOT isochron (Herpel, 2006), and in their GIS analysis of private wells, Kistemann et al. (2000) use a 50m radius to buffer their water samples.

Turning to studies that employ buffering techniques to study nitrate concentrations, researchers used buffer zones ranging from 250m (Barringer et al., 1990) to 800m (Eckhardt et al., 1995). Clearly, no firm consensus on an appropriate buffer size exists, so we thought it practical to test many different buffer radii in order to identify which buffer size might yield the most significant results in our study region. The buffer radii must be at least ~200m to capture variability in the environmental data layers, and 1000m or less for the computer to process the data without failing. Five different buffer sizes were created around each well in our 2003 *E. coli* sample, at distances of 200m, 300m, 650m, 900m, and 1000m, and were used to select environmental data surrounding the wells for further analysis. The methodologies employed in the selection process are the foci of the following section: Joining Environmental Data and Water Results.

3.4.2 Joining Environmental Data and Water Results

In order to compare water quality results, local environmental data, and aggregated census data, all datasets have been linked together to form a single table. Merging the aggregated census data with the water test results was a straightforward process: the attribute tables of the water samples were joined to the CCS and DA attribute tables using the *join by spatial location* function in ArcMap. Each well was assigned the complete set of census data values belonging to the administrative region the well is located in, making the comparison of census data and water test results possible. The approach to measuring local environmental characteristics is different, as continuous values (the total area in m^2 of the particular characteristic within a buffer zone) are being used as a metric, rather than densities or rates. First, each well buffer was assigned a unique ID code, then the *intersect* tool in ArcView 3.2 was used to clip (overlay) the environmental data layers with the buffers, saving only data in the same spatial location as the buffer zones. Then, the area of each polygon was calculated (see Fig. 3.13).

Figure 3.13: Well Buffers and Overlay



Some buffer zones contain more than one polygon of the same class[20]; therefore I aggregated the GIS database file by the unique well ID code, and summed together the area values of similar land classifications. First, the buffering/overlay process was carried out once for each of the five different buffer zones (200m, 300m, 650m, 900m, and 1000m) with just the 2003 *E. coli* sample, in order to compare the effects of varying buffer sizes on the associations with water quality. The results (to be discussed in section 4.1) indicate that a 300m buffer radius is an appropriate size for measuring the full set of land characteristics.

For the second round of analyses, a 300m buffer was used to repeat the buffering/overlay process 28 times: once for each of the seven environmental data layers, for each of the four water samples. Using the *merge data* function in SPSS, the 28 tables were consolidated to create four final tables, one for each sample. Each table contains one record per well sampled, with numerous fields listing the area (in m^2) of each environmental characteristic captured in the buffer zone, as well as the rates and densities of the aggregated census data. For a sample section of one table, see Appendix 5. The tables were imported into SPSS 12.0, and the associations between the microbial test results and the measured environmental data were examined using a number of different statistical analyses, to be outlined in detail in the following chapter.

[20] For example, two small polygon islands of agricultural land may exist within a larger forest class polygon.

4. RESULTS

The purpose of this chapter is to describe various descriptive and inferential statistical analyses used to summarize and explain the data, and the results garnered from these tests. The techniques are exploratory in nature as no protocol exists for applying statistics to describe the prevalence of *E. coli* in groundwater based on land characteristics. An in-depth discussion of the results can be found in the following chapter.

4.1 Buffer Size Comparisons

All analyses described in this chapter have been performed on four tables of data, one table per sample group. These tables contain continuous data quantifying the total area of an environmental variable measured within a well buffer zone. As discussed in section 3.4.1, there is contention among experts as to what the most appropriate radius of this buffer zone should be, although suggested radii range from 50m to 900m. Before embarking on an extensive analysis of the data, a suitable buffer size for capturing land characteristics local to wells had to be determined. To address this question, the independent samples t-test was applied to compare land characteristics surrounding clean versus contaminated wells, using data collected with five different buffer zones ranging from 200m – 1000m. The goal was to determine the buffer size that returned the most significant results. To increase efficiency, only 9 out of 26 variables were examined (agricultural land area, soils, and agricultural census data), and a smaller sub-sample (n=1512) was selected from the 2003 *E. coli* sample group. The mean values of each environmental variable were compared using the independent samples t-test, the results of which are listed in table 4.1.

Table 4.1: Buffer Size Comparison Results

Water Testing Data - 2003

Grouping variable: *E. coli* presence/absence p = 0.05
n = 1512 (756 *E. coli* positive wells, 756 *E. coli* negative wells)

	200m buffer		300m buffer		650m buffer		800m buffer		1000m buffer	
Variable	*t*	Sig (2-tailed)	*t*	Sig (2-tailed)	*t*	Sig (2-tailed)	*t*	Sig (2-tailed)	*t*	Sig (2-tailed)
Area of Agricultural Land in Buffer	**2.915**	**0.004**	**2.955**	**0.003**	**2.470**	**0.014**	**2.312**	**0.021**	**2.187**	**0.029**
Area of soils with high infiltration rates	**-4.403**	**0.000**	**-4.142**	**0.000**	**-4.311**	**0.000**	**-4.399**	**0.000**	**-4.242**	**0.000**
Area of soils with low infiltration rates	**3.837**	**0.000**	**3.788**	**0.000**	**4.219**	**0.000**	**4.243**	**0.000**	**4.173**	**0.000**
Human Population Density	-1.268	0.205	-1.268	0.205	-1.268	0.205	-1.268	0.205	-1.268	0.205
Cattle Density	-0.519	0.604	-0.519	0.604	-0.519	0.604	-0.519	0.604	-0.519	0.604
Pig Density	-1.007	0.314	-1.007	0.314	-1.007	0.314	-1.007	0.314	-1.007	0.314
Chicken Density	**2.401**	**0.016**	**2.401**	**0.016**	**2.401**	**0.016**	**2.401**	**0.016**	**2.401**	**0.016**
% Irrigated land in CCS	1.507	0.132	1.507	0.132	1.507	0.132	1.507	0.132	1.507	0.132
% Land fertilized by manure in CCS	-0.519	0.604	-0.519	0.604	-0.519	0.604	-0.519	0.604	-0.519	0.604

For this analysis, we are interested in the impact that a buffer radius size has on the results of a t-test, not the connotation of the results on well water quality. The t-test results displayed in Table 4.1 demonstrate that varying buffer size has little impact on returned *t*-values or *p*-values. The direction of association between independent and dependent variables (signified by the positive or negative value of the *t*-values) are the same for each variable, and the *t*-values are similar, no matter the buffer size. The same variables show significant group differences across all five buffer radii: agricultural land, soils with high infiltration rates, soils with low infiltration rates, and chicken density. Further, the *p*-values (and t-values) for each test are almost identical regardless of the area measured, differing only in the agricultural land category. The identical results suggest that at this time it would be logical to apply a static buffer size to analyze all variables. In reality, each environmental characteristic has a different influence on mediating the transport and survival of *E. coli*. In future research it may be useful to further investigate each variable[21] to determine whether assigning buffer radii of varying sizes to different environmental characteristics (based on their individual reported zones of influence on well water quality) may be a more suitable representation of the effects of local land characteristics on well water quality.

In our comparison, the 300m buffer radii return the most significant results overall, as this buffer size returns the smallest p-value (0.003) in the t-tests of the agricultural land data. A 300m radius allows the variability of continuous data layers to

[21] By conducting specific literature reviews, and/or applying alternative statistical tests to these data.

be captured, yet it is small enough to allow for efficient data processing in the GIS. Further, a 300m radius is supported repeatedly in current literature as being within the range of appropriate buffer sizes to delineate a WHPA using the arbitrary fixed radius method. For these reasons, buffers with 300m radii were placed around all wells in the four well sample groups for the selection and measurement of the environmental data in the main analyses.

4.2 Environmental Characteristic Comparisons

Once environmental data were selected with the 300m buffer zones, the resolved tables containing water test results and environmental characteristic measurements were imported into SPSS for analysis. First, independent samples t-tests were carried out to assess whether the mean values of environmental variables differ significantly between clean and contaminated wells. Secondly, bivariate logistic regressions were performed to test whether associations exist between *E. coli* presence and individual environmental variables. Multivariate logistic regressions were applied in an attempt to model *E. coli* or TC presence as a function of multiple environmental characteristics. Lastly, logit loglinear analyses were run in further exploration of inferential modeling as applied to this research problem.

4.2.1 Independent Samples T-tests

Similar to the buffer comparisons described in the previous section, independent samples t-tests were performed for 26 variables of interest in all four sample groups to compare land characteristics surrounding clean wells to contaminated wells. These tests compare the mean values of all dependent variables, grouped by *E. coli* presence/absence or TC presence/absence, in order to assess whether group differences are due to chance alone. These tests will address questions related to the hypotheses, such as: is agricultural land area more prevalent near contaminated wells? Or, is carbonate bedrock more prevalent near contaminated wells? The results of the *t*-tests are listed in Table 4.2

on the following page. In the *E. coli* sample groups, significant results with *positive* associations were returned for the following variables:

- Agricultural land
- Soils with low infiltration rates
- Geology with medium/variable permeability
- Carbonate bedrock
- Non-moraine material
- Sheep and lamb density

Significant results with *negative* associations were returned for the following variables:

- Developed land
- Soils with high infiltration rates
- Highly permeable geology
- Moraine material
- Human population density
- Intensive tilling of land
- Irrigated land
- Non-carbonate bedrock

The results of the t-tests for the TC sample groups were similar to those of the *E. coli* samples, except for positive associations with ancient bedrock (mesoproterozoic), and untilled land. The charts in Fig. 4.1 and Fig. 4.2 provide a graphical comparison of the mean values of eight selected environmental characteristics, grouped by *E. coli* presence/absence and TC presence/absence.

Table 4.2: T-test Results – E.coli and TC, 2003 & 2004

Comparison of Pilot data (2003), and Surveillance Data (2004)
with the Independent Samples T-test

Water Test Data — Grouping variable: ***E. coli* presence/absence**

	300m Buffer Radius	Pilot Study data - 2003 (n = 3718 (1859 pos, 1859 neg))				Surveillance data - 2004 (n = 5558 (2779 pos, 2779 neg))			
	Variable	Assoc	t	df	Sig (2-tailed)	Assoc	t	df	Sig (2-tailed)
1. Landcover	Area of Agricultural Land	**pos**	**2.655**	**3716**	**0.008**	**pos**	**4.266**	**5556**	**0.000**
	Area of Densly Forested Land	neg	-0.824	3715	0.410	**neg**	**-2.499**	**5556**	**0.012**
	Area of Mixed Forested Land	pos	1.872	3716	0.061	neg	-1.022	5556	0.307
	Area of Mixed Wetlands	neg	-1.782	3716	0.075	pos	0.008	5535	0.994
	Area of Developed Land	**neg**	**-5.004**	**3716**	**0.000**	**neg**	**-5.800**	**5556**	**0.000**
2. Soils	Area of Soils with High Infiltration rates	**neg**	**-4.586**	**2928**	**0.000**	**neg**	**-5.083**	**4647**	**0.000**
	Area of Soils with Low Infiltration rates	**pos**	**4.554**	**2933**	**0.000**	**pos**	**6.503**	**4674**	**0.000**
3. Surficial Geology	Area of High/Med-high Permeability	**neg**	**-7.136**	**3586**	**0.000**	**neg**	**-7.591**	**5389**	**0.000**
	Area of Medium/Variable Permeability	**pos**	**5.371**	**3586**	**0.000**	**pos**	**5.388**	**5389**	**0.000**
	Area of Low/Med-Low Permeability	pos	2.103	3585	0.360	pos	2.572	5389	0.100
4. Moraines	Area of Moraine Material	**neg**	**-2.952**	**3716**	**0.003**	**neg**	**-2.856**	**5556**	**0.004**
	Area of Non-Moraine Material	**pos**	**2.801**	**3716**	**0.005**	**pos**	**3.406**	**5556**	**0.001**
5. Census 2001	Human Population Density	**neg**	**-2.806**	**3703**	**0.005**	**neg**	**-3.882**	**5541**	**0.000**
6. Agricultural Census 2001	Cattle Density	neg	-0.143	3706	0.886	**neg**	**-1.302**	**5556**	**0.003**
	Pig Density	neg	-1.484	3716	0.138	**neg**	**-2.005**	**5556**	**0.045**
	Chicken Density	pos	1.619	3716	0.106	pos	0.947	5556	0.344
	Sheep and Lamb Density	pos	0.966	3713	0.334	**pos**	**2.139**	**5548**	**0.032**
	% Land tilled intensively in CCS	**neg**	**-2.323**	**3716**	**0.020**	**neg**	**-3.019**	**5556**	**0.003**
	% No till land in CCS	pos	0.431	3716	0.666	pos	1.09	5556	0.276
	% Irrigated land in CCS	**neg**	**-2.054**	**3716**	**0.040**	**neg**	**-2.398**	**5556**	**0.016**
	% Land fertilized by manure in CCS	neg	-0.723	3715	0.470	neg	-1.713	5556	0.087
7a. Bedrock Material	Carbonate Bedrock	**pos**	**2.228**	**3711**	**0.026**	**pos**	**4.804**	**5544**	**0.000**
	Non-carbonate Bedrock	**neg**	**-2.311**	**3711**	**0.021**	**neg**	**-4.909**	**5544**	**0.000**
7b. Bedrock Age	Youngest (Paleozoic)	neg	-0.420	3711	0.966	pos	1.838	5544	0.066
	Middle (Neo to Meso Proterozoic)	neg	-0.985	3704	0.325	neg	-1.600	5544	0.110
	Oldest (Mesoproterozoic)	pos	1.123	3711	0.261	neg	-1.188	5544	0.235

Water Test Data — Grouping variable: **Coliform presence/absence**

	300m Buffer Radius	Pilot Study data - 2003 (n = 23504 (11752 pos, 11752 neg))				Surveillance data - 2004 (n = 31850 (15925 pos, 15925 neg))			
	Variable	Assoc	t	df	Sig (2-tailed)	Assoc	t	df	Sig (2-tailed)
1. Landcover	Area of Agricultural Land	**pos**	**6.875**	**23502**	**0.000**	**pos**	**7.301**	**31848**	**0.000**
	Area of Densly Forested Land	neg	-0.132	23502	0.895	neg	-1.508	31848	0.132
	Area of Mixed Forested Land	neg	-0.817	23502	0.414	neg	-0.511	31848	0.609
	Area of Mixed Wetlands	pos	1.174	23502	0.895	pos	1.286	31848	0.198
	Area of Developed Land	**neg**	**-8.850**	**23502**	**0.000**	**neg**	**-10.647**	**31848**	**0.000**
2. Soils	Area of Soils with High Infiltration rates	**neg**	**-3.315**	**18127**	**0.001**	pos	1.150	25335	0.250
	Area of Soils with Low Infiltration rates	**pos**	**6.099**	**18127**	**0.000**	**pos**	**5.556**	**25628**	**0.000**
3. Surficial Geology	Area of High/Med-high Permeability	**neg**	**-4.486**	**14016**	**0.000**	**neg**	**-2.150**	**18562**	**0.032**
	Area of Medium/Variable Permeability	pos	0.757	13897	0.449	pos	0.314	12030	0.754
	Area of Low/Med-Low Permeability	**neg**	**-2.262**	**13681**	**0.024**	neg	-0.728	19654	0.467
4. Moraines	Area of Moraine Material	**neg**	**-2.274**	**23502**	**0.023**	**neg**	**-2.845**	**31848**	**0.004**
	Area of Non-moraine Material	**pos**	**2.335**	**23502**	**0.020**	**pos**	**2.572**	**31848**	**0.010**
5. Census 2001	Human Population Density	**neg**	**-5.774**	**23425**	**0.000**	**neg**	**-4.912**	**31770**	**0.000**
6. Agricultural Census 2001	Cattle Density	pos	0.828	23502	0.408	**neg**	**-3.234**	**31848**	**0.001**
	Pig Density	pos	0.082	23502	0.935	**neg**	**-4.148**	**31848**	**0.000**
	Chicken Density	**pos**	**2.961**	**23502**	**0.003**	pos	1.497	31848	0.134
	Sheep and Lamb Density	**pos**	**2.844**	**23495**	**0.004**	**pos**	**4.588**	**31848**	**0.000**
	% Land tilled intensively in CCS	pos	0.333	23498	0.739	**neg**	**-17.453**	**31848**	**0.000**
	% No till land in CCS	**pos**	**3.655**	**23502**	**0.000**	**pos**	**2.462**	**31848**	**0.014**
	% Irrigated land in CCS	**neg**	**-2.638**	**23502**	**0.008**	**neg**	**-3.969**	**31848**	**0.000**
	% Land fertilized by manure in CCS	pos	0.680	23495	0.497	**neg**	**-3.602**	**31848**	**0.000**
7a. Bedrock Material	Carbonate Bedrock	**pos**	**6.587**	**23488**	**0.000**	**pos**	**14.211**	**31818**	**0.000**
	Non-carbonate Bedrock	**neg**	**-6.573**	**23488**	**0.000**	**neg**	**-14.255**	**31818**	**0.000**
7b. Bedrock Age	Youngest (Paleozoic)	neg	-0.821	23488	0.412	pos	1.200	31818	0.230
	Middle (Neo to Meso Proterozoic)	neg	-1.686	23488	0.092	**neg**	**-2.803**	**31818**	**0.005**
	Oldest (Mesoproterozoic)	**pos**	**3.639**	**23488**	**0.000**	**pos**	**2.060**	**31818**	**0.039**

Figure 4.1: Comparison of Mean values, *E. coli*, 2003

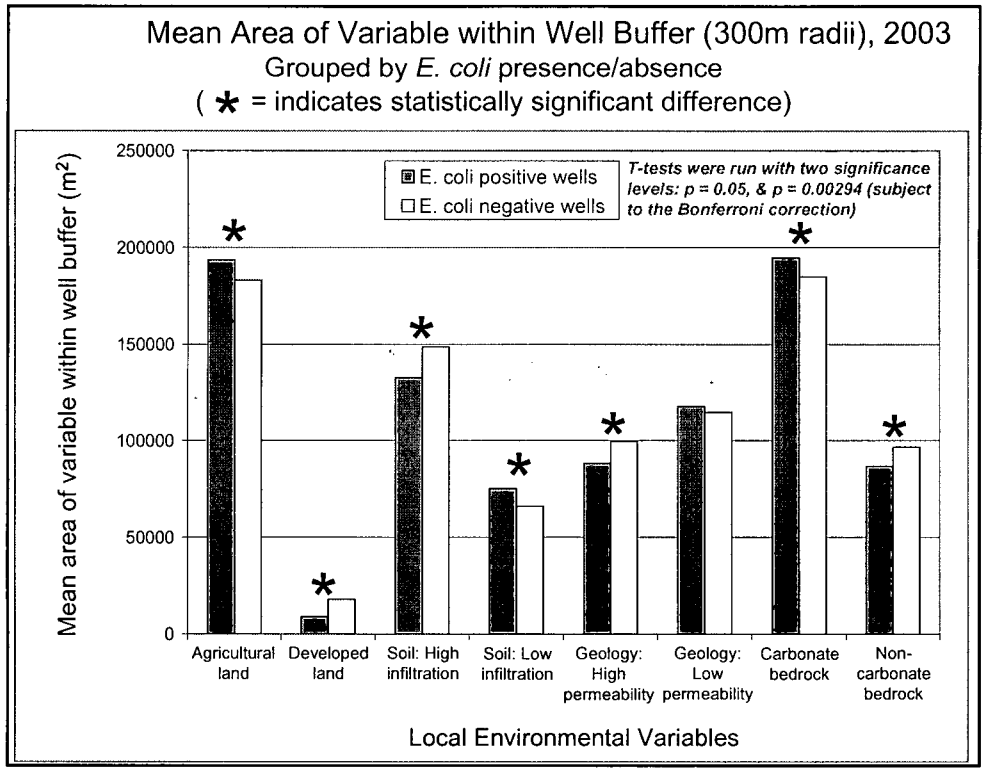
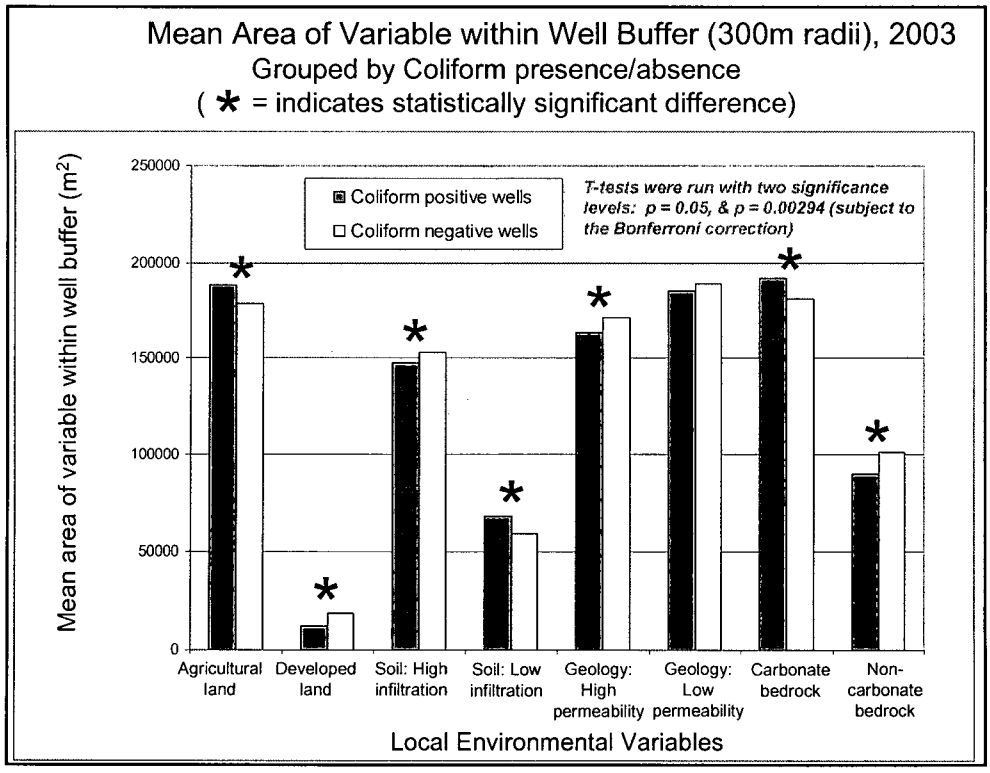


Figure 4.2: Comparison of Mean Values, TC,



These visual summaries demonstrate the clear differences in mean values of environmental variables between clean and contaminated wells. It is interesting to note that the mean values of developed land are lower than the mean values of agricultural land. This discrepancy reflects both the high prevalence of farming and agriculture in southwestern Ontario in comparison to the lower overall amount of developed land. Further, in developed areas where municipalities generally supply drinking water, there are fewer private wells per capita than in rural and agricultural regions, where populations rely more heavily on private wells.

The test results displayed in Table 4.2, and Figs. 4.1 and 4.2 are consistent – the same variables show significant differences across both TC and *E. coli* samples and across two years. However, when many statistical tests are performed, there is a possibility that some results may be erroneous. A *p*-value of 0.05 indicates that there is a 95% chance that the results are statistically significant. 26 t-tests were performed for each sample to obtain the aforementioned results, so there is a possibility that the outcome of at least 1 of the 26 tests is the result of type I or type II errors. Accordingly, the Bonferroni correction is a multiple-comparison correction that can be applied in a situation where many tests are performed simultaneously. I applied the correction to the thirteen variables that returned significant results in the first round of t-tests. To obtain a new confidence level, alpha was simply divided by the number of variables tested:

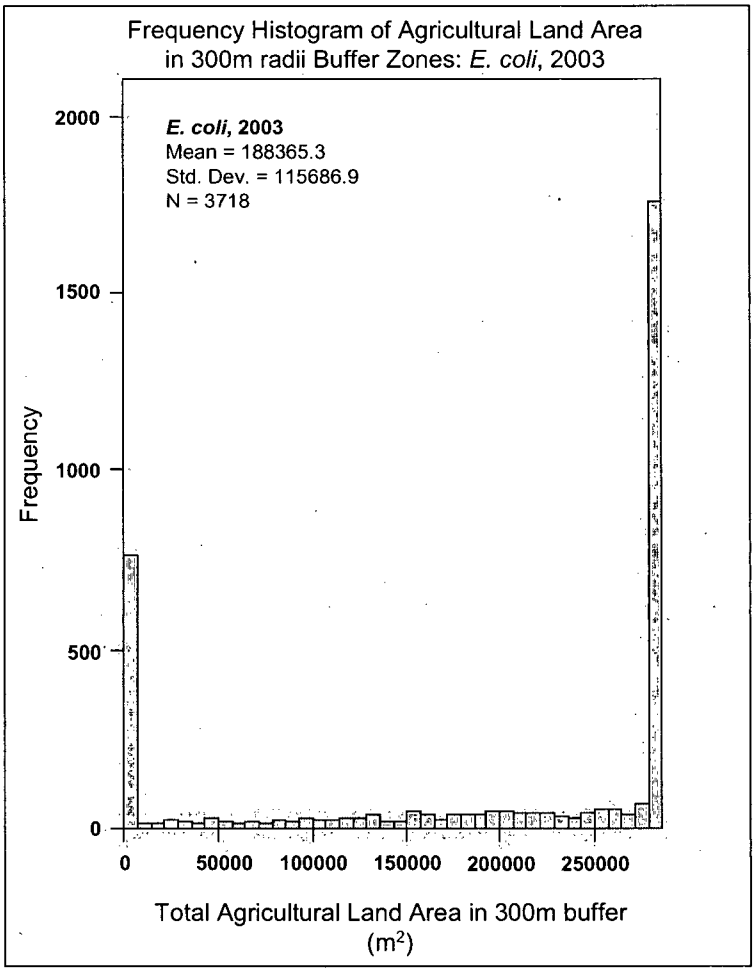
$$\alpha / n = \alpha_\beta$$
$$0.05 / 13 = 0.003846$$
$$1 - 0.00294 = 0.996153$$

Therefore, the new confidence interval = 99.62

The t-tests were re-examined with the new (higher) confidence interval ensuring that out of thirteen tests, the overall chance of making a type I error is still less than 5%. The results of the correction show that t-test results returned from nine of the thirteen variables continue to reject the null hypothesis. The 4 variables that fail to return significant results under the Bonferroni correction are: land tilled intensively, irrigated land, carbonate and non-carbonate bedrock.

Another method to improve the validity of the t-tests is to ensure the fundamental assumptions of the tests are not violated by the input data, potentially rendering the results incorrect or misleading. The t-test requires input data to follow the normal distribution. However, due to the nature of the sample selection (buffers) and the nature (resolution, variability) of the original data, the frequencies for 17 out of 26 environmental variables are bimodal. This is the case because a well buffer zone is usually either composed of 100% of an environmental characteristic, or 0%, and it is not as common for variability to occur within a buffer (owing in part to the 300m buffer size used). The histogram in Fig. 4.3 shows the bimodal nature of the frequency distribution of agricultural land, in the 2003 *E. coli* sample.

Figure 4.3: Histogram of Agricultural Land Values, *E. coli*, 2003



When assumptions of parametric tests are violated by input data, nonparametric tests can be employed instead of their parametric counterparts, in this case, the Mann-Whitney U Test. All variables were re-tested using the Mann-Whitney test (with the Bonferroni correction) in SPSS, and nine of the thirteen continued to return significant results. The four variables for which returned results did not reject the null hypothesis are: soils with low infiltration rates, geology with medium/variable permeability, and both carbonate and non-carbonate bedrock.

The results of these tests consistently return significant differences between clean and contaminated wells in terms of these surrounding variables:

- Agricultural land
- Developed land
- Soils with high infiltration rates
- Surficial Geology with high permeability
- Moraine material
- Non-moraine material
- Human population density

4.2.2 Bivariate Logistic Regressions

The results of the t-tests show that certain variables are significantly more prevalent near contaminated wells, as opposed to clean wells. However, the t-test does not indicate an *association* between the dependent and the independent variables. Regression analysis attempts to model the relationship between two variables indicating the extent to which variables are associated with one another. Logistic regression is part of a group of statistical models called generalized linear models (GLM). Where linear models predict the best line to fit a series of continuous data points, logistic regression allows for the prediction of a probability of categorical outcomes, in this case discrete outcomes (i.e. *E. coli* presence versus. *E. coli* absence), based on either continuous or categorical data.

Logistic regression has a historical place in both health research and hydrogeology, as noted by Gardner and Vogel (2005, p. 345): "logistic regression has been used extensively in the health sciences since the late 1960s to predict a binary response from explanatory variables (e.g. Truett, Cornfield, and Kannel, 1967; Hosmer and Lemeshow, 2000), and more recently in the environmental sciences to assess multiple variables that may explain the occurrence of contamination in ground water." Applied to our research question, logistic regression generates coefficients to calculate a logit transformation (the natural log odds of the probability of *E. coli* appearing in a well) based on environmental characteristics within a 300m range of a well. The formula used for logistic regressions is as follows:

$$\text{logit}(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

Where

p is the probability of *E. coli* presence

b_0 is the intercept

X is the area of an environmental variable within the well buffer zone

The logit transformation is defined as the logged odds:

$$\text{Odds} = \frac{P}{1-p} = \frac{\textit{probability of presence of E. coli}}{\textit{probability of absence of E. coli}}$$

and

$$\text{logit}(p) = \ln\left[\frac{p}{1-p}\right]$$

The null hypothesis is that a particular logit coefficient is zero (i.e. that the independent variable does not contribute to the outcome of the dependent variable). Logistic regression makes no assumption of the distribution of predictor variables, so bimodal data distributions are not problematic, and both categorical and continuous variables can be included in the analysis.

In SPSS, bivariate logistic regressions were individually run with the thirteen environmental variables that showed significant group differences in the t-tests. The tests returned two figures of interest for each variable: the Wald statistic and an R square (R^2) value. The Wald statistic tests the statistical significance of each component of the model to discern whether an effect exists between two variables. The results listed in Table 4.3 show that every variable tests significant in at least one of the samples, meaning that each independent variable significantly contributes to the prediction of the outcome of the dependent variable. The second figure of interest is the R^2 values.

Table 4.3: Bivariate Logistic Regression Results

	E. coli				Total Coliforms			
300m Buffer Radius	**2003 season**		**2004 season**		**2003 season**		**2004 season**	
Variable	Wald	Sig.	Wald	Sig.	Wald	Sig.	Wald	Sig.
Area of Agricultural Land	**7.029**	**0.008**	**18.102**	**0.000**	**47.124**	**0.000**	**53.161**	**0.000**
Area of Developed Land	**23.784**	**0.000**	**25.634**	**0.000**	**76.655**	**0.000**	**110.515**	**0.000**
Area of Soils with High Infiltration rates	**13.895**	**0.000**	**23.634**	**0.000**	10.981	0.001	1.325	0.250
Area of Soils with Low Infiltration rates	**5.158**	**0.023**	**41.626**	**0.000**	**37.050**	**0.000**	**30.798**	**0.000**
Geology - High/Med-high Permeability	**8.635**	**0.003**	**56.643**	**0.000**	**20.052**	**0.000**	**4.623**	**0.032**
Area of Moraine Material	**8.376**	**0.004**	**7.994**	**0.005**	**5.109**	**0.024**	**8.022**	**0.005**
Area of Non-Moraine Material	**7.623**	**0.006**	**11.334**	**0.001**	**5.437**	**0.020**	**6.602**	**0.010**
Human Population Density	**7.289**	**0.007**	**14.116**	**0.000**	**31.185**	**0.000**	**28.219**	**0.000**
Chicken Density	2.610	0.106	2.927	0.411	8.747	0.003	2.239	0.135
Sheep and Lamb Density	**0.931**	**0.334**	**4.557**	**0.033**	**8.075**	**0.004**	**20.974**	**0.000**
% Land tilled intensively in CCS	**5.375**	**0.020**	**9.069**	**0.003**	**0.111**	**0.739**	**28.219**	**0.000**
% No till land in CCS	0.186	0.666	1.188	0.276	**5.833**	**0.016**	**6.059**	**0.014**
Carbonate Bedrock	**4.971**	**0.026**	**22.933**	**0.000**	**43.273**	**0.000**	**199.994**	**0.000**
Non-carbonate Bedrock	**5.236**	**0.022**	**23.937**	**0.000**	**43.091**	**0.000**	**201.218**	**0.000**

Logistic regression does not actually return an R^2, however SPSS has a function that returns two derived measurements: the Cox and Snell R^2, and the Nagelkerke R^2, designed to mimic R^2 values in linear regression. Out of the 52 regressions run in SPSS, the highest R^2 value returned was 0.014, for the variable 'surficial geology with high permeability.' This means that the best model can only predict 1.4% of the variability displayed by the dependent variable (*E. coli* presence/absence), based on the amount of highly permeable surficial geology found in a buffer zone. The remaining variables all returned lower R^2, so although the regressions determine that the thirteen variables singled out by the t-test results are significant components of a bivariate logistic regression model, a model based the coefficients of just one environmental variable clearly does not predict the presence or absence of *E. coli* or TC in our dataset.

4.2.3 Multivariate Logistic Regressions

A bivariate analysis assesses whether the independent and dependent variables covary, but "provides insufficient evidence for causality. Multivariate analysis may provide better, though still incomplete, evidence for causality because it allows checking for possible spuriousness" (Hamilton, 1992, p. 29). Adding multiple independent variables to the model can help identify whether covariation between independent and dependent variables found by the bivariate regressions do not result from the effects of some other environmental characteristic in our GIS[22]. Multivariate regression analysis also helps to account for the inherent complexity of the multifaceted environmental processes responsible for fecal coliforms appearing in well water.

The problem of selecting the 'best' number and combination of independent variables to use in a multiple regression is a longstanding discussion among statisticians (McQuarrie and Tsai, 1998). Including too few variables in the model may result in misspecification bias and an over-simplified model, yet including many variables could cause spurious results and an inflated R^2. The method of model construction used in this analysis is a forward-inclusion regression, where the investigator first includes the single X variable that has the largest effect on the dependent (Y) variable. At each subsequent step, one more X variable is introduced, until the model produces the largest possible R^2. In order to select independent variables for inclusion to the multivariate model, we turn to the previous results of the independent samples t-tests and the binary regressions. These tests indicate that thirteen of the variables are associated with the presence or absence of *E. coli* and TC in private wells.

In constructing a multivariate regression, it is preferable to add variables to the model based on the strength of association between X and Y. A test called discriminant function analysis (DFA) is useful for this purpose, as it ranks the relative importance of each independent variable in predicting the dependent variable based on the Wilks

[22] Although it does not help us in any way if the model's spuriousness is a result of an environmental characteristic not included in our analysis.

lambda statistic[23]. In SPSS, DFA was performed on all four samples, with the results of the top six variables[24] listed in Table 4.4 in canonical order. Each sample has a different ranking pattern, although many of the same variables appear in the top 6. Regions of

Table 4.4: Discriminant Function Analysis

Discriminant Function Analysis

***E.coli* 2003**

Rank	Variable	Wilks Lambda
1	High Infiltration Soil	0.995
2	Developed Area	0.992
3	Moraine Area	0.988
4	High Permeability Geology	0.986
5	Tilled Soil	0.984
6	Agricultural Area	0.981

TC 2003

Rank	Variable	Wilks Lambda
1	Developed Area	0.995
2	Untilled Soil	0.994
3	Low Permeability Geology	0.994
4	Non-Moraine Area	0.993
5	Low Permeability Soil	0.992
6	Carbonate Bedrock	0.991

***E.coli* 2004**

Rank	Variable	Wilks Lambda
1	High Permeability Geology	0.984
2	Low Infiltration Soil	0.980
3	Tilled Soil	0.975
4	Low Permeability Geology	0.972
5	Agricultural Area	0.967
6	Non-Carbonate Bedrock	0.964

TC 2004

Rank	Variable	Wilks Lambda
1	Tilled Soil	0.995
2	Moraine Area	0.993
3	Developed Area	0.990
4	Low Infiltration Soil	0.989
5	Carbonate Bedrock	0.988
6	High Infiltration Soil	0.987

developed land appear to be comparatively strong predictors of the absence of both *E. coli* and TC, and both soil infiltration rates and geological permeability (high and low) figure prominently in all four samples. The results of the DFA indicate the order in which variables should be included in the model, but the presence of multicollinearity among independent variables must still be tested for.

The presence of multicollinearity (or the lack of independent variation between data) leads to unreliable or misleading coefficient estimates (Hamilton, 1992) and an imprecise model. For example, among our data, the soils layer illustrates the infiltration

[23] In this analysis, I added the variables to the model one by one based on the strength of association between X and Y. In other models (i.e. forward-inclusion stepwise regression), the software uses algorithms to test different combinations of independent variables one by one, and automatically selects the order of variable inclusion based on algorithm results. In future research, a stepwise regression could be applied to these data (for further discussion on analytical possibilities see section 5.2.2).

[24] Although DFA was performed on all 13 variables, only the top six are listed in table 4.4. The number six is an arbitrary cutoff, although after the top six variables the Wilks Lambda values noticeably plateau.

rate of the top one meter of surficial geologic materials, whereas the geology layer describes the permeability of all geologic materials overlying bedrock, including the same top one meter of surficial materials. To test whether multicollinearity is a problematic factor among our data, the Variance Inflation Factor (VIF) was examined for all thirteen variables in SPSS. The output of the test is a VIF score with a range from one to infinity. The most commonly used VIF threshold is ten, and any value above this number indicates the presence of collinearity among variables. Among the thirteen variables tested, the highest VIF returned is 69.1, for both carbonate and non-carbonate bedrock. Moraine and non-moraine areas have the second highest VIF of 23, and the remainder of the variables had VIF scores below the threshold of ten, near one. The reason for the high VIF values in the geology and moraine layers is because these two layers are both dichotomous categories (i.e. in any point in the study area, there is either a moraine or there is not), therefore collinear as they directly reflect one another. The findings of these collinearity diagnostics show that when entering variables into the model, entering both carbonate and non-carbonate bedrock data, or moraine/non-moraine data into the same model may cause confounding.

The results of the DFA indicate the order variables should be entered into the model, and the results of the VIF confirm that the most of the independent variables are not collinear. Multiple logistic regressions were performed on each of the four samples, adding each variable into the model separately in the order prescribed by the DFA analysis. A VIF analysis was performed alongside the regressions, the results of both listed in Table 4.5. The highest VIF returned was 2.1, well below the threshold of ten. As indicated by the R^2 values, the models do not have strong inferential power, with the most favourable model explaining only 4.6% of the variability of the dependent variable[25]. Berry and Feldman (1985) caution the use of R^2 as a sole measure of goodness-of-fit, as "R^2 will always increase (to some degree) when new variables are added to the equation, even when they may have no effect on the dependent variable. In fact, as the number of independent variables (k) gets close to the number of cases in the sample (n), R^2 will necessarily get close to 1.0 (p. 16)." Calculating an adjusted R^2 is one

[25] See Appendix 6 to find the formula for the strongest model.

Table 4.5: Multiple Logistic Regression Results

Results from Multiple Logistic Regressions

***E.coli* 2003**

# entered into model	Variable	Cox and Snell R^2	Nagel-kerke R^2
1	High Infiltration Soil	0.005	0.006
2	Developed Area	0.008	0.001
3	Moraine Area	0.012	0.016
4	High Permeability Geology	0.014	0.019
5	Tilled Soil	0.016	0.021
6	Agricultural Area	0.017	0.023

VIF <= 1.384

TC 2003

# entered into model	Variable	Cox and Snell R^2	Nagel-kerke R^2
1	Developed Area	0.003	0.004
2	Untilled Soil	0.004	0.005
3	Low Permeability Geology	0.005	0.006
4	Non-Moraine Area	0.007	0.009
5	Low Permeability Soil	0.007	0.009
6	Carbonate Bedrock	0.007	0.009

VIF <= 1.172

***E.coli* 2004**

# entered into model	Variable	Cox and Snell R^2	Nagel-kerke R^2
1	High Permeability Geology	0.011	0.014
2	Low Infiltration Soil	0.020	0.027
3	Tilled Soil	0.027	0.037
4	Low Permeability Geology	0.029	0.040
5	Agricultural Area	0.032	0.043
6	Non-Carbonate Bedrock	0.035	0.046

VIF <= 2.135

TC 2004

# entered into model	Variable	Cox and Snell R^2	Nagel-kerke R^2
1	Tilled Soil	0.001	0.001
2	Moraine Area	0.001	0.001
3	Developed Area	0.005	0.006
4	Low Infiltration Soil	0.007	0.010
5	Carbonate Bedrock	0.010	0.014
6	High Infiltration Soil	0.012	0.016

VIF <= 1.082

method of solving this problem, as the equation adjusts R^2 to compensate for number of variables added to the model. For the model with the highest R^2 of 0.046, the adjusted R^2 is 0.045. So although the R^2 values increased as variables were added, even the most favourable adjusted R^2 value is so low that re-running the DFA and VIF tests in order to add the 7 additional variables is redundant.

4.2.4 Logit Loglinear Analyses

As the low R^2 values demonstrate, the bivariate and multivariate logistic regression models do not account for the variability of *E. coli* and TC in our samples. One final method of analysis (involving data transformations) was applied in an effort to improve upon the inferential capacity of previous models. Logit loglinear analysis is a multivariate method for analyzing relationships between variables through data crosstabulation, when both X and Y variables are categorical, nominal, or ordinal in nature. Unlike the regressions previously explored, loglinear analysis involves the

transformation of continuous data to ordinal categories. Data transformations are defined as the mathematical modification of the values of a variable (Osborne, 2002), and are commonly used tools in statistics to remedy modeling problems related to outliers, homoscedasticity, and non-normality.

The crosstabulations performed as part of loglinear analysis renders it a labour-intensive process, and experts suggest that the analysis should include five or less variables (Hosmer et al., 2000). This analysis was performed solely on the 2003 *E. coli* sub-sample, from which five independent variables were selected on the basis of the rank order prescribed by the DFA (see section 4.2.4 for details). These variables are: high infiltration soil type, developed area, moraine area, geology with high permeability, and tilled soil. These data are continuous in nature, and represent either an area value or a rate. I used the 'recode' function in SPSS to reclassify each variable into nine different categorical combinations, ranging from two to ten categories per variable. Then, logit loglinear analyses were performed nine times (once for each categorical breakdown) in order to determine the transformation that yields the strongest results. Although the results were almost identical among the nine different categorical combinations, transforming the data into three categories returned the highest measure of association. Loglinear analyses return entropy and concentration values (similar to R^2), and the highest values returned overall were 0.036 and 0.048, respectively. These values are slightly higher than the R^2 found in the bivariate and multivariate regressions, yet are still too small to have any inferential power.

4.3 Summary

Although lacking inference, the statistical analyses described in this chapter show results that address questions raised in our hypotheses. Multiple tests were applied to the same data, and the results demonstrate that specific variables are statistically important to the presence of *E. coli* and TC. The results of independent samples t-tests, Mann-Whitney U-tests, bivariate logistic regressions, and discriminate function analysis all

show that *positive* relationships exist between *E. coli* and TC presence in tested private wells and the following variables:

- agricultural land
- soil with low infiltration rates

The same analyses maintain that *negative* relationships exist between *E. coli* and TC presence in tested private wells and the following variables:

- developed land
- human population density
- soil with high infiltration rates
- surficial geology with high permeability
- areas of moraine material
- tilled soils

Moreover, these analyses serve as a useful exploration of the role of applied statistics in characterizing and estimating microbial groundwater quality based on local environmental characteristics. Using the information presented here as a starting point for discussion, the following chapter addresses how this research can be improved upon. Specifically, problems relating to the lack of inference in the multivariate models (such as autocorrelation, coarse-resolution data, and the lack of inclusion of temporal and climactic data) will be discussed.

5. DISCUSSION AND CONCLUSION

5.1 Discussion

Our hypotheses, methodologies, and study area are together relatively understudied, therefore the spatial analyses reported here are exploratory and experimental in nature. Although the inferential power of the models are weak, the results of t-tests and bivariate regressions identify a list of environmental characteristics associated with contaminated wells. Many of the results are consistent across sample groups, seasons, and statistical tests, and are in agreement with our hypotheses. The purpose of this chapter is to discuss key findings reported in chapter 4, and identify limitations in data and methods that may have contributed to a lack of inferential power in the multivariate models. I also offer suggestions for future research based on our experiences and results.

5.1.1 Key Findings

In the opening analysis, well buffer zones of five different radii were examined to determine an appropriate buffer size for the selection of local land characteristics. As reported in section 4.1, varying the buffer radii has a negligible impact on t-test results. The 300m buffer returns the most significant results (lowest p-values) overall, and is within the 200m-900m range of buffer sizes utilized in similar studies (see section 3.4.1). Our decision to use a 300m buffer zone was corroborated (after the fact) by a July 2006 article on private well water by Wang et al. The investigators applied similar geo-statistical methods in their study of agriculture-derived nitrate concentrations on the North China Plain, and tested buffer radii ranging from 200m-2000m. The study reports that a 400m radius returned optimal results in a comparison of land characteristics to nitrate concentrations in groundwater. Although the geological characteristics, climate, and agricultural practices of the North China Plain most likely differ to our study region, the findings of Wang et al.'s study support our results.

The 300m buffer zones were used to overlay and select environmental variables local to sampled wells in order to answer questions related to the hypotheses. To review, we hypothesize that the following environmental characteristics are associated with *E. coli* presence in the samples of private well water:

- Agricultural land use
- Agricultural animal densities (Cattle, sheep, poultry, hens)
- Low human population densities (rural areas)
- Soils with low water infiltration rates
- Surficial geology with low permeability
- Bedrock composed of carbonate materials
- Ancient (mesoproterozoic) bedrock
- Non-moraine material
- Land subject to manure application
- Lack of tilled soil
- Irrigated land

The t-tests show that seven of these eleven variables are significantly more prevalent near *E. coli* contaminated wells as compared to clean wells. The exceptions are: animal density, ancient bedrock, manure application, and irrigated land. The results of the bivariate logistic regressions show that the same seven variables are significantly associated with *E. coli* presence in wells. The Bonferroni correction and the Mann-Whitney U-Test emphasize these results with the exception of the carbonate bedrock layer.

These seven variables consistently return significant results in four different statistical tests across two separate sampling seasons. Logical complimentary patterns exist in the results. For example, low infiltration soils and surficial geology with low permeability both show positive associations with *E. coli* and TC. These two layers originate from different sources, yet describe similar environmental phenomena. Likewise, agricultural land shows a positive association with *E. coli*, while areas of developed land and high population density show negative associations with well contamination. These consistencies add strength to the argument that wells located

within 300m of any of the noted 7 variables may be at higher risk of contamination than other wells.

Four layers are not associated with *E. coli* presence: manure application, animal densities, amount of irrigated land, and bedrock age. Three of these four variables originate from the agricultural census, and consist of data aggregated to municipal CCS tracts. This suggests that in our analyses, regional-level data are not as important as local data to the presence of *E. coli* in individual wells. In future studies, mixed (multilevel) modeling methods could allow for the analyses of the effects of different sample groupings on the model results (see section 5.2.2 for further discussion). It would be useful to compare agricultural census data with water test results aggregated to the same geographic boundaries for analysis. Instead of comparing individual wells to CCS data, the rate of well contamination per CCS could be used as the dependent variable[26]. Conversely, if information on anthropogenic processes had been collected at the same local level as environmental data, there is a possibility that the analytical outcomes would be more definitive (see section 5.2.1). A further discussion on the effects of scale in geographic analyses will continue in section 5.2.2. The lack of significant results among census data may also indicate that anthropogenic effects are less important to well water quality in our study area than hydrogeologic characteristics. In continuation with this idea, it is interesting to observe that the only census variable found to be associated with *E. coli* is the proportion of the CCS area composed of tilled land. Although tillage is a human-induced process, it impacts the hydrogeology of a region by encouraging the natural filtration of microbes from groundwater through soil disturbance.

One surprising result returned by the agricultural census data, based on the analytical methods used in this research, is that agricultural animal densities at the CCS level are not associated with *E. coli* presence. Although a majority of evidence suggests otherwise, the study by Conboy and Goss cited earlier (2001) corroborates this finding. In their household-level survey of wells in Southern Ontario, the researchers found that

[26] The 2nd half of the ARO Geospatial team, based in Quebec, is using this approach for their regional analysis of AR E. coli in well water.

> "many low risk wells also housed livestock indicating that the presence of livestock alone did not result in a well being vulnerable to contamination. Logistic regression analysis did not find the presence of livestock to be an important predictor of well vulnerability" (2001, p. 18).

Yet, Conboy and Goss found that locally, manure application was an important indicator of well contamination, whereas our results (based on regional data) did not. An extensive body of literature reports that manure storage sites, septic lagoons, and feedlot sites are key sources of environmental *E. coli* (Gagliardi and Karns, 2000; Macler and Merkle, 2004; Rahe, Hagedorn, McCoy, et al., 1978). On the basis of this knowledge, and our findings, it is possible that locations of mass storage of fecal matter are more important to *E. coli* presence in groundwater than the total number of animals living in a region. Because we do not have access to detailed (or regional) data on the location of feedlots, manure storage sites, and septic fields, there is a strong likelihood that we are neglecting to address important indicators of *E. coli* in well water. These data would permit a comparison of the geology of different manure storage sites, septic fields, and feedlots, to assess whether certain hydrogeologic characteristics mitigate the harmful effects of such hazards. This information would be fruitful from a landuse planning perspective.

In the case of *local* environmental variables, the only layer listed in the hypotheses that that did not return significant results is bedrock age. The oldest bedrock in the study region is found in the Canadian Shield, and is composed mostly of erosion-resistant materials such as granites, gneisses, metasedimentary and metavolcanic rocks. The insignificant t-test results indicate that in Southern Ontario, bedrock *age* is not as important as bedrock *material* in terms of providing protection to the contamination of bedrock aquifers by bacteria. Our results indicate that a well located in newer bedrock composed of carbonate material is at greater risk for *E. coli* contamination than a well located on the older non-carbonate materials of the Canadian Shield.

For the purposes of this research we are most interested in the results of the *E. coli* samples. However, the results returned by the two TC sample groups are telling. Every variable that returned significant results in the *E. coli* tests returned significant values in the TC tests. Three variables showed significance in the TC t-tests but not the *E. coli* t-tests: ancient bedrock, % irrigated land, and sheep and lamb density. This

difference in results is most likely due to the larger sizes (and statistical power) of the TC samples (n 2003 =23504, n 2004 =31850) as opposed to the *E. coli* samples (n 2003 =3718, n 2004 =5558); because TC are more prevalent in the natural environment. However, there is a strong similarity in the results of the *E. coli* and TC samples, across both sample years. Although not mentioned in the hypotheses, these findings are in line with current literature as reports consistently show that the TC is reliable indicator of fecal coliform bacteria and *E. coli* in groundwater (Atherholt et al., 2003; US EPA, 2006, Drinking Water Pathogens and Their Indicators, ¶3). Our results demonstrate that the environmental characteristics associated with *E. coli* are also associated with the presence of TC in private well water, confirming TC an appropriate indicator for *E. coli* in future investigations into the landscape epidemiology of *E. coli*.

5.1.2 Limitations

The results of the t-tests and binary logistic regressions show certain variables are more prevalent near contaminated wells versus clean wells. However, the low R^2 values returned by the binary logistic regressions also demonstrate that data from one variable alone is not enough to predict the presence of *E. coli*. Although an individual environmental variable may be important to well water quality, the presence of *E. coli* in private well water requires a host of processes to occur concurrently (i.e. deposition of fecal matter on the surface, transportation to an aquifer, lack of filtration). Multivariate methods were applied in an attempt to model the complex biological, ecological, and spatial relationships between bacteria and the environment. Despite large sample sizes and multiple statistical analyses, none of the models returned an adjusted R^2 of higher than 0.045, meaning that the best model is only able to predict 4.5% of the variability of *E. coli* presence in private wells. The possible reasons for the low predictability of the models are many, but are rooted in the basic tenets of the modeling process.

The first assumption of logistic regression is that the true conditional probabilities are a logistic function of the independent variables. However, misspecification of the logistic function does not usually result in specification error, in comparison to using

alternative function choices (UCLA, 2006, Regression Diagnostics, ¶2), such as those based on a normal distribution (i.e. linear regression). The second possible violation of the tests' assumptions is the existence of multicollinearity among independent variables, which has been ruled out by testing the VIF of each of the variables separately (and together). Thirdly, we know that extraneous variables were not included, as the only variables added to the models showed significant results across both sample years, in both the t-tests and binary logistic regressions. Further, DFA was applied to ensure that the independent variables were added in hierarchical order of importance to the dependent variable.

This leaves us with two remaining assumptions: that no important variables are omitted from the model, and that the observations are independent. Both of these assumptions have likely been violated in our attempts, causing misspecification errors that have rendered the models ineffectual (Berry and Feldman, 1985). As described, the biogeography of *E. coli* is complex and can not be explained by a small handful of independent variables. In current literature, many variables in addition to those included in our models have been cited as being important to *E. coli* presence in groundwater, such as: climate, time, well characteristics, septic tank locations, manure storage sites, drainage basin topography, feedlot locations, groundwater flow, and aquifer characteristics. The reasons for the exclusion of important variables in the modeling process are fourfold:

- The data do not exist in digital format for input to the GIS
- The data exist but were not released for confidentiality purposes
- We have access to digital data, but time constraints and project boundaries restrict further analyses
- We may not be aware of certain variables of importance

A detailed discussion of causes and implications of missing data layers will be continued in section 5.2.1.

The second assumption of logistic regression that may be violated in our analyses is that our observations (well water tests) may be not spatially independent. Spatial

autocorrelation refers to patterns that exist in data collected in the vicinity of other samples. If a nonrandom pattern (i.e. clustering) is identified in the spatial arrangement of the samples, then spatial autocorrelation exists. The same can be said for samples clustered in time. Because clustered samples may bear a greater resemblance to each other than samples collected randomly over space and time, grouped data may cause model confounding. Two point pattern analyses can be applied to assess a sample for spatial autocorrelation: the average nearest neighbor test, and the Moran's I test. Both tests were applied to the 2003 and 2004 *E. coli* samples (using ArcToolbox in ArcMap v. 9.0). The results of the nearest neighbor analyses showed that *E. coli* positive wells were no closer together than *E. coli* negative wells. However, the Moran's I tests report that moderate clustering does exist in both samples[27].

The combination of spatial autocorrelation and, perhaps more importantly, omitted independent variables are the principal problems of the multivariate logistic models, logit loglinear analyses, and of this research in general. The limitations of this project are the following:

- A lack of important independent environmental variables, specifically:
    - Local agricultural data (feedlots, manure storage)
    - Hydrogeologic characteristics (aquifer boundaries, drift thickness)
    - Septic tank and outhouse locations
- Lack of attention to climate (rainfall events, temperature)
- Lack of attention to time
- Lack of attention to watershed characteristics (multi-scale analyses)
- The spatial (and possible temporal) autocorrelation of samples
- Lack of exact well locations (well points were geocoded to road networks)

With additional data and time, most of these limitations could be addressed. The potential for further analyses is unlimited, and unfortunately many of the questions raised here are beyond the scope of this MA thesis. However, it is imperative to identify project

[27] A Moran's I value near 1.0 indicates clustering, a value near -1.0 indicates dispersion, and a value near 0 indicates a random spatial distribution. Moran's I returned values ranging from 0.51 to 0.78 in both *E. coli* sample groups, indicating moderate clustering among samples.

constraints so that future work can continue where this research ends. In the following section, the above-listed limitations are discussed, with a view to offer direction for further investigations on this topic.

5.2 Suggestions for Future Research

5.2.1 Data

The quality of any analysis depends on the characteristics of the original data. Sub-standard data and accessibility issues are hallmark problems in GIS-related research. Unfortunately, high-quality digital data are difficult to obtain as a Canadian researcher - bureaucratic roadblocks have certainly affected the direction and outcomes of this research project. And, as mentioned above, some data simply don't exist or are not in digital format for input to a GIS. Data missing from our analyses can be broken down into three categories: sources of environmental *E. coli* (i.e. manure storage sites, feedlots, and septic fields), mediums through which *E. coli* travel to contaminate a private well (i.e. aquifer and well characteristics), and larger influential processes (i.e. climate and time).

Extensive research shows that wells located on or near agricultural land are at risk of contamination by *E. coli* because many agricultural activities are sources of environmental *E. coli*. The MNR landcover layer represents the location of all agricultural land in the study area, but is not specific to different types of agriculture. This is problematic and may introduce error into the analyses because not all agricultural activities encourage the release of *E. coli* into the environment. For example, fruit orchards and fallow fields would pose negligible risks to nearby wells, in comparison to livestock operations or even a field of corn or legumes[28]. It would be useful to have information on specific agricultural land uses in a finer-grain resolution than the currently available CCS-level data, and this may be a possibility as data collection via satellite

[28] Nitrates hinder the production of fruit in plants and are not typically applied to orchards or other fruit crops. Opposite to this, nitrates encourage the production of grains and legumes and are frequently applied to these crops.

imagery improves[29]. Statistics Canada publishes non-sensitive human demographic data at a fine resolution (DA-level), yet similar-scale agricultural landuse data are not available. This means that the public can know the human population of every DA in the country, but not the number of commercial animals that live in the same small areas.

From a public health perspective, information on the location of high-risk agricultural operations (like feedlots or manure storage sites) would be valuable. Research has shown that groundwater is at risk for contamination by *E. coli* and other fecal coliforms near feedlots (Olson, Miller, Rodvang, et al., 2005), defined as buildings or yards where a large number of animals are fattened for slaughter. OMAF possesses data on the locations of high-density animal operations in the Province; yet will not circulate it for public (or research) use. Similarly, MNR refused to release data locating septic drying beds, septic fields, sewage disposal, tile beds, and transfer stations for use in this project, despite evidence that human fecal waste contributes to *E. coli* in groundwater (George et al., 2004), and poorly-maintained septic systems have been identified as the cause of numerous outbreaks of waterborne disease in developed countries (Goss et al., 1998; Francey, Helsel, and Nalley, 2000). If these data ever come available in the future, the distance between a well to the nearest feedlot or septic field could be measured, and added to a logistic regression model as a continuous independent variable. The last source data that would be beneficial (but do not exist as far as I am aware) are non-agricultural animal densities, since fecal matter from wild animals and domestic animals such as horses, dogs, and cats also contribute to the contamination of groundwater (Mallin, Williams, Esham, et al., 2000).

It has been established that both topographic and subsurface characteristics affect the movement and survival of *E. coli* from the surface to underground drinking water sources. The relief of the natural landscape can contribute to well water quality, as landuse in an upland water catchement zone has a direct effect on downland water quality (Crowther et al., 2002; Hunter and McDonald, 1991). Topography affects both surface

[29] Environment Canada's current data on crop cover was collected by RADARSAT-1 (the first radar satellite system), and does not contain enough detail to identify crop types. However, RADARSAT-2, launched in 2005, is currently collecting detailed data to provide this information (Wagner, 2005).

and subsurface flow gradients, therefore influencing the spatial diffusion of fecal pollutants. The flow direction and bacterial loading of surface waters can also affect groundwater quality when natural exchange occurs between surface and groundwaters[30] (Brunke and Gonser, 1997). Most importantly, water drainage is compartmentalized by watersheds, and water quality in one watershed often differs to that of an adjacent basin. In future work, Digital Elevation Models (DEM) could be used to delineate watershed boundaries in order to compare land characteristics and well test results by basin. Other researchers have found multi-scale analyses (incorporating watershed boundaries) fruitful in determining land uses associated with fecal coliforms in source waters (Kistemann et al., 2001). Data describing subsurface aquifer boundaries would be useful for the same reasons.

The 'depth to bedrock' of overlying surficial geology is another hydrogeological factor important to well water quality, and could easily be introduced to the analyses. We would expect to see a higher prevalence of shallow depths near contaminated wells, because

> "if the depth of soil over the bedrock was shallow, there would be little opportunity for soil to interact with water and any contaminants percolating with it. Consequently, a relatively unrestricted flow of water would take place, allowing contaminants to enter the groundwater" (Conboy and Goss, 2000, p.4).

This also explains why the depth of a well is significant to its water quality. Deep wells offer greater protection than shallow wells because microbial (and other) contaminants are subject to increased filtration. Waters in shallow wells are less likely to have encountered geologic media that provide natural barriers to contaminant transport (O'Conner, 2002b). Apart from depth, numerous well characteristics affect water quality. The method of well construction is important because sandpoint, dug, and bored wells have been found to be subject to greater contamination than drilled wells (Goss et al., 1998). The age and casing material of a well are also relevant, because older wells are more likely to be under-maintained and have cracks in the casing, and casings made out of concrete generally crack more readily than plastics or metal. We are also

[30] The nature of surface/groundwater exchange depends on complex hydrogeological processes and is beyond the scope of this paper.

interested in the sites of abandoned or decommissioned wells, because improperly capped wells can provide an unmediated pathway for contaminants into an aquifer (Novokowski et al., 2006).

Fortunately, well-level data do exist. MOE maintains a Province-wide database called the Water Well Information System (WWIS), to which well custodians and contractors must log the construction or decommissioning of a well. The records date back to the 1940s and are available to the public at cost. The following data variables are available for each well: UTM coordinates, municipality, lot, elevation, casing material, casing diameter, pump rate, pump recovery rate, depth of water found, drill method, primary water use, and contractor. The WWIS is a wealth of information regarding Ontarian wells, yet we were unable to use these data at this time because the process of matching our geocoded records from the MOHLTC database to the (approximate) UTM coordinates in the MOE WWIS would be time-consuming and inexact[31]. Further, because there are no data collection standards, the accuracy of the WWIS data has been called into question in the past (Galal and Sarvas, 2005). Accurate well-level data will be obtained from the ARO case-control study questionnaire, and comparing these data with water tests will be a more reliable route for investigation.

Most of the local independent variables used in our analyses are geologic, and static in nature. Landuse is the only data that requires updating. However, we know that the presence of *E. coli* in well water can be sporadic due to precipitation events, temperature, and other fluctuating environmental conditions that impact *E. coli* survival and transport (Atherholt et al. 2003). For example, research shows that *E. coli* concentrations in groundwater spike in the 48-hour period after a heavy rainfall (Long and Plummer, 2004; Kelsey, Scott, Porter, et al., 2003). Likewise, *E. coli* thrive in warmer environments therefore are more likely to be present in a well later in the sampling season as temperatures rise (Chalmers et al., 2000). Charron et al. report that "there is mounting evidence that weather is often a factor at triggering waterborne disease outbreaks" (2004, p. 1667). The researchers warn that under conditions of global

[31] The matching process would be an approximation because there is no link between the two data to validate whether correct matches have been made.

warming these risks will increase, as "drought increases the demand for water when the supply is significantly reduced and vulnerable. Heavy rain following drought can lead to more severe runoff and risk of water contamination" (Charron et al., 2004, p.1674).

Although undoubtedly important, time and climate have been excluded from our analyses, first because existing climate data are sparse and incompatible for entry to the GIS at this time. Secondly, although a time stamp is attached to each well test, incorporating time into our analyses would mean re-conceptualizing our study design. Our analyses regard any well that tested positive for *E. coli* or TC during the course of an entire four month sampling season as a 'positive' case, regardless of the date the positive sample was taken. The samples would require disaggregation and re-coding in order to account for the temporal location of positive samples in the May 1 – Oct 1 season. Spending additional time recoding climate data and re-establishing our analyses to incorporate climate and time are beyond the scope of this project. Thirdly, well locations are static over time – and from a landuse planning perspective, it is important to identify fixed environmental characteristics that naturally protect or endanger well water quality throughout a range of weather conditions in the long term.

5.2.2 Spatial Analyses

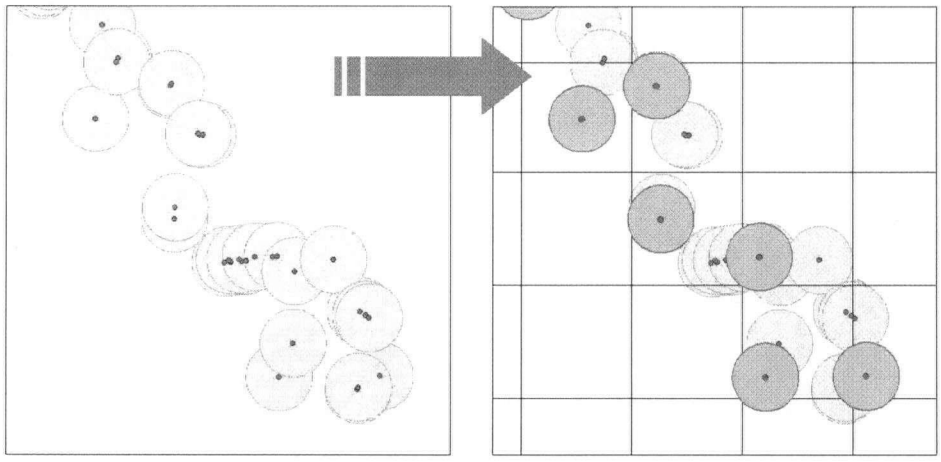
The section above describes data that we know are important to the presence of *E. coli,* but are absent from our analyses. However there may be other key environmental characteristics missing that we are not aware of, subsequently causing misspecification errors in our models. Additional geo-statistical analyses could help unpack this problem, and other limitations of this project. In this section, I will call attention to limitations in our analytical methodologies and suggest practical solutions; such as residual analyses, stratified sampling techniques, and multi-scale methodologies, that would benefit future investigations on this topic.

All observations in our models are referenced spatially, thus visualizing the geography of the residuals "may reveal insights into mechanistic relationships or other

kinds of associations between the predictors and the response" (White and Sifneos, 2002, p. 602). In plain terms, a residual is the estimated difference between a predicted and observed value. In logistic regression, the residual is calculated as: the observed value of Y minus the predicted probability, divided by the standard deviation of the predicted probability. The residuals could be visualized (by well location) in the GIS, and compared to other continuous environmental data to evaluate whether similar patterns may exist between layers. If a non-random pattern is identified in the residuals, then there is a chance that the exclusion of the mirroring data layer may be causing misspecification bias in our models.

Autocorrelation may also be causing error in the models. Although nearest-neighbor tests did not detect clustering, Moran's I coefficients indicate positive spatial autocorrelation in all samples, likely caused by grouping in human settlement patterns. The cause and effects of autocorrelation are complex issues, and providing in-depth analyses and solutions for potential spatial dependence in our samples is outside of the realm of this paper. However, I investigated the effects of autocorrelation on our results by applying a stratified random sub-sample to our data, a technique sometime used to mitigate effects of autocorrelation in simple random samples (McGrew and Monroe, 1993). A sub-sample of the 2004 *E. coli* group was selected by dividing the study area into sections based on a 1000m grid (100ha sections). I then joined the grid layer to the well points in the GIS, and randomly selected one well from each square to retain for analyses. See Fig. 5.1 for a diagram of this process.

Figure 5.1: Stratified Sub-sampling of Wells with a 1000m Grid



The stratified sub-sample reduced the original sample size by 24%, from n=5558 to n=4189. The same series of t-tests and logistic regressions as described in chapter 4 were applied to this sub-sample. The results collectively show no significant differences between the two sample groups, and the R^2 values returned by the regressions were all lower than those returned by the full 2004 *E. coli* sample. One possible reason why the sub-sampling did not improve the inferential power of the models is because the observations were still found to be positively autocorrected. The stratified sampling reduced clustering but did not eradicate it: in the sub-sample, the Moran's I coefficient is 0.36 compared to 0.78 in the full sample. For future analyses, I would suggest using a larger grid to stratify the sample, as 1000m does not separate groupings sufficiently. Further, samples could be stratified by boundaries other than a square grid, such as CCS boundaries, natural landforms, or geology.

Stratifying samples by environmental data may also help eliminate autocorrelation resulting from missing data values. Incomplete cases in a dataset can cause error if information loss is not random and occurs in one region more than in another. An example of data potentially subject to this problem is our soils layer. These data were collected by OMAF for the purposes of surveying regions for agricultural viability in the Province, and as a result data were not collected in areas where landcover is not suitable for cultivation (i.e. exposed bedrock or muskeg). The map of the OMAF soils layer (Fig. 3.6) shows evidence that large portions of data are missing in the northern parts of the study area. This northern region is located on the Canadian Shield, a geologic formation characterized by exposed bedrock, shallow soils, and low infiltration rates. The problem therein is that the incidence of missing data is not random, as areas with higher infiltration rates have more complete data than do areas with low rates. To address the potential error caused by this autocorrelation, sub-samples could be stratified by the two infiltration rate groups, with a balanced number of wells located in both high and low infiltration rate soils (proportionate to the total area of each variable). One benefit of having such a large sample size is that it allows for such flexibility in sub-sampling.

Well test results could also be aggregated to spatial boundaries delineated by environmental characteristics for an alternative scalar perspective. Likewise, land characteristics could be measured by zones other than circular buffers (i.e. watersheds). So far, our study has focused on the large scale (small area) effects on well water quality. While there is no doubt that local environmental factors play an important part in determining the water quality of a well, regional watershed characteristics have also been found to be important contributor to water quality (Bolstad and Swank, 1997, Sekhar and Raj, 1995; Kistemann et al. 2001; Sliva et al., 2001). Deciding on an 'appropriate' scale of analyses is an issue for any investigation of spatial data. Landscape ecologists are particularly involved with discussions on scale as they typically examine the structure, function, and patterns of a landscape – all of which are affected by different levels of observation. Levin (1992) argues that

> "there is no single natural scale at which ecological phenomena should be studied; systems generally show characteristic variability on a range of spatial, temporal, and organizational scales. The observer imposes a perceptual bias, a filter through which the system is viewed" (Levin, 1992, p. 1943).

Levin is referring to a phenomenon geographers term the Modifiable Areal Unit Problem (MAUP). When data are grouped to areal units of different sizes, analyses can obtain variable results. This is referred to as the scale effect (or aggregation effect). It would be interesting to learn whether the importance of environmental variables to well contamination changes when the landscape is analyzed at multiple scales. In their comparison of land characteristics selected by buffer zones and by watersheds, Johnson, Richards, Host, et al. (1997) found that data collected with their 100m buffer zones explained slightly more of the variability in water quality than did the characteristics of the entire watershed. Sliva et al. (2001) found *opposite* results, but comment that "the influence of buffer landscape composition in our study...may be underestimated due to the low resolution of digitized data used" (Sliva et al., 2001, p. 3471). Interestingly, these authors used the same geology and landuse layers in their study of three Ontarian watersheds as we did in this project.

In a regional analysis, insight into the effects of the MAUP on test outcomes could be offered by aggregating well water data to different regional boundaries (i.e. drainage basins, aquifer boundaries) and comparing results across data groupings.

Adding multi-scalar observations would add depth to this research by addressing both the local and regional environmental processes responsible for groundwater contamination. Multilevel models (also known as mixed models, or hierarchical models) include variables measured at more than one level of organization, and are useful for analyzing hierarchically-nested data. Applied to the ARO well water data, a multilevel model could potentially include data on local land characteristics (area of land characteristics inside the well buffer zone), and also include data on regional land characteristics that a well is located in (area of land characteristics inside a watershed, for example).

Classification and Regression Tree (CART) modeling is a multilevel approach that shows potential for use in this project. CARTs are inferential models based on decision trees where binary (yes/no) questions are posed to classify data, and where the 'leaves' of the tree predict membership of cases based on one or more independent variables. Applied to our research question, the goal of a CART would be to predict the presence of *E. coli* in a well based on a hierarchical system for sorting cases, accomplished by asking binary questions such as: 'does this well lie on a moraine?' Or, 'does this well lie on or near agricultural land?" Algorithms can be applied to determine the optimal order and combination of binary decisions, providing further insight into the hierarchical effects of the independent variables on well water results.

In considering analytical methods to apply to our data, the possibilities are endless, and as outlined above there are a number of opportunities for continued work on this project. The large sample sizes, and extensive environmental data contained in the GIS to date provide a solid basis for further investigation of the research questions raised here.

5.3 Conclusion

In summary, our analyses of the relationships between private well water quality and local landscape characteristics have proved productive on many levels. First, a GIS database containing the most current available data on environmental characteristics relevant to microbial water quality in Southern Ontario was compiled and used as a

foundation for analyses. This database is ready for the use of other researchers working on the ARO project. Secondly, this work demonstrates the value of applying concepts in spatial epidemiology to groundwater studies. In this project, geo-statistical methods for measuring land characteristics surrounding private wells were explored, and circular buffer zones with 300m radii were found to be the most relevant for our analyses.

Third, the results of numerous statistical tests identify a list of specific local land characteristics that are more prevalent near contaminated wells than to clean wells. We found that areas of agricultural land, low population density, low-infiltration rate soils, surficial geology with low permeability, untilled land, and carbonate bedrock are more prevalent near wells tested positive for *E. coli* or TC at least once in the May 1- Oct 1 sampling season, for both 2003 and 2004. These results indicate that private wells located in or near any of the aforementioned characteristics may be at an increased risk of contamination by fecal coliforms. The following characteristics were found to be more prevalent near clean wells: developed land, high population density, high-infiltration rate soils, surficial geology with high permeability, tilled soils, and non-carbonate bedrock. These characteristics may offer natural protection to groundwater sources located nearby.

Fourth, multivariate regression modeling was explored in tandem with various other statistical tests in order to derive models to predict the presence of *E. coli* and TC in private wells based on land use and population attributes. Although the models identified individual environmental covariates (listed above) associated with the presence or absence of fecal coliforms in well water, the models did not have any significant inferential power. The weak R^2 values indicate that misspecification errors exist in the models, possibly caused by absent independent variables and/or autocorrelation of observations. The problems associated with the inferential modeling have highlighted some key steps to take to improve and build upon these analyses. Through these combined explorations, we have furthered our understanding of the potential processes responsible for the effective transfer of microbes from the environment and animals to humans, specific to our study region. This research is exploratory by nature, and the limitations identified simply serve to strengthen future work in this field and in the larger ARO project.

The applications of this research are many, within and beyond the ARO project. On the basis of this study, questions regarding the identified risk factors have been added to a questionnaire used by the ARO case-control group to obtain qualitative information on the condition of informants' wells and the surrounding environment. An exciting opportunity to 'ground-truth' our findings exists here. The quantitative area measurements in our tables could be compared to the qualitative information provided by well custodians on a well-by-well basis to assess how representative our digital data is of the real world. The informants' perceptions of the natural environment could also be incorrect, so for increased accuracy the interviewing team could conduct a survey of the well locality. The qualitative information provided by informants and interviewers on land characteristics adjacent to wells could be recoded and entered into a logistic regression model for group analyses similar to those described in sections 4.2.2 and 4.2.3.

The spatial database stands as a basis for other researchers, and components of these data will be used by another graduate student working with Dr. Buzzelli in group 4 of the ARO team. This student will be involved with 3-dimensional computer modeling of groundwater flow, adding further complexity to the research and addressing many of the data and analytical limitations outlined in section 5.2 of this thesis, such as lack of addition of important time, climate, and geological data. In this respect, the identification of the shortcomings of this project has helped pave the direction for more advanced work on the environmental covariates of *E. coli* in Southern Ontario.

On a broader level, this thesis adds to understanding the effects of non-point source pollution on private groundwater quality in Ontario. In the Inquiry Report following the Walkerton crisis, Justice Dennis O'Connor comments that research in this field has been limited in the past:

> "Non-point source contamination has received significantly less attention and less funding, in part because some non-point contaminants are considered less dangerous, and because non-point application of fertilizer, manure and pesticides is a necessary part of agriculture" (O'Connor, 2002b, p. 2).

The fallout from the Walkerton crisis has had much to do with changing the tide of research in a different direction, this thesis being an indirect product of this shift in

thought. The public health problem associated with private water sources can be further reduced by the continued identification and understanding of risk factors such as those identified here, and by the proper protection of water sources through this knowledge (Said, Wright, Nichols, et al., 2003).

The methods explored in this project also contribute to the evolving field of landscape epidemiology. Studies that apply GIS and statistical analyses to investigate the effects of local land characteristics on private well water sources are scarce, especially in our study region. However, as a whole, we will likely see an increase in local area health studies. Digital data are becoming available at a more detailed level; therefore research in spatial epidemiology is beginning to focus on large-scale variability within populations as opposed to more traditional regional approaches. The limitations of this project also continue to constrain other work in this field and, collectively, spatial epidemiologists are calling for the rejection of static space/time/health research models to improve disease process analyses (Rushton, 2000; Meade and Earickson, 2000; Jacquez et al., 2000; Jarup, 2004). Further work is needed to refine models to represent the unpredictable bio-physical world in order to understand disease ecology – and this will require a transcendence of traditional fixed space/time research models that has only just begun to take place.

Trends in health geography are also shifting towards the incorporation of more complex study designs. Kearns and Moon (2002) suggest that longitudinal or repeated cross-sectional studies should be performed to better understand processes of change in health and place. Today, in the face of global warming, the importance of longitudinal approaches to studying climate events, regional ecology, and human health is emphasized. Existing drinking water infrastructure is designed to operate within expected climate and landuse dynamics, and as Charron et al. warn,

> "the frequency and severity of drought, flood, sea-level rise, extreme rainfall, and changes in snow cover and timing of snowmelt may change in some parts of Canada under conditions of climate change. As a result, pathogen entry and behavior in source and finished water will also be subject to change" (Charron et al, 2004, p. 1668).

Clearly, it is imperative to understand the drawbacks of our existing systems in order to prepare for how potential landscape changes may impact drinking water quality.

In summary, this project is a study in environmental health, a term the WHO defines as: "the theory and practice of assessing, correcting, controlling, and preventing those factors in the environment that can potentially affect adversely the health of present and future generations" (WHO, 2006, ¶1). By identifying environmental characteristics that facilitate the transfer of microbes from animals to humans through the contamination of source waters, the research described in this thesis contributes to one facet of a multi-barrier approach to public health protection. The continuation of this work is important here in Canada, but is especially vital in other nations, as there is a current bias in the spatial clustering of similar research examples in developed countries. The worldwide public health problem associated with the contamination of groundwater sources by *E. coli* could be reduced by landscape-specific identification and understanding of risk. As landuse and agricultural production intensifies, as populations grow, and as weather events become more severe due to global warming; the risk of human exposure to fecal coliforms will increase worldwide, certainly highlighting the importance of continuing such research.

REFERENCES

Ali, M., Emch, M., Donnay, J. P., Yunus, M., & Sack, R. B. (2002). Identifying environmental risk factors for endemic cholera, a raster GIS approach. *Health and Place*, 8, 201-210.

Allen, M. J., & Morrison, S. M. (1973). Bacterial movement through fractured bedrock. *Ground Water*, 11(2), 6-10.

An, Y. J., & Breindenbach, G. P. (2005). Monitoring *E. coli* and total coliforms in natural spring water as related to natural recreation mountain areas. *Environmental Monitoring and Assessment*, 102(1-3), 131-137.

Andrews, G. J. (2002). Towards a more place-sensitive nursing research, an invitation to medical and health geography. *Nursing Inquiry*, 9(4), 221-238.

Atherholt, T., Feerst, E., Hovendon, B., Kwak, J., & Rosen, J. D. (2003). Evaluation of indicators of fecal contamination in groundwater. *American Water Works Journal*. 95(10), 119-131.

Barringer, T., Dunn, D., Battaglin, W., & Vowinkel, E. (1990). Problems and methods involved in relating land use to groundwater quality. *Water Resources Bulletin*. 26, 1-9.

British Columbia Water and Waste Association (BCWWA). (2006). Source To Tap Assessment Guide, Module 1 – Delineate and Characterize Drinking Water Sources. Retrieved February 9, 2006, from BCWWA Web site: http://www.bcwwa.org/source-to-tap/documents/mod-1-delineate-&-characterize-source.pdf

Berg, H. C. (2003). E. *coli in Motion*. New York: Springer.

Berry W. D., & Feldman, S. (1985). *Multiple Regression in Practice*. London: Sage Publications.

Bingham, P. (2004). John Snow, William Farr and the 1849 outbreak of cholera that affected London, a reworking of the data highlights the importance of the water supply. *Public Health*, 118(6), 387-394.

Bocking, S. (2002, November). *Linking Science and Policy for Urban Nonpoint Source Pollution in the Great Lakes Region*. Retrieved June 17, 2006, from International Association for Great Lakes Research Web site: http://www.iaglr.org/scipolicy/nps/nps_iaglr02.pdf

Bolstad, P. V., & Swank, W. T. (1997). Cumulative impacts of landuse on water quality in a southern Appalachian watershed. *Journal of the American Water resources Association*, 33(3), 519-533.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., & Freudenheim, J.L. (2003). Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology*, 14(4), 408-412.

Brunke, M., & Gonser, T. (1997). The ecological significance of exchange processes between rivers and groundwater. *Freshwater Biology*, 37(1), 1-33.

Chalmers, R. M., Aird, H., & Bolton, F.J. (2000). Waterborne Escherichia coli O157. *Journal of Applied Microbiology*, Suppl. S, 88, 124S-132S.

Charron, D. F., Thomas, M. K., Waltner-Toews, D., Aramini, J. J., Edge, T., Kent, R. A., Maarouf, A. R., & Wilson, J. (2004). Vulnerability of waterborne diseases to climate change in Canada: A review. *Journal of Toxicology and Environmental Health, Part A*, 67, 1667-1677.

Clark, K. C., McLafferty, S. L., & Tempalski, B. J. (1996). On epidemiology and Geographic Information Systems: A Review and Discussion of Future Directions. *Emerging Infectious Diseases*, 2(2), 85-92.

Conboy, M. J., & Goss, M. J. (2001). Identification of an assemblage of indicator organisms to assess timing and source of bacterial contamination in groundwater. *Water, Air, and Soil Pollution,* 129, 101-118.

Conboy, M. J., & Goss, M. J. (2000). Natural protection of groundwater against bacteria of fecal origin. *Journal of Contaminant Hydrology*, 43, 1-24.

Crane, S. R., Westerman, P. W., & Overcash, M. R. (1980). Die-off of fecal indicator organisms following land application of poultry manure. *Journal of Environmental Quality*, 9(3), 531-537.

Crane, S. R., Moore, J. A., Grismer, M. E., & Minter, J. R. (1983). Bacterial pollution from agricultural sources: a review. *Transactions of the American Society of Agricultural Engineers*, 26(3), 858-866.

Crowther, J., Kay, D., & Wyer, M. D. (2002). Faecal-indicator concentrations in water draining lowland pastoral catchement in the UK: relationships with land use and farming practices. *Water Research*, 36, 1725-1734.

Dangendorf , F., Herbst, S., Reintjes, R., & Kistemann, T. (2002). Spatial patterns of diarrhoeal illnesses with regard to water supply structures – a GIS analysis. *International Journal of Environmental Health*, 205, 183-191..

De Loe, R. C., & Kreutzwiser, R. D. (2005). Closing the groundwater protection gap. *Geoforum,* 36, 241-256.

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

Edge, T., Byrne, J. M., Johnson, R., Robertson, W., & Stevenson, R. (2003). *Waterborne Pathogens: A report prepared for the National Water Research Institute, Environment Canada.* Retrieved on May 4, 2006, from the National Water Research Institute Web site: http://www.nwri.ca/threatsfull/ch1-1-e.html

Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, 112(9), 998-1006.

Francy, D. S., Helsel, D. R., & Nally, R. A. (2000). Occurrence and distribution of microbiological indicators in groundwater and stream water. *Water Environment Research*, 72(2), 152-161.

Gagliardi, J. V., & Karns, J. S. (2000). Leaching of Escherichia coli O157:H7 in diverse soils under various agricultural management practices. *Applied and Environmental Microbiology*, 66(3), 877-883.

Galel, D., & Sarvas, P. (2005). *Summary of Field Work and Other Activities 2005.* Ontario Geological Survey, Open File Report 6172, p.32-1 to 32-4. Retrieved on June 18, 2006, from the Ontario Ministry of Mines and Northern Development, Groundwater Resource Project for the Halton Area Watershed Web site: http://www.mndm.gov.on.ca/mndm/mines/ims/pub/sfw/sfwpdf/6172-32.pdf

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

George, I., Anzil, A., & Servais, P. (2004). Quantification of fecal coliform inputs to aquatic systems through soil leaching. *Water Research*, 38, 611-618.

Girard, M. P., Steele, D., Chaignat, C. L., & Kieny, M. P. (2006). A review of vaccine research and development, human enteric infections. *Vaccine*, 24(15), 2732-2750.

Goss, M. J., Barry, D. A. J., & Rudolph, D. L. (1998). Contamination in Ontario farmstead domestic wells and its association with agriculture, 1. Results from drinking water wells. *Journal of Contaminant Hydrology*, 32, 267-293.

Government of Ontario (2006). *About Ontario: Climate.* Accessed on September 30, 2006, from http://www.gov.on.ca/ont/portal/!ut/p/.cmd/cs/.ce/7_0_A/.s/7_0_252/_s.7_0_A/7_0_252/_l/en?docid=004190

Graham, A. J., Atkinson, P. M., & Danson, F. M. (2004). Spatial analysis for epidemiology. *Acta Tropica*, 91(3), 219-225.

Griffin, P. M., & Tauxe, R. V. (1991). The epidemiology of infections caused by Escherichia coli O157:H7, other enterohemorrhagic E. coli, and the associated hemolytic uremic syndrome. *Epidemiology Reviews*, 13, 60-98.

Haack, J. P., Jelacic, S., Besser, T. E., Weinberger, E., Kirk, D. J., McKee, G. L., Harrison, S. M., Musgrave, K. J., Miller, G., Price, T. H., & Tarr, P. I. (2003). Escherichia coli O157:H7 Exposure in Wyoming and Seattle, Serologic Evidence of Rural Risk. *Emerging Infectious Diseases*, 9(10), 1226-1231.

Hamilton, L. (1992). Regression *with Graphics: A Second Course in Applied Statistics.* Belmont, California: Duxbury Press.

Heath Canada. (2004). *Summary of Guidelines for Canadian Drinking Water Quality.* Prepared by the Federal-Provincial-Territorial Committee on Drinking Water, Accessed April 1, 2005, from http://www.hc-sc.gc.ca/hecs-sesc/water/pdf/summary.pdf

Herpel, R. (2006). *Source Water Assessment and Protection: Workshop Guide, 2nd Ed.* Retrieved on March 8, 2006, from the Groundwater Foundation Web site: http://www.groundwater.org/gi/swap/SWAP2_FullGuide3.pdf

Holme, R. (2003). Drinking water contamination in Walkerton, Ontario: positive resolutions from a tragic event. *Water Science and Technology*, 47(3), 1-6.

Horsley, S., & Witten, J. (1995). *A Guide to Wellhead Protection.* Chicago: APA Publications.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* New York: Wiley.

Hrudey, S. E., Payment, P., Huck, P. M., Gillham, R. W., & Hrudey, E. J. (2003). A fatal waterborne disease epidemic in Walkerton, Ontario: comparison with other waterborne outbreaks in the developed world. *Water Science and Technology*, 47(3), 7-14.

Hunter, C., & McDonald, A. (1991). The occurrence of coliform bacteria in the surface soils of two catchment areas in the Yorkshire Dales. *Journal of the Institute of Water and Environmental Management*, 5, 534-538.

Innocent, G. T., Mellor, D. J., McEwen, S. A, Reilly, W. J., Smallwood, J, Locking, M. E., Shaw, D. J., Michel, P., Taylor, D. J., et al. (2005). Spatial and temporal epidemiology of sporadic human cases of Escherichia coli O157 in Scotland, 1996-1999. *Epidemiology of Infection*, 133(6), 1033-41.

Jacquez, G. M. (2000). Spatial analysis in epidemiology: Nascent science or a failure of GIS? *Journal of Geographic Information Systems,* 2, 91-97.

Jarup, L. (2004). Health and environment information systems for exposure and disease mapping, and risk assessment. *Environmental Health Perspectives*, 112, 995-7.

Jones, K., & Moon, G. (1987). *Health, disease and society: An introduction to medical geography.* London, England: RKP.

Johnson, L. B., Richards, C., Host, G. E., & Arthur, J. W. (1997). Landscape influences on water chemistry on Midwestern stream ecosystems. *Freshwater Biology*, 37, 193–208.

Kearns, R. A. & Gesler, W. M. (1998). *Putting health into place, landscape, identity, and well being*. New York: Syracuse University Press.

Kearns, R. A., & Moon, G. (2002). From medical to health geography: novelty, place and theory after a decade of change. *Progress in Human Geography*, 26(5), 605-625.

Kelsey, R. H., Scott, G. I., Porter, D. E., Thompson, B., & Webster, L. (2003). Using multiple antibiotic resistance and land use characteristics to determine sources of fecal coliform bacterial pollution. *Environmental Monitoring and Assessment*, 81, 337-348.

Kistemann, T., Dangendorf, F., & Exner, M. (2001). A Geographical Information System (GIS) as a tool for microbial risk assessment in catchment areas of drinking water reservoirs. *International Journal of Hygiene and Environmental Health,* 203, 225-233.

Kolpin, D.W. (1997). Agricultural chemicals in groundwater of the Midwestern United States, relations to land use. *Journal of Environmental Quality*, 26, 1025-1037.

Lee, S. M., Min, K. D., Woo, N. C., Kim, Y. J., & Ahn, C. H. (2003). Statistical models for the assessment of nitrate contamination in urban groundwater using GIS. *Environmental Geology*, 44, 210-221.

Levin, S.A. (1992). The problem of scale and pattern in ecology. *Ecology*, 73(6), 1943-1967.

Lim, C. H., & Flint, K. P. (1989). The effects of nutrients on the survival of Escherichia coli in lake water. *The Journal of Applied Bacteriology,* 66(6), 559-569.

Lin, S. (2001). *Water and Wastewater Calculations Manual.* New York: McGraw-Hill.

Long, S. C., & Plummer, J. D. (2004). Assessing land use impacts on water quality using microbial source tracking. *Journal of the American Water Resources Association,* 40(6), 1433-1448.

Macler, B. A, & Merkle, J. C. (2000). Current knowledge on groundwater microbial pathogens and their control. *Hydrogeology Journal*, 8(1), 29-40.

Mallin, M. A., Williams, K. E., Esham, E. C., & Lowe, R. P. (2000). Effect of human development on bacteriological water quality in coastal watersheds. *Ecological Applications,* 10(4), 1047-1056.

McGrew, J. C., & Monroe, C. B. (1993). *An Introduction to Statistical Problem Solving in Geography*. Boston: McGraw-Hill.

McLay, C. D., & Dragten, R. (2001). Predicting groundwater nitrate concentrations in a region of mixed agricultural land use, a comparison of three approaches. *Environmental Pollution*, 115, 191-204.

McQuarrie, A. D., & Tsai, C. (1998). *Regression and Time Series Model Selection.* New Jersey: World Scientific Publishing Co.

Meade, M. S. & Earickson, R. (2001). *Medical Geography.* (2nd Ed.). New York: The Guilford Press.

Meade, M. S., Florin, J. W., and & Gesler, W. M. (1988). Landscape Epidemiology. In M. S. Meade & R. Earickson (Eds.) *Medical Geography* (pp. 59-106). New York, The Guilford Press.

Miller G. (2000). *The Protection of Ontario's Groundwater and Intensive Farming: Special Report to the Legislative Assembly of Ontario.* Retrieved July 28, 2006, from the Environmental Commissioner of Ontario Web site: www.eco.on.ca/english/publicat/sp03.pdf

Miller, J. J., Beasley, B. W., Yanke, L. J., Larney, F. J., McAllister, T. A., Olson, B. M., Selinger, L. B., Chanasyk, D. S., & Hasselback, P. (2003). Bedding and Seasonal Effects on Chemical and Bacterial Properties of Feedlot Cattle Manure. *Journal of Environmental Quality,* 32, 1887-1894.

Ministry of Environment (MOE). (2001). *Delineation of wellhead protection areas for municipal groundwater supply wells under the direct influence of surface water.* Report PIBS 4168e. Retrieved November 11, 2005, from the MOE Web site: http://www.ene.gov.on.ca/envision/techdocs/4168e.htm

Ministry of Environment (MOE). (2003). *Protocol of Accepted Drinking-Water Testing Methods: A Report prepared by the Ministry of Environment Laboratory Services Branch.* Retrieved on October 1, 2006, from the MOE Web site: http://www.ene.gov.on.ca/envision/gp/4465e.pdf

Nygard, K., Andersson, Y., Rottengen, H. A., Svensson, A., Lindback, J., Kistemann, T., & Giesecke, J. (2004). Association between environmental risk factors and campylobacter infections in Sweden. *Epidemiology and Infection,* 132, 317-325.

Na, S. H., Miyanaga, K., Unno, H., & Tanji, Y. (2006). The survival response of Escherichia coli K12 in a natural environment. *Applied Microbiology and Biotechnology*, 72(2), 386-392.

New Jersey Department of Environmental Protection (NJDEP). (2003). *Guidelines for Delineation of Well Head Protection Areas in New Jersey*: New Jersey Geological Survey Open-File Report OFR 03-1. Retrieved on January 18, 2006, from the NJDEP Web site: http://www.state.nj.us/dep/njgs/whpaguide.pdf

Novokowski, K., Beatty, B., Conboy, M. J., & Lebedin, J. (2006). *Water Well Sustainability in Ontario: Expert Panel Report*. Prepared for the Ontario Ministry of the Environment Sustainable Water Well Initiative. Retrieved on April 4, 2006, from the Ontario Centres of Excellence Web site: www.oce-ontario.org/pages/PDF/SWWI_Final_Jan30.pdf

O'Connor, D. R. (2002a). *Report of the Walkerton Inquiry, Part One: The Events of May 2000 and Related Issues*. Toronto, Ontario: Queen's Printer for Ontario.

O'Connor, D. R. (2002b). *Report of the Walkerton Inquiry, Part Two: A Strategy for Safe Drinking Water*. Toronto, Ontario: Queen's Printer for Ontario.

Olson, B. M., Miller, J. J., Rodvang, S. J., & Yanke, L. J. (2005). Soil and groundwater quality under a cattle feedlot in southern Alberta. *Water Quality Research Journal of Canada*, 40(2), 131-144.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, 8(6). Retrieved August 2, 2006 from http://PAREonline.net/getvn.asp?v=8&n=6

Ostfeld, R. S., Glass, G. E., & Keesing, F. (2005). Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in ecology and evolution*, 20(6), 328-336.

Pavlovsky, E. N. (1966). *Natural Nidality of Transmissible Diseases, With Special Reference to the Landscape Epidemiology of Zooanthroponoses*. Chicago, University of Illinois Press.

Postgate, J. (2000). *Microbes and Man* (4th Ed.). Cambridge, UK: Cambridge University Press.

Rahe, T. M., Hagedorn, C., McCoy, E. L., & Kling, G. F. (1978). Transport of antibiotic-resistant Escherichia coli through western Oregon hillslope soils under conditions of saturated flow. *Journal of Environmental Quality*, 7(4), 487-494.

Raina, P. S., Pollari, F. L., Teare, G. F., Goss, M. J., Barry, D. A. J., & Wilson, J. B. (1999). The relationship between E. coli indicator bacteria in well-water and

gastrointestinal illnesses in rural families. *Canadian Journal of Public Health*, 90(3), 172-175.

Raper, J. (2004). Book Review: Geographical and environmental epidemiology, Methods for small-area studies. *International Journal of Geographical Information Science*, 18 (6), 629-629.

Regional Municipality of Waterloo. (2003). *Protecting Significant Moraines in Waterloo Region: A Supplementary Report in Support of Waterloo Region's Growth Management Strategy.* Retrieved on June 22, 2006, from http://www.region.waterloo.on.ca/web/region.nsf/0/5E9A98FD583B353F852570000056FA92/$file/protectingsuprep.pdf?OpenElement

Rosen, B. H. (2000). *Waterborne Pathogens in Agricultural Watersheds, Watershed Science Institute, Conservation Technical Note 2.* Retrieved on Dec 12, 2005, from: http://nature.berkeley.edu/forestry/rangelandwq/pdfs/Atwillwssitn21.pdf

Rosenberg, M.W. (1998). Medical or Health Geography? Populations, people, places. *International Journal of Population Geography*, 4, 211-226.

Rushton, G. (2000). GIS to improve public health: Guest Editorial. *Transactions in GIS*, 4, 1-4.

Said, B., Wright, F., Nichols, G. L., Reacher, M., & Rutter, M. (2003). Outbreaks of infectious disease associated with private water supplied in England and Wales 1970-2000. *Epidemiology and Infection*, 130, 469-479.

Schaffter, N., Zumstein, J., & Parriaux, A. (2004). Factors influencing the bacteriological water quality in mountainous surface and groundwaters. *Acta Hydrochimica et Hydrobiologica*, 32(3), 225-234.

Sekhar, M. C., & Raj, P. A. (1995). Landuse – water quality modeling: A case study. *Water Science and Technology*, 31(8), 383-386.

Sliva, L., & Williams, D. D. (2001). Buffer zone versus whole catchement approaches to studying land use impact on river water quality. *Water Research*, 35(14), 2463-3472.

Statistics Canada. (1996). A *geographical profile of manure production in Canada: 1996.* Retrieved on August 18, 2006, from: http://www.statcan.ca/english/freepub/16F0025XIB/m/manure.htm

Statistics Canada. (2001). A *geographical profile of manure production in Canada: 2001.* Retrieved on August 18, 2006, from: http://dsp-psd.pwgsc.gc.ca/Collection/Statcan/21-601-MIE/21-601-MIE2006077.pdf

Truett, J., Cornfield, J., & Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Disease*, 20, 511-524.

University of California, Los Angeles (UCLA). (2006). *Logistic Regression Diagnostics*. Retrieved on August 12, 2006, from UCLA's online Web book on Logistic Regression: http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter3/statalog3.htm

United Nation Environment Programme (UNEP). (2003). *Groundwater and its susceptibility to degradation, a global assessment of the problem and its options for management: Early Warning and Assessment Report Series*. Retrieved on Sept 25, 2005, from the United Nations Environment Programme Groundwater Protection Web site: http://www.unep.org/DEWA/water/groundwater/pdfs/Groundwater_Prelims_SCREEN.pdf

United States Environmental Protection Agency (US EPA). (2006). *Drinking Water Pathogens and Their Indicators: A Reference Resource*. Retrieved on Sept 29, 2005, from the EPA Web site: http://www.epa.gov/enviro/html/icr/gloss_path.html

United States Environmental Protection Agency (US EPA). (1991). *Protecting Local Ground-Water Supplies through Wellhead Protection*. Report 570/9-91-007. Retrieved April 24, 2005, from the EPA Web site: http://www.epa.gov/r10earth/offices/water/whpgprnt.pdf

Vine, M. F., & Degnan, D. (1998). Geographic Information Systems: their use in environmental epidemiology. *Journal of Environmental Health*, 61(3), 7-17.

Wagner, M. J. (2005). *Agribusiness Prepares to Harvest RADARSAT-2 Data*. Retrieved on September 30, 2006, from The Earth Imaging Journal Web site: http://www.eijournal.com/radarsatagri.asp

Wang, M. X., Liu, G. D., Wu, W. L., Bao, Y. H., & Liu, W. N. (2006). Prediction of agriculture derived nitrate distribution in North China Plain with GIS-based BPNN. *Environmental Geology*, 50, 637-644.

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., Mix, W. Colt, J., & Hartge, P. (2005). Positional Accuracy of Two Methods of Geocoding. *Epidemiology*, 16(4), 542-547.

Washington State Department of Health. (2006). Fact Sheet: Coliform Bacteria and Drinking Water. Retrieved on June 19, 2006, from the Washington State Department Web site, Division of Environmental Health, Office of Drinking Water: http://www.doh.wa.gov/ehp/dw/programs/coliform.htm

White, D., & Sifneos, J.C. (2002). Regression Tree Cartography. *Journal of Computational and Graphical Statistics*, 11(3), 600-614.

World Health Organization (WHO). (2006). Public *Health and Environment*. Retrieved on September 25, 2005, from the World Health Organization Environmental Health Web site: http://www.who.int/phe/en/

World Health Organization (WHO). (1993). *Guidelines for Drinking-water Quality: 2nd Edition*, Volume 1: Recommendations. Retrieved on June 18, 2006, from the World Health Organization Web site: http://www.who.int/water_sanitation_health/GDWQ/Chemicals

York Region. (2006). Frequently Asked Questions About Private Well Water. Retrieved on February 14, 2006, from the York Region Web site, Public Health and Safety: http://www.region.york.on.ca/Services/Public+Health+and+Safety/Safe+Water+Program/Frequently+Asked+Questions+About+Private+Well+Water.htm

APPENDICES

APPENDIX 1: Ontario Farm and Population Data, 1996-2001
Tables adapted from Statistics Canada (1996, 2001).

Ontario Farm Data, 1996 and 2001 Census of Agriculture

Item	1996	2001	% change, 1996 to 2001
Number of census farms	67,520	59,728	-13.0
Total acreage of farms	13,879,565	13,507,357	-2.8
Average farm size acres	206	226	8.8
Cropland acres	8,759,707	9,035,915	3.1
Pasture acres	2,502,478	2,087,985	-19.9
Summerfallow acres	48,492	35,175	-37.9
All other land acres	2,568,888	2,348,282	-9.4
Total of above crops acres	8,739,902	9,004,267	2.9
Greenhouse area thousand ft^2	63,303	98,374	35.7
Total number of cattle and calves	2,285,996	2,140,731	-6.8
Total number of pigs	2,831,082	3,457,346	18.1
Number of sheep and lambs	231,087	337,625	31.6
Number of horses	76,553	83,337	8.1
Number of hens and chickens	35,596,946	43,624,696	18.4
Number of tractors	180,213	183,704	1.9
Number of combines	19,855	17,677	-12.3
Number of balers	38,329	35,385	-8.3
Market dollar value of land and buildings	33,167,842,178	40,898,278,324	18.9
Dollar value of machinery and equipment	5,410,519,342	6,564,007,507	17.6
Dollar value of livestock and poultry	2,282,574,515	3,067,497,674	25.6
Total capital value	40,860,936,035	50,529,783,505	19.1

Ontario Population Data, 1996 and 2001 Census of Population

Item	1996	2001	% change, 1996 to 2001
Ontario	10,753,373	11,410,046	5.8
Rural Southern Ontario	3,792,719	3,998,060	5.1
Urban Southern Ontario	6,128,529	6,616,361	7.4
Total Southen Ontario	9,927,297	10,623,603	6.6

References:

Statistics Canada

APPENDIX 2: Geocoding Results[32]

2003

Of the 181 558 entries in the initial database,

43,903 (~24%) entries were geocoded using the civic address
11,392 (~ 6%) entries were geocoded according to lot/concession numbers
17,373(~10%) entries were geocoded using the six digit postal code
34,883 (~19%) entries were geocoded using the five digit postal code
30,340 (~17%) entries were geocoded using the tree digit postal code
34,681 (~ 19%) entries were geocoded to the centroid of their municipality

In total: 172,572 entries (~95%) were geocoded while 8,986 (~5%) were not.

2004

Of the 280,139 entries in the initial database,

70,073 (~25%) entries were geocoded using the civic address
14,182 (~ 5%) entries were geocoded according to lot/concession numbers
27,937 (~10%) entries were geocoded using the six digit postal code
55,927 (~20%) entries were geocoded using the five digit postal code
51,091 (~18%) entries were geocoded using the tree digit postal code
47,964 (~ 17%) entries were geocoded to the centroid of their municipality

In total: 267,174 entries (~95%) were geocoded while 12,965 (~5%) were not.

[32] Adapted from Caroline Guenette's Geocoding Reports entitled: Geocoding a database related to water testing: ARO Study (2003 v.2.1, 2004 v.1).

APPENDIX 3: Water Sampling Guidelines[33]

How do I take a water sample properly?

Use the sampling bottles provided by the laboratory. These will be sterile, 250-ml bottles containing the preservative sodium thiosulphate to prevent the collected water from degrading. Do not touch or handle the preservative. Do not rinse the bottle as this will remove some or all of the preservative and ruin the sample.

Choose an inside tap that is not connected to a treatment device such as a chlorinator or ultra-violet light treatment system. Then follow the sampling procedure below:

1. Wash hands carefully.
2. Remove screens, aerators or other attachments from the faucet
3. Run cold water for 3-5 minutes to ensure a constant temperature and to clear stagnant water that may have been sitting in the lines/pipes.
4. Reduce the water flow to a steady stream.
5. Take the cap off the bottle and hold it in one hand and the bottle in the other. Do not rinse the bottle.
6. Do not set the cap down or drop it. Do not touch the inside of the cap or the mouth of the bottle. Bacteria on your hands will contaminate the sample.
7. Carefully fill the bottle to line indicated, approximately one inch (2.54 cm) from the top.
8. Put the cap back on the bottle so that the inside of the cap and the mouth of the bottle are untouched.
9. Fill in the Bacteriological Analysis of Drinking Water for Private Citizen, SINGLE HOUSEHOLD ONLY form and bring the water sample to the laboratory as soon as possible.
10. Refrigerate the samples and maintain at a temperature of about 4 ° C (40 ° F). This will slow bacterial growth and maintain the target bacterial population at the level that existed at the time of sample collection. Keep the samples chilled during transport to the lab. Do not freeze the samples.

Please Note: The result is unreliable if the sample is improperly collected or improperly stored. If the water sample takes more than 48 hours to reach the laboratory, it will not be tested.

It is recommended that well water be tested:

1. After well construction is completed and the well has been disinfected,
2. When a well has not been in use for long periods, e.g., seasonal residences, and
3. 2-3 times during the year, preferably after a heavy rain or snow melt.

It is also recommended that 3 samples are taken 1-3 weeks apart to determine well water quality. Please note that the bacterial stability of water cannot always be determined from a single sample.

[33] Adapted from the York Region Website (2006).

APPENDIX 4: List of Data and Sources

Data Description	Originator	Year
Well Data		
Well water test results – ARO Pilot data	ON MOHLTC	2003
Well water test results – ARO Surveillance data	ON MOHLTC	2004
Water well borehole logs	MOE	1946-present
Administrative Boundaries		
ON Public Health Unit	Statistics Canada	2003
ON District Health Council Areas	Statistics Canada	2003
ON Census Agricultural Regions (CAR)	Statistics Canada	2001
ON Agricultural Census Consolidated Subdivisions (CCS)	Statistics Canada	2001
ON Dissemination Areas (DA)	Statistics Canada	2001
ON Census Tracts (CT)	Statistics Canada	2001
ON Regional Districts	Statistics Canada	2001
Provinces	Statistics Canada	2001
Postal Geographies		
ON Forward Sortation Areas (FSA)	Statistics Canada	2003
ON 6-digit postal code points	Statistics Canada	2003
Land Cover		
Land Cover	ON MNR	1986-1997
Land Use	CLI	1966
Soils		
Soils – ON (type, drainage, infiltration rates)	CANSIS	1970-1990
Soils – Southern ON, Municipal Soil Surveys	OMAF	1950-present
Geology		
Bedrock Geology – ON	ON Geological Survey	1970 – present
Quaternary Geology – ON	ON Geological Survey	1970 – present
Surficial Geology – Southern ON	ON Geological Survey	1970 – present
Moraine Deposits	ON Geological Survey	1970 - present
Surficial Geology - Canada (1880A)	Geological Survey of Canada	1960 - present
Geological Map of Canada (1860A)	Geological Survey of Canada	1960 - present
Topography		
DEM (30m resolution) - ON	DMTI	---
Hydrology		
Canadian Water	Statistics Canada	2001
ON Hydrology (Lakes & Rivers)	Statistics Canada	2001
Meteorology		
Meteorology	Environment Canada	2000-2003
Meteorology	Environment Canada	1840-present

APPENDIX 5: Sample of a hypothetical table ready for analysis (real values altered)

x	y	ecoli	colifom	agri	develop	soil_high	soil_low	geo_high	geo_low	moraine	no_morai	cow_dens
-82.839207636	42.951157076	0	0	282660.66	0.00	0.00	282660.66	0.00	282660.65	0.00	282660.66	0.0120771
-82.371190309	42.147800229	0	1	282660.65	0.00	0.00	242181.93	0.00	221607.25	0.00	282660.66	0
-82.617938202	42.989988434	0	0	1966.32	0.00	88479.34	194181.32	0.00	282660.64	0.00	282660.66	0.0193273
-82.184265670	42.003358008	0	0	282660.66	0.00	10340.61	272320.05	282660.67	0.00	0.00	282660.66	0.0065725
-82.058055094	42.933426279	1	1	282660.64	0.00	0.00	282660.66	0.00	282660.64	0.00	282660.66	0.0065725
-82.551290000	42.785940000	0	0	282660.65	0.00	0.00	282660.66	0.00	282660.66	42265.25	240395.40	0.0094727
-82.748820000	42.353847768	0	1	207920.45	74740.21	0.00	146539.33	282660.65	0.00	0.00	282660.66	0.0094727
-81.871110000	42.199610000	0	1	282660.67	0.00	0.00	282660.66	0.00	282660.71	0.00	282660.66	0.0549425
-81.100647000	42.713730000	1	1	163438.03	0.00	151362.14	131298.53	282660.67	0.00	0.00	282660.66	0.0626677
-81.934875789	42.382629460	0	1	282660.66	0.00	181995.55	100665.11	170247.45	81502.36	0.00	282660.66	0.0626677
-81.795046108	42.483623783	0	1	282660.66	0.00	0.00	266844.24	156862.70	125798.01	282660.63	0.00	0.0791804
-81.360568960	42.700602064	0	0	282660.66	0.00	226271.33	56389.33	282660.64	0.00	0.00	282660.66	0.0626677
-81.690782836	42.868962267	0	0	201983.02	0.00	17024.07	114461.38	0.00	146922.72	0.00	282660.66	0.0791804
-81.745010150	42.619499547	1	1	282660.67	0.00	74187.54	193857.99	11843.74	270816.88	0.00	282660.66	0.0224816
-81.583699820	42.816005982	1	1	282660.65	0.00	0.00	282660.66	0.00	282660.64	0.00	282660.66	0.0791804
-81.435641270	42.781781899	0	1	282660.66	0.00	0.00	282660.66	282660.63	0.00	0.00	282660.66	0.0626677
-81.192223753	42.619327069	0	0	282660.67	0.00	0.00	137771.65	0.00	177180.62	0.00	282660.66	0.0224816
-81.442740390	42.117664000	1	1	282660.66	0.00	22035.66	260625.00	0.00	282660.68	0.00	282660.66	0.0626677
-81.114725480	43.952108896	0	0	177836.72	0.00	0.00	188669.88	91520.05	118804.05	0.00	282660.66	0.105281
-81.872220314	45.227191209	0	1	0.00	0.00	0.00	0.00	171238.64	92331.05	0.00	282660.66	0.0449031

where:

x =	X coordinate (UTM)
y =	Y coordinate (UTM)
ecoli =	*E. coli* presence or absence (binary code 1 = presence, 0 = absence)
colifom =	Total Coliform (TC) presence or absence (binary code 1 = presence, 0 = absence)
agri =	area of agricultural land in buffer zone (m^2)
develop =	area of developed land (m^2)
soil_high =	area of high infiltration soil type (m^2)
soil_low =	area of low infiltration soil type (m^2)
geo_high =	area of surficial geology with high permeability (m^2)
geo_low =	area of surficial geology with low permeability (m^2)
moraine =	area of moraine material (m^2)
no_morai =	area of non-moraine material (m^2)
cow_dens =	cow density (number of animals per km^2 in CCS)

APPENDIX 6: Multivariate Logistic Regression Formula

E. coli, 2004

Formula for the strongest multivariate logistic regression model produced:

logit(p) = 0.33478 + (0.00000166 x A) + (0.00000163 x LS) + (-0.00000286 x HG) + (-0.00000183 x LG) + (-1.59 x TS) + (-0.000000772 x NC)

Where:
p = probibility of the presence of *E. coli*
A = agricultural land
LS = Low porosity soil
HG = Highly permeable geology
LG = Low permeability geology
TS = Tilled Soil
NC = Non-carbonate geology

Entropy = 0.036
Concentration = 0.048