SPATIALLY EXPLICIT LOAD ENRICHMENT CALCULATION TOOL AND CLUSTER ANALYSIS FOR IDENTIFICATION OF *E. coli* SOURCES IN PLUM CREEK WATERSHED, TEXAS

A Thesis

by

AARIN ELIZABETH TEAGUE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2007

Major Subject : Biological and Agricultural Engineering

SPATIALLY EXPLICIT LOAD ENRICHMENT CALCULATION TOOL AND CLUSTER ANALYSIS FOR IDENTIFICATION OF *E. coli* SOURCES IN PLUM CREEK WATERSHED, TEXAS

A Thesis

by

AARIN ELIZABETH TEAGUE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by :

Co-Chairs of Committee,	Raghupathy Karthikeyan
	Russell Persyn
Committee Member,	Raghavan Srinivasan
Head of Department,	Gary Riskowski

August 2007

Major Subject : Biological and Agricultural Engineering

ABSTRACT

Spatially Explicit Load Enrichment Calculation Tool and Cluster Analysis for Identification of *E. coli* Sources in Plum Creek Watershed, Texas. (August 2007)

Aarin Elizabeth Teague, B.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. R. Karthikeyan
Dr. R. Persyn

According to the 2004 303(d) List, 192 segments are impaired by bacteria in the State of Texas. Impairment of streams due to bacteria is of major concern in several urban watersheds in Texas. In order to assess, monitor and manage water quality, it is necessary to characterize the sources of pathogens within the watershed. The objective of this study was to develop a spatially explicit method that allocates *E.coli* loads in the Plum Creek watershed in East Central Texas. A section of Plum Creek is classified as impaired due to bacteria. The watershed contains primarily agricultural activity and is in the midst of an urban housing boom.

Based on a stakeholder input, possible sources *E. coli* were first identified in the different regions of the watershed. Locations of contributing non-point and point sources in the watershed were defined using Geographic Information Systems (GIS). By distributing livestock, wildlife, wastewater treatment plants, septic systems, and pet sources, the bacterial load in the watershed was spatially characterized. Contributions from each source were then quantified by applying source specific bacterial production rates. The rank of each contributing source was then assessed for the entire watershed. Cluster and discriminant analysis was then used to identify similar regions within the watershed for assistance in selection of appropriate best management practices. The results of the cluster analysis and the spatially explicit method were compared to identify regions that require further refinement of the SELECT method and data inputs.

ACKNOWLEDGEMENTS

I would like to thank my committee, Dr. Karthikeyan, Dr. Srinivasan, and Dr. Persyn for all of their guidance and support.

Many thanks also go to Dr. Meghna Babbar-Sebens who patiently assisted with understanding of the technical details of this research. Thank you, also, to Jenifer Jacobs of the Spatial Sciences Laboratory, for running SWAT, managing the custom land use digitization, and assisting with data presentation and management for the stakeholder committee. In addition, the Plum Creek Extension team, including Nikki Diction, Matt Berg, and Dr. Mark McFarland were instrumental in data acquisition and coordinating with the stakeholder committees.

TABLE OF CONTENTS

Page

Chapter Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Waterborne Diseases Due to Pathogens

Water is essential to the preservation and flourishing of all life forms, making it the fundamental need of every human being. Only 2.5% of the world's water resources are freshwaters and 30% of this freshwater is physically available for human use in the form of groundwater and surface water. At present, approximately 15% of the world's population live in areas of water stress, struggling to meet their drinking, cooking, and sanitation needs (Fenwick, 2006). The United Nations (U.N.) predicts that the world population will grow by 40% to 9.1 billion (UN, 2005) by 2050 resulting in increased demand for water. Moreover, each year approximately 1.1 billion people do not have access to safe water and 2.2 million die due to waterborne disease (Mintz et al., 1995). Waterborne diseases such as typhoid, cholera, hepatitis, and diarrheal diseases are caused by bacteria, viruses, and protozoa. In the United States alone, the estimated cost of waterborne illness ranges from $269 to $806 million for medical costs and $40 to $107 million in lost work and productivity (Payment and Hunter, 2001).

1.2 Total Maximum Daily Load (TMDL) Program

The United States addresses water quality issues through the Clean Water Act of 1977. This legislation authorized the Environmental Protection Agency (EPA) to set water quality standards. The EPA requires water to be monitored for pathogens, nutrients, metals, organic contaminants and other physical and chemical characteristics. Pathogen monitoring includes testing for the presence of *Escheria coli (E. coli)*, cryptosporidium, giardia, legionella, and enteric viruses (USEPA, 2006). Each state is obligated to assess

---

This thesis follows the style of *Transactions of ASABE*.

the quality of the water every two years and publish a report of all waterbodies that do not meet the water quality standards. The Texas Commission on Environmental Quality (TCEQ) does this water quality assessment through the publication of the Texas Water Quality Inventory and 303(d) list. This inventory describes the status of all evaluated surface waters and classifies the surface waters into five categories (see Table 1.1). Category five has three sub-designations:

5a) A TMDL program is underway

5b) Water quality standards are reviewed before a TMDL program is scheduled

5c) Additional data are collected before the water quality standard is reviewed or TMDL program scheduled

A TMDL stands for Total Maximum Daily Load. A TMDL program is a process that includes a scientific model and implementation plan designed to bring the water body into compliance with the water quality standards. Stream segments that are classified into Category 5 are listed on the 303(d) list as impaired water bodies. The most recent five years of monitoring data are used for the classification of streams. A stream segment is classified as impaired due to pathogens if 25% of its samples exceed 394 cfu/dL or if the geometric mean of the samples exceeds 126 cfu/dL (TCEQ, 2004). The indicator organism for pathogen impairment is *E. coli*.

Table 1.1. Categories of Water Quality Inventory

Classifications	Classification Description
Category 1	Waterbody has attained the water quality standard and its use not threatened
Category 2	Some of the designated uses are attained; there is insufficient data to evaluate remaining uses
Category 3	There is insufficient data to determine if any of the designated uses are threatened
Category 4	The water quality standard is not attained ; a TMDL program is not required
Category 5	Waterbody is on the 303(d) list and the water quality standard is not attained for multiple pollutants

Once a stream segment is listed on the 303(d) list, the state is required to establish the TMDL. The TMDL is the total amount of a pollutant that a water body can receive each day from contributing sources and still maintain the water quality standard. First, the daily load must be divided amongst the various pollutant sources present in the watershed. Then an implementation plan that addresses decreasing the pollutant load from these sources is developed. The goal of this plan is to achieve the water quality standard for the impaired segment. The steps in the TMDL process include the quantification of sources, modeling of existing conditions, and definition of reduction activities that will bring an impaired stream into compliance with the state standards (USEPA, 1999).

The implementation plan, also called a watershed protection plan (WPP), is defined by the EPA as a strategy that provides assessment and management information for a geographically defined watershed, including the analyses, actions, participants, and resources related to development and implementation of the plan (USEPA, 2005). The developed plan addresses watershed pollution in a holistic manner. Additionally, it involves interested parties or stakeholders in the process of selecting the reduction strategies. These reduction strategies, also known as Best Management Practices (BMPs), should be chosen to efficiently and economically address the pollutant sources according to local conditions.

According to the USEPA, pathogen contamination is the second most frequent reason for waterbody impairment classification on the 303(d) list, comprising over 13% of the total impairments (USEPA, 2006). In Texas, 42% of water bodies did not meet water quality standards (TCEQ, 2005). Of these impaired water bodies, 61% were listed on the 303(d) list due to pathogens (TCEQ, 2002). Out of the impaired water bodies on the 2006 Texas 303(d) list, 77% of water bodies were impaired due to bacteria (TCEQ, 2007), an increase from the 2004 303(d) list. When a waterbody is listed as impaired, it impacts the local economy due to loss of the designated use, such as recreation activities.

1.3 Characterization of Contaminant Sources

The first step in the TMDL process and development of a WPP for a pathogen-impaired stream is to characterize the sources of contamination. There are both direct and indirect methods applied to characterize the pathogen contamination of a stream. Direct methods such as bacterial source tracking and load duration curves use direct monitoring data. In contrast, indirect methods characterize pathogen sources within the watershed using census and self-reporting data.

Methods such as bacterial source tracking (BST) are used to identify the sources of *E. coli* within a stream but do not quantify nor spatially characterize the sources. Ribotyping identifies unique genetic sequences of host-specific *E. coli* for development of a watershed genetic library. Then *E. coli* in the water samples are compared to this library for identification of its source. This technique has demonstrated the ability to distinguish between *E. coli* from humans, cattle, swine, horses, chickens, turkeys, dogs, and migratory geese (Carson et al., 2001). Although this method definitively identifies the source of the *E. coli*, it is highly expensive, and does not allocate the load amongst the sources.

Another method to identify the source of *E. coli* contamination is load duration curve (LDC) analysis. Load duration curves are used to characterize water quality concerns and to describe patterns associated with the impairment (Cleland, 2003). Load duration curve methodology is designed to assess the sources of exceedances in relation to stream flow conditions. First the daily flow data is ranked in descending order. Then, for each flow instance, the percent number of days for which that flow was exceeded is calculated. The cumulative frequency curve of the flow data is plotted against the percent days exceeded to create the flow duration curve. Then the load duration curve is developed by multiplying the stream flow by the water quality standard for fecal contamination with a safety factor of 10%. This is the maximum load of *E. coli* the

stream can receive and still achieve the water quality standard at different flow conditions. In order to find the instances where exceedances occurred, the actual daily loads are calculated by multiplying the measured concentration of *E .coli* by that day's stream flow. Then the actual loads are plotted against the load duration curve. The points above the curve are exceedances. The load duration curve is divided into different flow conditions (extremely high flow, high flow, dry, and drought conditions) based on the percent days exceeded of the stream-flow. The flow condition where most of the exceedances occur is be used to characterize the source of the exceedances. This is based on the assumption that exceedances occurring in high flows are due to non-point sources and exceedances occurring during low flows are due to point sources (Cleland, 2002). This method assists in differentiating between point and non-point sources; however, it does not give further insight into the sources within each category. Furthermore, it does not provide any spatial information about the sources.

1.4 Load Allocation to Various Sources

Indirect methods allocate loads based on modeling the sources of contamination. Several load allocation methods have been developed to quantify the sources for the next step in the TMDL process. The EPA has published recommendations for assessing the source contributions including identification of sources, characterization of the sources, and estimation of the load produced by each source (USEPA, 2001). Using previously published literature values, estimations of the *E. coli* load relate the source population to the number of *E. coli* excreted per day (cfu/day).

Methods such as Bacterial Load Source Calculator (BLSC), predict the contaminant output through estimates of the source populations (Zeckoski et al., 2005), where as watershed models simulate *E .coli* transport through the watershed to the stream based on runoff estimates. The BLSC combines spreadsheet calculation of loads based on animal inventories and human populations with Hydrologic Simulation Program in

Fortran (HSPF) to simulate accumulation and die off of *E. coli* (Zeckoski et al., 2005). The watershed is divided into sub-watersheds and source populations are assigned to each sub-watershed. Then HSPF is used to model the dynamics of the *E. coli*. Division of the watershed into sub-watersheds introduces a spatial component into the analysis. However within each sub-watershed the loads are not spatially allocated. Thus BLSC is not spatially referenced throughout the watershed.

Watershed models such as the Soil and Water Assessment Tool (SWAT) and HSPF are based on modeling the runoff from a rainfall event. Based on the runoff through the watershed, the amount of contaminant entering the stream is calculated. *E. coli* fate and transport are determined by environmental conditions. Watershed models consider the spatial and temporal aspects of microbial movement into the stream (Fraser et al., 1998). These models need extensive spatially referenced input data describing the potential sources (Tian et al., 2002).

1.5 Spatially Explicit Methodology

Spatially explicit analysis is needed to investigate the location of the sources of a specific contaminant. By spatially referencing *E. coli* sources, the potential load at each location in the watershed is determined. Information of the load distribution throughout the watershed can then be combined with watershed modeling to determine the amount of *E. coli* that will be transported by runoff to the stream. With this information, BMPs can specifically target areas that will contribute to stream impairment. In addition, the BMPs are designed to target the prominent sources thus increasing the efficiency of the watershed protection plan. Unfortunately, detailed date concerning the distribution and population of sources is scarce.

The proposed spatially explicit tool, Spatially Explicit Load Enrichment Calculation Tool (SELECT), identifies and distributes the various potential sources of *E. coli* in the

watershed. The locations of point sources of *E .coli* such as Wastewater Treatment Plants are first identified. Then the density or total populations of the non-point source populations are estimated. Data pertaining to livestock and human populations can be acquired from census inventories. Wildlife data can be obtained based on wildlife studies and local knowledge. In addition, characteristics of the locations and the distribution of theses sources, such as appropriate habitat are incorporated. These characteristics are then used to identify the spatial areas and densities of the point and non-point sources. Application of SELECT is particularly useful in areas with limited data concerning the population and location of contaminant sources.

1.6 Statistical Clustering

In order to extend the utility of the spatially explicit methodology, statistically unique areas are identified through clustering. This statistical characterization of the watershed supports WPP development and implementation. Factor and principal component analysis examine the different variables associated with the allocation and calculation of the load in order to reduce the number of variables, while retaining the variability of the data. This will decrease the cost of data acquisition in the TMDL process. Then the identified factors of variables are evaluated in discriminant analysis to determine their ability to distinguish the different clusters identified by cluster analysis. The unique factors identified in this statistical process will be considered and addressed in development of the WPP.

1.7 Objectives of the Research

The objective of this research study was to develop a spatially explicit tool that would statistically allocate loads from different contaminant sources. This tool was applied to Plum Creek Watershed in Texas to specifically allocate *E. coli* loadings from various contributing sources. Then the characteristics of the watershed and the allocated loads

were statistically analyzed to determine clusters of similar regions and the variables that distinguish these clusters. This holistic process provides decision support for the development and successful implementation of the WPP.

CHAPTER II

DEVELOPMENT AND APPLICATION OF SPATIALLY EXPLICIT LOAD ENRICHMENT CALCULATION TOOL (SELECT) FOR IDENTIFICATION OF *E. coli* SOURCES

2.1 Introduction

The Clean Water Act authorizes the United States Environmental Protection Agency (USEPA) to set water quality standards. To ensure compliance with the standards set by the USEPA, the Total Maximum Daily Load (TMDL) process was developed. It establishes the allowable pollutant loading for a waterbody based on the relationship between pollutant sources and water quality conditions (USEPA, 1991). The steps in the TMDL process include quantification of sources, modeling of existing conditions, and the definition of reduction activities that will bring an impaired stream into compliance with state standards (USEPA, 1999). In Texas, surface water quality standards are set by the Texas Commission on Environmental Quality (TCEQ) (TCEQ, 1997) as codified by the Texas Administrative Code Title 30 Chapter 307. To assess water quality conditions, five years of a stream segment's monitoring data is reviewed. The data is compared to a standard criterion for support of a particular water use. The number of exceedances, determines whether a stream's quality fully supports, partially supports, or does not support its designated water use. If a stream segment does not support its designated use it will be listed as impaired on a list known as 303(d) list. The 303(d) list is published biannually, to report a state's impaired surface waters.

In Texas 61% of the stream segments listed on the 303(d) list are impaired due to pathogens (TCEQ, 2005). *E. coli* is used as the indicator organism for pathogens from fecal contamination (USEPA, 1986). Indicator organisms are used because they eliminate the need to test water for all potential pathogens. They should be easy to

detect and quantify as well as have similar survival characteristics as the pathogens of concern (Zhang and Lulla, 2006). The TCEQ set an *E. coli* limit of a geometric mean of 126 cfu/dL or a single grab sample of 394 cfu/dL (TCEQ, 2004). For the TMDL process addressing pathogen contamination, the USEPA published recommendations to assess *E. coli* source contribution and identification, characterize of the sources, and estimate the *E. col*i load produced by each source (USEPA, 2001). The location and densities of *E. coli* contributing sources are identified in order to characterize the loads.

The USEPA recommends characterizing non-point sources by multiplying individual species' excretion rates by corresponding species' population (USEPA, 2001). Then the estimates of non-point sources are combined with calculated point source contributions. Previous efforts have automated this non-spatial methodology using a spreadsheet program by dividing the watershed into smaller management units or sub-watersheds (Zeckoski et al., 2005). Direct methods estimate the bacterial sources by stream monitoring including ribotyping, which use genetic testing to find the source of the bacteria (Carson et al., 2001; Ahmed et al., 2005). Load duration curves narrow the cause of potential exceedances to either point or non-point sources. This method uses direct monitoring data of the stream flow and bacterial concentrations (Cleland, 2002; Bonta and Cleland, 2003). These two methods do not spatially reference the sources and thus limit the application within the Watershed Protection Plan (WPP) because they do not provide information regarding the optimal placement of BMPs.

The objectives of this study were to develop a Spatially Explicit Load Enrichment Calculation Tool (SELECT) for the characterization of *E. coli* sources and to apply this tool to Plum Creek Watershed in Texas for the TMDL development process.

2.2 Plum Creek Watershed

The Plum Creek Watershed is a part of the Guadalupe River Basin and is located in East Central Texas. It encompasses a drainage area of 1028 km^2 in the counties of Hays, Caldwell, and Travis (Figure 2.1). Plum Creek has a length of 83 river km and joins the San Marcos River and eventually the Guadalupe River. The watershed ranges in latitude from 29º38'33.94"N to 30º5'20.11"N and in longitude from 97º54'36.29"W to 97º27'13.60"W. Within the watershed there are several rapidly growing towns including Lockhart, Kyle, and Luling. The populations of Kyle, Lockhart, and Luling are 19,335, 12,978, and 5,704 respectively (Office of Texas State Demographer, 2006). Land use varies from urban to agriculture and oil field activities. The northern part of the watershed is primarily urban whereas the southern section has crop and animal agriculture along with oil wells. The landscape is characterized as rolling hills of pasture and cropland surrounded by scrub oak forest (GBRA 2006).

2.3 SELECT Methodology

The SELECT methodology was developed using ArcGIS 9.0 with the Spatial Analyst extension available from ESRI. This spatially explicit method divides the watershed into a raster grid of 30 m × 30 m cells. For each of the cell locations within the watershed the *E. coli* loads are estimated from the sources that are potentially present at each location. Custom land use classification was performed by the Texas A&M University Spatial Sciences Laboratory, using the 2004 National Agricultural Imagery (NAIP) aerial photographs (Figure 2.2). The Soil and Water Assessment Tool (SWAT) was used to delineate the sub-watersheds within Plum Creek (Figure 2.3).

The SELECT method identifies point and non-point sources throughout the urban and rural areas. The identified point sources are active wastewater treatment plants. Non-point sources from urban areas include urban runoff, septic failure, and dogs. Non-point

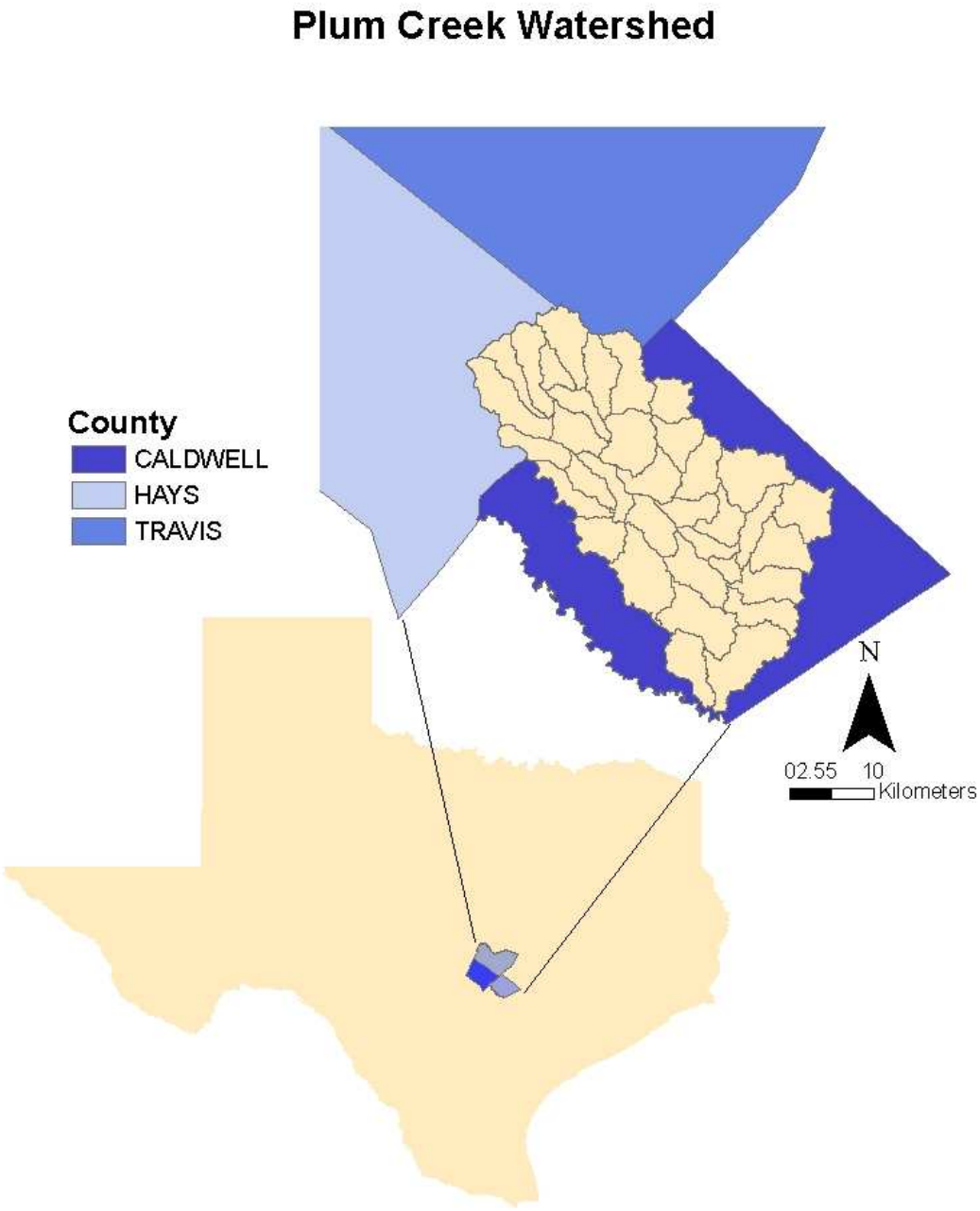

Figure 2.1. Location of Plum Creek Watershed in Central Texas.

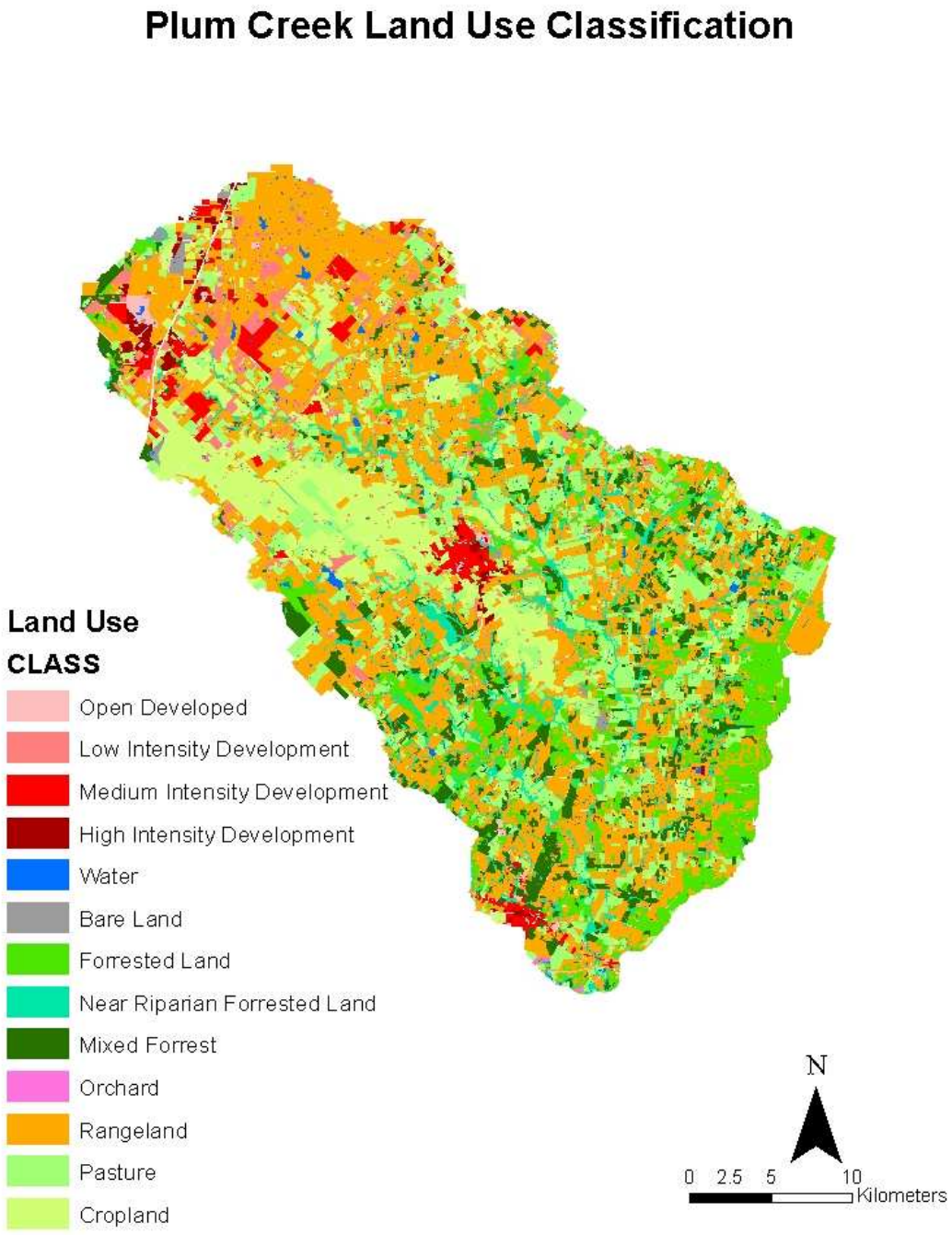

Figure 2.2. Land Use Classification of Plum Creek.

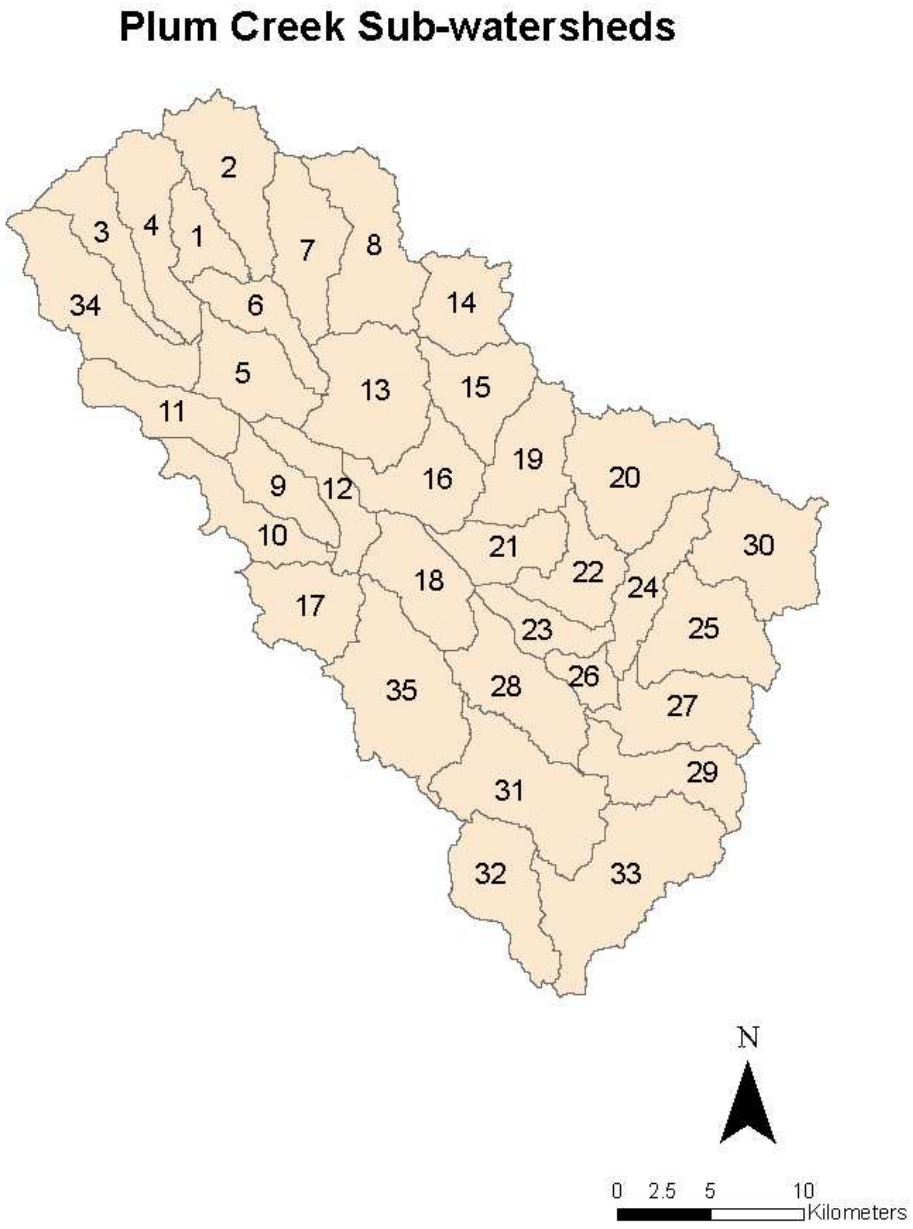

Figure 2.3. Thirty-Five Sub-Watersheds in Plum Creek

sources from rural areas include runoff from livestock, dogs (Schueler, 1999), wildlife (Weiskel et al., 1996), and septic failure (Reed, Stowe & Yanke LLC, 2001). Wildlife sources can include many types of wild animals and birds. In this study area the known wildlife includes feral hogs, whitetail deer, raccoons, rodents, opossums, and migratory birds. Feral hogs and deer were the only wildlife sources included within SELECT because they are the only populations of concern with available data. Livestock production within the study area is primarily cattle, horses, sheep, and goats.

2.3.1 Potential *E. coli* Load Calculation

Each *E. coli* source was first distributed to the appropriate locations within the watershed and then the load calculated. The average daily potential load was calculated according to USEPA guidance (USEPA, 2001). The population of sources was multiplied by a daily average fecal coliform excretion rate and then multiplied by 0.5. This 50% is a rule of thumb conversion that estimates that fifty percent of fecal coliform (FC) are *E. coli* (Doyle and Erikson, 2006).

2.3.1.1 Point Sources

Waste Water Treatment Plants

Wastewater Treatment Plants (WWTPs) are point sources permitted to discharge treated effluent into Plum Creek. There are thirteen permitted WWTPs in the watershed, however only five release effluent into the streams. Each WWTP is permitted to release effluent at the water quality standard of 126 cfu/dL. The load from each WWTP is calculated by multiplying the permitted concentration by the permitted flow in MGD (Table 2.1).

Table 2.1. Wastewater Treatment Plant Permitted Flow.

WWTP	Permitted Flow (MGD)
City of Lockhart	1.1
Lockhart 2	1.5
Luling North	0.9
City of Buda	0.3
City of Kyle	1.5

2.3.1.2 Non-Point Sources

Urban Runoff

Urban runoff includes *E .coli* that accumulates on surfaces from various sources. A study was performed by the engineering consultants PBS&J in the nearby city of Austin, Texas to measure the *E. coli* concentrations in runoff from different locations (PBS&J, 2000). Based on this data, an empirical relationship was developed to correlate the drainage area's percent impervious cover and the concentration of *E. coli* in the runoff. The percent impervious cover for Plum Creek's urban areas was determined based on the land use classification.

Using the empirical relationship reported by PBS&J (2000) the *E. coli* concentrations in the runoff were calculated. This concentration was transformed to a load by multiplying the concentration by a volume of runoff. The runoff coefficient was assumed to be one across the sub-watershed, meaning that each location contributes runoff equally to the stream. Then the average daily potential runoff was calculated from precipitation reported by the National Weather Service for an Austin weather station (NCDC, 2007). This rainfall depth was multiplied by the area of each raster cell to calculate the volume of water that would drain from each cell. This volume was then multiplied by the calculated runoff concentration for each city, resulting in an *E. coli* load from urban runoff.

Septic Failure

Septic systems can contribute pathogens to a water body due to system failure and surface or subsurface malfunction (USEPA, 2001). According to the stakeholder feedback to the Plum Creek Watershed Protection Plan Team there are a number of older failing systems within the study area however there is no local data concerning the distribution or number of failing systems. Based on a report for the Texas On-Site Waste Water Treatment Research Council, it was assumed that regulated septic systems would have a failure rate of 12% and unregulated systems would have a 50% failure rate (Reed, Stowe & Yanke LLC, 2001). On-site wastewater treatment systems were regulated starting in 1989, while systems installed prior to that remained unregulated (Lesikar, 2005).

First, the households that would utilize septic systems were estimated. Households outside of a city limit were assumed to use a domestic septic treatment system. All census blocks that fell within the watershed and were outside of a city limit were selected to calculate the number of households using septic systems. Next the number of failing systems was calculated. Subdivision data containing the number of lots and the date the subdivision was built was obtained from Caldwell and Hays counties. Both the number of houses inside and outside of a subdivision were estimated. Based on each subdivision's date built, the number of failing systems in each subdivision was calculated. All households outside of a subdivision were assumed to be non-regulated and the number of failing systems calculated accordingly.

The number of systems in each subdivision was checked to ensure that they did not exceed the number of households reported in the census. If the number of households found from subdivision data did exceed the number of households reported by the census, then the number of households reported by the census was assumed to be equal to the number of households in the subdivision.

Next the density of failing systems per raster cell was assessed. The area of each census block was found, and the density of failing systems per 900 m^2 calculated. With an estimated 70 gal/person/day discharge and a 5×10^6 cfu /dL concentration in this discharge, the *E. coli* load was calculated according to the equation in Table 2.2 and parameters converted to appropriate units. The average number per household is the average number of people in each household as reported by the 2000 U.S. Census. Then potential *E. coli* load was aggregated for each sub-watershed.

Table 2.2. Calculation of *E. coli* Loads from Non-Point Source Populations.

Source	Calculation
Cattle	$EC = \#Cattle * 2.7 * 10^9 \, cfu / day$
Horses	$EC = \#Horses * 2.1 * 10^8 \, cfu / day$
Sheep & Goats	$EC = \#Sheep * 9 * 10^9 \, cfu / day$
Deer	$EC = \#Deer * 1.75 * 10^8 \, cfu / day$
Feral Hogs	$EC = \#Hogs * 4.45 * 10^9 \, cfu / day$
Dogs	$EC = \#Households * \frac{0.8 dogs}{Household} * 2.5 * 10^9 \, cfu / day$
Failing Septic	$EC = \#FailingSystems * \frac{5x10^5 \, cfu}{100 mL} * \frac{70 \, gal}{person / day} * \frac{Ave\#}{Household} * \frac{3758.2 mL}{gal}$
WWTP	$EC = PermittedMGD * \frac{126 cfu}{100 mL} * \frac{10^6 \, gal}{MGD} * \frac{3758.2 mL}{gal}$

Dogs

Of the many pets kept by owners in Plum Creek, only dogs were considered to contribute to urban pet waste. Dog waste is a significant source of pathogen contamination of water resources (Geldreich, 1996). According to the American Veterinary Medical Association, Texans own 5.4 million dogs (AVMA, 2002, pp 1, 2, 13, 19). By dividing by the number of households in Texas, the average number of dogs

per household was found to be 0.8 dogs per household. This average is multiplied by the number of households in each block to find an estimated number of dogs per census block. Using the area of each census block, a density of dogs per 900 m^2 is found. Then the census polygons were converted to a raster and the dog density was assigned to each 30 m × 30 m cell. Published values report that dogs produce 5×10^9 fecal coliform organisms per day (USEPA, 2001). Again the 50% rule of thumb is applied to find the *E. coli* load per day from each household. The *E. coli* load was calculated according to the equation in Table 2.2. The potential *E .coli* load contribution from dogs was aggregated for each sub-watershed.

Agriculture

Rural non-point sources include agricultural range animals and wildlife. *E. coli* in animal manure can either be directly deposited into the stream or can be carried by runoff from the fields to the streams (Benham et al., 2006). Range animals such as cattle, sheep, and goats are primarily kept in pasture and on rangeland. Horses are principally confined to pasture areas. Watershed areas that were classified as pasture and rangelands were selected from digitized land use data and the areas within the city limits eliminated. The animal populations obtained from the United States Department of Agriculture (USDA) 2002 Agricultural Census were aggregated per county (USDA-NASS, 2002). This data was uniformly distributed across the remaining appropriate area of each county. Based on this distribution, a density of animals per 900 m^2 is calculated. The appropriate lands in Plum Creek were assigned these densities and multiplied by the fecal coliform excretion rate and then converted to *E .coli* potential (see equations in Table 2.2). Then *E .coli* loads were aggregated to the sub-watershed level.

Wildlife

Wildlife also contribute to the *E. coli* within Plum Creek watershed. Within the watershed major wildlife contributors include deer and feral hogs. There are many other wildlife sources, such as birds, opossums, raccoons, and coyotes. However, at the time

of analysis there was not a reliable method to estimate these populations. Deer habitat includes shrubland and forest areas. Feral hogs primarily use riparian corridors of undeveloped land uses. To distribute the deer population within Plum Creek watershed, appropriate land use areas with a continuous area of greater than 20 acres were first selected. Texas Parks and Wildlife Department (TPWD) annual surveys report a density of deer per 1000 acres for resource management units (RMUs) (Lockwood, 2005). Plum Creek falls in RMUs 7, 19, and 20. The total number of deer was calculated based on the area of Plum Creek in each RMU. With the area of appropriate land use within each Plum Creek section of the appropriate RMU, a density of animals per 900 m^2 is calculated. The RMU vector data was converted to raster format using the same extent and cell size as the land use data, with the cells assigned the deer density per 900 m^2. Then a fecal coliform excretion rate of $3.5x10^8$cfu/day-animal (Zeckoski et al., 2005) was multiplied by the deer per unit area in order to then find the *E. coli* load throughout the area (see the equation in Table 2.2). Then the potential *E. coli* load was aggregated to the sub-watershed level.

Feral hog population densities and distribution data is scarce for Plum Creek watershed. Estimates of feral hog densities for the Rio Grande Plains and lower coastal prairie of Texas ranges from 3.2 to 6 hogs/km^2 (Hellgren, 1997). Plum Creek habitat is comparable to the landscape of the Rio Grande Plains and lower coastal prairies. A landscape wide density of 5 hogs/km^2 is applied to the entire watershed to produce an estimate of 5,141 hogs for the entire watershed. These hogs were then uniformly distributed to riparian corridors, or the undeveloped and undeveloped land within 100 m to a stream. Feral hogs utilize nearly all types of landscape, but primarily use forested and shrublands adjacent to river bottomlands. Based on the number of cells with appropriate habitat, the density of hogs per cell was determined and multiplied by the fecal coliform excretion standard. This is calculated according to the equation found in Table 2.2, where 4.45×10^9 cfu/animal-day is the fecal coliform excretion rate

multiplied by the 50% rule of thumb. Then the distributed *E .coli* load was aggregated to the sub-watershed level.

2.3.2 *E. coli* Load Aggregation in Sub-watersheds

In order to give the relative ranking of the sources on a spatial basis, all sources were summed for each sub-watershed. This allows for the ranking of total potential *E .coli* load on a spatial basis. In addition, the sources were ranked by total contribution.

2.3.3 Comparison of Potential *E .coli* with Actual Monitoring

The results of SELECT were compared to the actual monitoring data collected by Guadalupe Blanco River Authority. First the sampling dates were compared with meteorlogical data reported by the National Weather Service. The dates with a reported precipitation event that was large enough to result in runoff were selected. This was determined by the NRCS curve number method (Haan et al., 1994, pp 63-65) based on the average curve number for each monitoring station. With the daily precipitation depth, a runoff depth was calculated using this method. Multiplying by the drainage area for each monitoring station resulted in a runoff volume. This runoff volume was added to the daily volume of effluent that is discharged into the drainage area for a total stream flow volume. The daily average potential *E. coli* load was divided by the total flow volume to calculate the potential concentration. This potential concentration was then compared to the actual monitored concentration. It is important to note that SELECT results and actual monitoring data can only be compared when there is a runoff event. This is because the *E. coli* loads estimated from the non-point sources in SELECT only enter the streams when there is a runoff event.

2.4 Results and Discussion

The results from SELECT for all sources are found in Figures 2.4 through Figure 2.14. The larger loads are found in the darker shaded sub-watersheds. The mid-range loads are in the medium shaded sub-watersheds and the lowest loads in the lightest shaded sub-watersheds.

2.4.1 Urban Sources

The estimation of potential *E. coli* loads from WWTPs are found in Figure 2.4. The five sub-watersheds in which the WWTPs are located are highlighted (Figure 2.4). The higher the permitted effluent discharge, the higher the estimated potential load and the darker sub-watersheds in Figure 2.4. Best management practices such as tertiary treatment (Godfree and Farrell, 2005) or overflow monitoring , for WWTP would most efficiently be designed for the sub-watersheds that fall near the cities of Lockhart and Kyle.

The potential contribution from urban runoff is shown in Figure 2.5. The sub-watersheds with large septic loads correspond to the high population areas around Kyle, Lockhart, and Luling. The largest loads are estimated for the sub-watersheds near the city of Kyle. Table 2.3 shows sub-watershed 34 (see Figure 2.3), which includes portions of the city of Kyle, has higher levels of high and medium intensity development and has the largest potential *E. coli* load (Figure 2.5). Low intensity developed land is defined as having from 20% to 49% impervious cover. Medium and high intensity development are defined as having impervious cover ranging from 50% to 79% and 80% to 100%. The city of Lockhart falls partially into sub-watershed 16 (Figure 2.3), which also has high percent land uses classified as high and medium intensity development (Table 2.3). The higher percent of low, medium, and high intensity development (Table 2.3) centered around urban centers corresponds to large *E. coli* load allocations (Figure

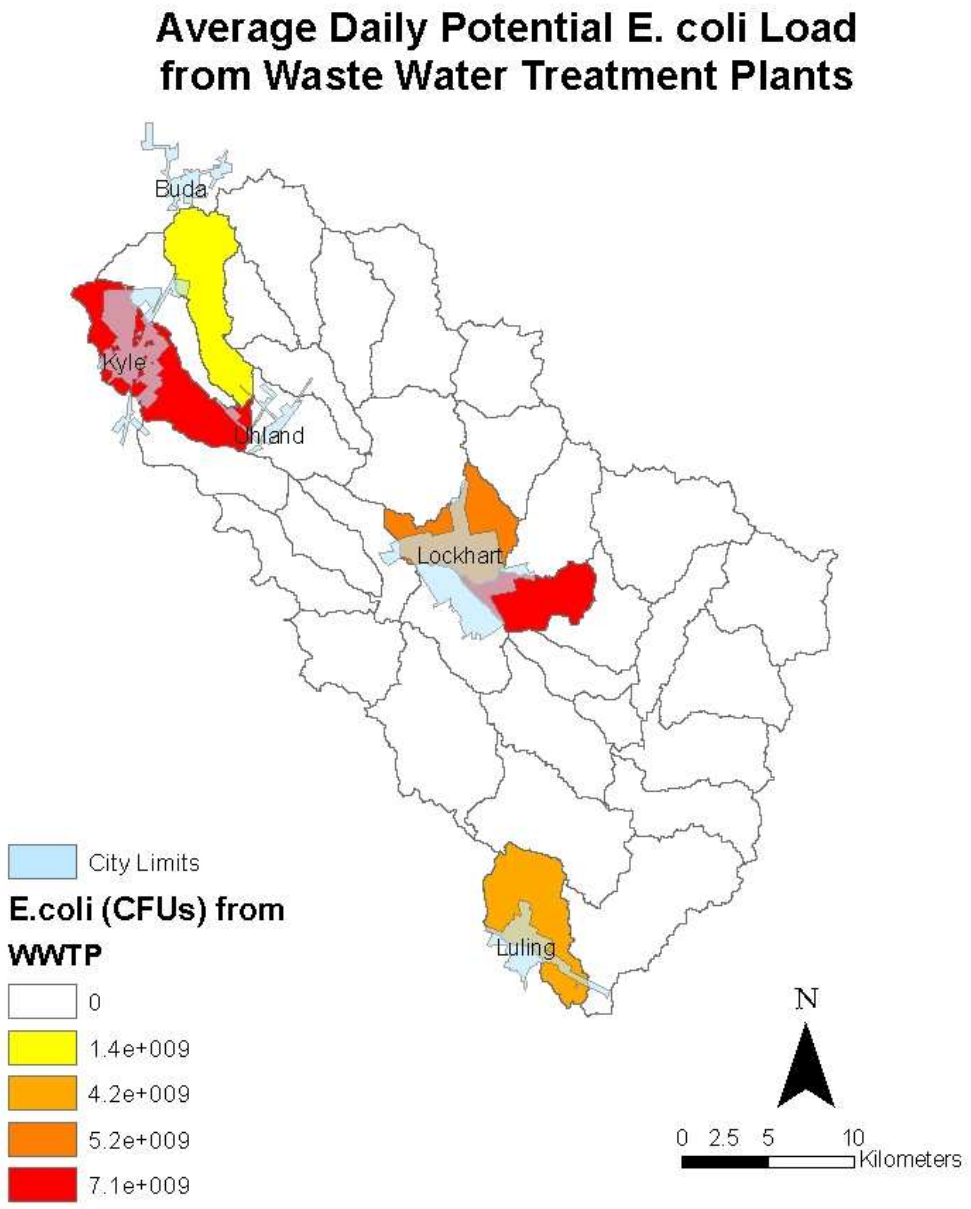

Figure 2.4. Average Daily Potential *E. coli* Loads Resulting from WWTP in Plum Creek Watershed.

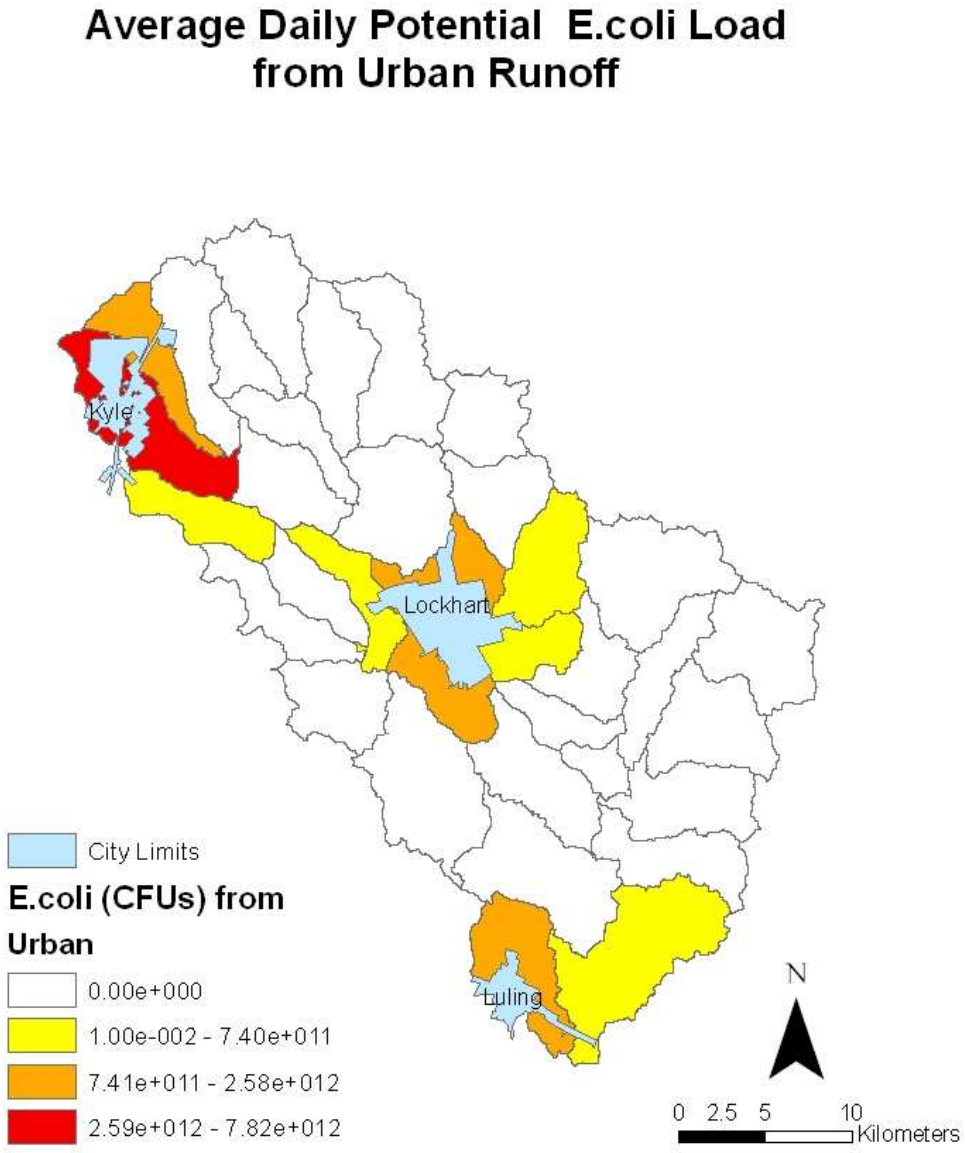

Figure 2.5. Average Daily Potential *E. coli* Load from Urban Runoff.

Table 2.3 Breakdown of Sub-Watershed Land Use

	Percent Land Use in Each Sub-Wwatershed											
Sub-Watershed	**Dev. Open Space**	**Low Intensity Dev.**	**Med. Intensity Dev.**	**High Intensity Dev.**	**Open Water**	**Bare Land**	**Forrest**	**Riparian Forrest**	**Mixed Forrest**	**Range**	**Pasture**	**Crops**
1	1.16	16.37	19.95	0.00	2.28	0.00	0.00	1.56	0.00	50.50	3.95	4.22
2	1.06	9.22	0.17	0.00	3.21	0.28	0.00	0.50	0.00	68.51	3.53	13.53
3	1.27	5.24	2.56	2.79	1.14	6.69	0.01	2.57	0.70	52.40	6.79	17.84
4	1.21	15.75	7.09	5.71	1.64	1.28	0.00	0.42	0.45	42.12	14.84	9.49
5	0.32	12.07	3.57	0.00	1.49	0.00	0.43	7.44	3.31	46.20	1.87	23.29
6	0.00	7.01	1.88	0.00	1.86	0.00	0.00	4.59	1.21	56.07	17.71	9.68
7	0.07	9.23	10.54	0.00	1.79	0.00	0.55	0.66	1.65	57.68	7.95	9.88
8	0.13	6.04	0.65	0.00	2.58	0.00	0.74	2.43	4.56	59.59	17.36	5.93
9	0.00	3.66	0.00	0.00	0.03	0.00	0.33	2.09	0.10	4.26	32.25	57.28
10	0.00	5.65	1.23	0.00	3.78	0.00	1.61	1.68	0.72	28.51	17.81	39.01
11	0.14	4.57	2.55	0.01	0.25	0.31	0.32	0.52	0.04	9.07	4.47	77.76
12	0.00	2.55	0.00	0.00	0.19	0.06	0.26	3.20	0.28	12.84	12.09	68.53
13	0.18	3.12	0.14	0.00	1.97	0.00	1.77	6.06	3.51	47.40	24.58	11.27
14	0.00	3.87	1.70	0.00	1.42	0.00	14.34	7.21	4.36	41.29	12.14	13.67
15	0.00	4.45	0.00	0.00	1.01	0.00	8.40	7.68	11.66	42.01	21.17	3.61
16	1.56	3.04	17.13	2.88	0.51	1.06	3.83	3.64	2.57	16.83	19.19	27.76
17	0.00	4.05	0.00	0.00	1.06	0.00	2.64	5.66	13.53	55.32	15.38	2.35
18	0.60	3.48	6.10	2.84	0.45	0.11	2.67	14.26	6.44	25.53	20.96	16.59
19	0.00	2.61	0.03	0.03	1.05	0.34	9.76	7.75	7.37	32.88	31.13	7.05
20	0.00	2.12	0.07	0.00	1.34	0.03	12.59	12.34	14.07	37.76	17.55	2.10
21	0.00	2.21	3.57	3.04	0.79	0.27	4.72	10.37	6.90	19.92	31.80	16.41
22	0.00	2.41	0.00	0.00	0.76	0.06	4.49	11.53	8.22	43.97	23.65	4.91
23	0.00	2.10	0.19	0.00	0.50	0.14	6.70	3.53	0.63	32.71	20.27	33.22
24	0.00	1.13	0.02	0.00	1.41	0.00	7.73	7.30	13.86	34.49	28.70	5.35
25	0.00	1.15	0.00	0.00	0.64	0.01	25.19	6.50	11.18	28.91	20.27	6.15
26	0.00	0.81	0.00	0.00	0.41	7.16	19.61	6.29	6.84	15.85	39.56	3.46

Table 2.3 Continued.

Sub-Watershed	Percent Land Use in Each Sub-Watershed											
	Dev. Open Space	Low Intensity Dev.	Med. Intensity Dev.	High Intensity Dev.	Open Water	Bare Land	Forrest	Riparian Forrest	Mixed Forrest	Range	Pasture	Crops
27	0.00	0.61	0.00	0.00	0.78	0.33	22.31	7.34	16.20	34.49	17.61	0.35
28	0.00	1.27	0.00	0.00	0.34	0.94	18.63	10.84	8.85	26.35	18.20	14.58
29	0.00	0.73	0.00	0.63	1.52	0.22	27.55	5.21	6.29	37.60	14.92	5.33
30	0.00	1.57	0.00	0.00	1.87	0.00	11.03	8.29	12.19	44.33	19.75	0.97
31	0.33	1.24	0.12	0.00	0.98	0.04	12.81	8.09	13.64	44.03	18.44	0.27
32	2.44	1.93	7.71	0.94	0.81	0.54	5.18	6.16	24.13	40.44	9.72	0.00
33	0.32	1.24	0.32	0.00	0.87	14.11	6.64	13.93	0.10	43.19	18.75	0.53
34	6.42	8.31	15.82	9.64	1.32	0.34	0.67	2.73	2.41	29.67	9.23	13.44
35	0.00	1.42	0.09	0.00	1.32	0.09	21.93	11.83	9.78	39.50	13.98	0.08

2.5). Best management practices for urban runoff, such as detention ponds, filter strips, and artificial wetlands (Braune and Wood, 1999) should be designed for urban centers with high percentages of medium to high intensity development.

The estimated potential *E. coli* load from septic failure is shown in Figure 2.6. The darker sub-watersheds indicate the larger estimated potential *E. coli* load. Larger loads (2.13×10^{10} to 2.34×10^{12} cfu) are associated with sub-watersheds that correspond to the cities of Lockhart and Kyle. However, large loads are also associated with sub-watersheds one, two, four, and seven (Figure 2.6). These sub-watersheds have high percentages of low intensity development as shown in Table 2.4. The area in sub-watersheds one, two, four, and seven (Figure 2.3) in the north of the watershed have a large population reported in the 2000 census, which is not yet incorporated into a city (Figure 2.6) and thus not provided with sewer service. In addition, the average age of the subdivisions in sub-watersheds one, four, and seven are all pre-1988. As a result, the septic systems in these sub-watersheds are unregulated. Therefore BMPs should be designed to address regulation of septic systems, focusing on proper operation and owner maintenance of the system (Lesikar, 2005) within this region.

The potential *E. coli* load estimated from dogs is shown in Figure 2.7. Sub-watersheds with large allocations are associated with the cities of Kyle, Lockhart , and Luling. This can be attributed to the large number of households in the urban areas. In addition, like the septic estimation, the sub-watersheds of one, two, four, and seven are estimated to have higher potential loads of *E. coli*. This area has higher population in comparison to the rest of the sub-watersheds, despite the lack of urban centers. The higher population of this area is attributed to urban sprawl from the nearby metropolitan area of Austin. Best management practices such as pooper scooper programs and dog owner education (Kemper, 2000) should be implemented not only in the cities of Kyle, Lockhart, and Luling, but also in the areas where urban sprawl is a concern, primarily in the northern portion of the watershed.

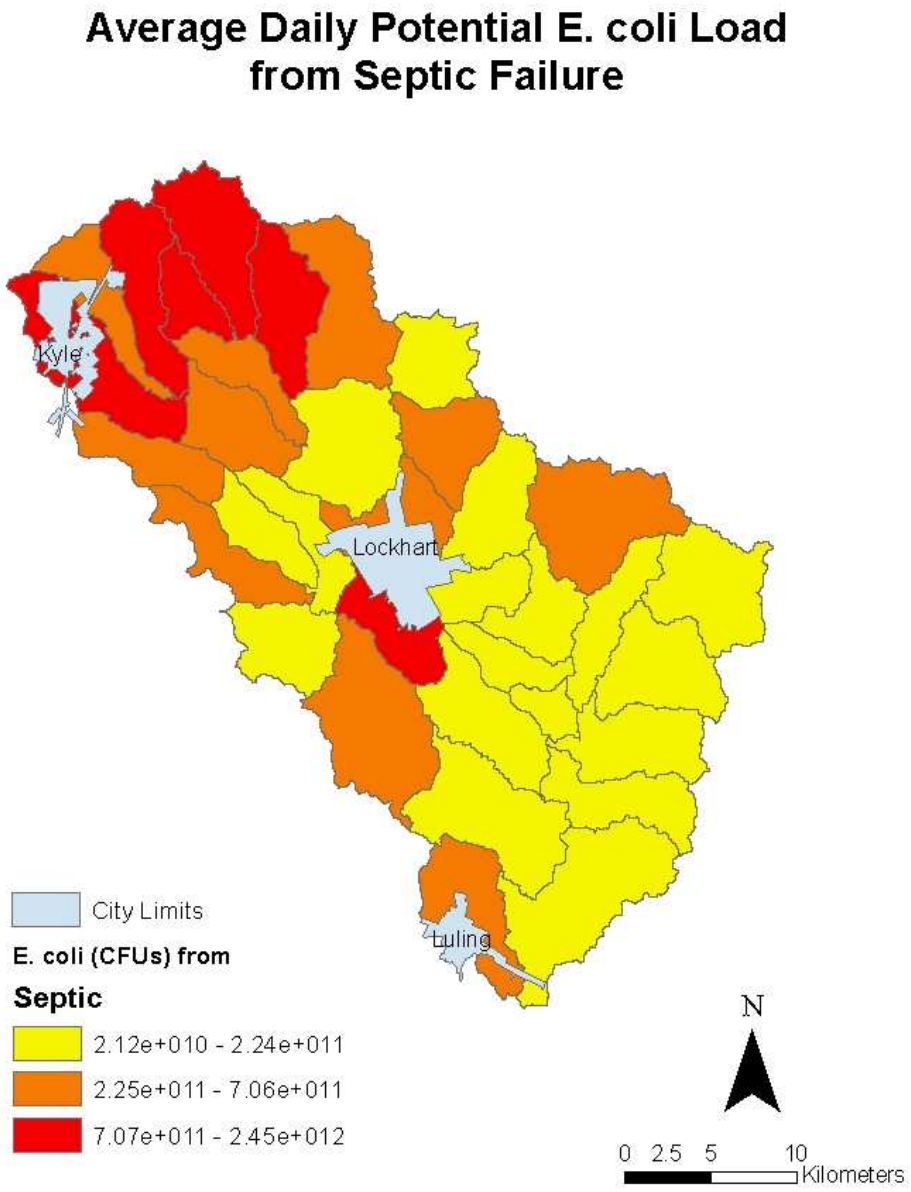

Figure 2.6. Average Daily Potential *E. coli* Loads Resulting from Septic Failure.

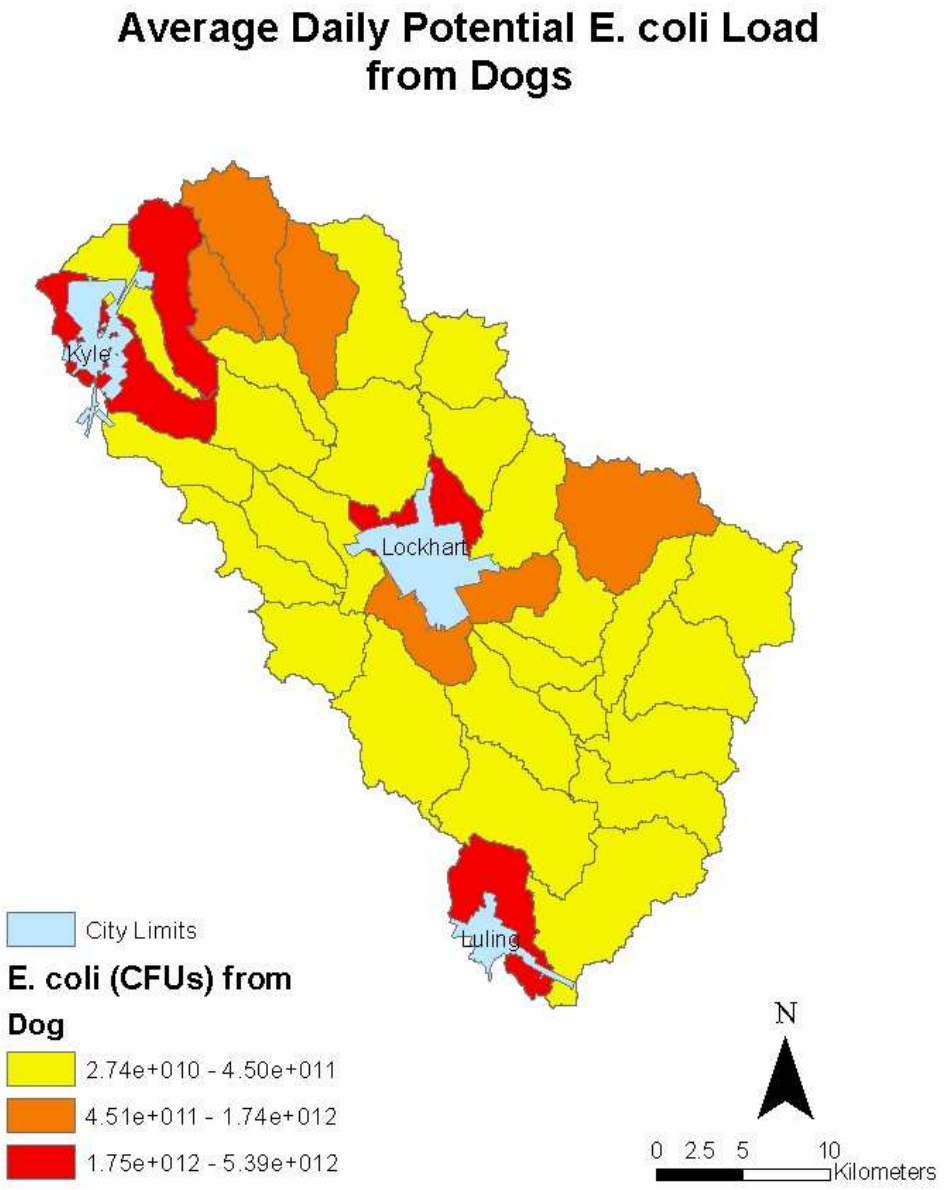

Figure 2.7. Average Daily Potential *E. coli* Load Resulting from Dogs.

2.4.2 Rural Sources

Rural sources include agricultural animals, wildlife, dogs, and septic failure. The load allocations from cattle, sheep and goats, and horses are in Figures 2.8, Figure 2.9, and Figure 2.10, respectively. Greater *E. coli* loads from cattle are estimated for sub-watersheds on the southwestern portion of the watershed and along the southeastern edge (Figure 2.8). The sub-watersheds which have larger estimated loads of *E. coli* from cattle have higher percentage of land used for pasture and rangeland (Table 2.3) and generally are larger in area. In contrast the high *E. coli* potential sub-watersheds for sheep and goats are in the north of the watershed (Figure 2.9). Like cattle, these sub-watersheds have a high percentage of pasture and rangeland (Table 2.3). Sub-watershed 34 is the exception with a low percentage of pasture and rangeland, however its large load is due to sub-watershed 34 having a larger area. The *E. coli* loads from sheep are estimated to be primarily in the northern part of the watershed, whereas the *E. coli* loads from cattle are estimated to be primarily in the southern portion of the watershed because according to the USDA census there is greater sheep and goat production in Hays and Travis counties and greater cattle production in Caldwell county (USDA-NASS, 2002). Potential loads estimated from horses are primarily found in the southern and middle section of the watershed (Figure 2.10). These high potential sub-watersheds have large areas of pasture lands (Table 2.3). When the total loads allocated to cattle, sheep and goats, and horses are compared (Figures 2.8, 2.9, and 2.10), the magnitudes are quite different. The total estimated potential loads for cattle (Figure 2.8) and sheep and goats (Figure 2.9) are two orders of magnitude larger than the estimated load for horses (Figure 2.10). Because of the higher population of cattle, cattle have a larger potential load than sheep and goats. Agricultural BMPs such as riparian fencing, vegetative filter strips, and alternative watering (Anderson and Flaig, 1995), should be prioritized in the southern section of the watershed for cattle producers and sheep and goat producers in the northern section of the watershed.

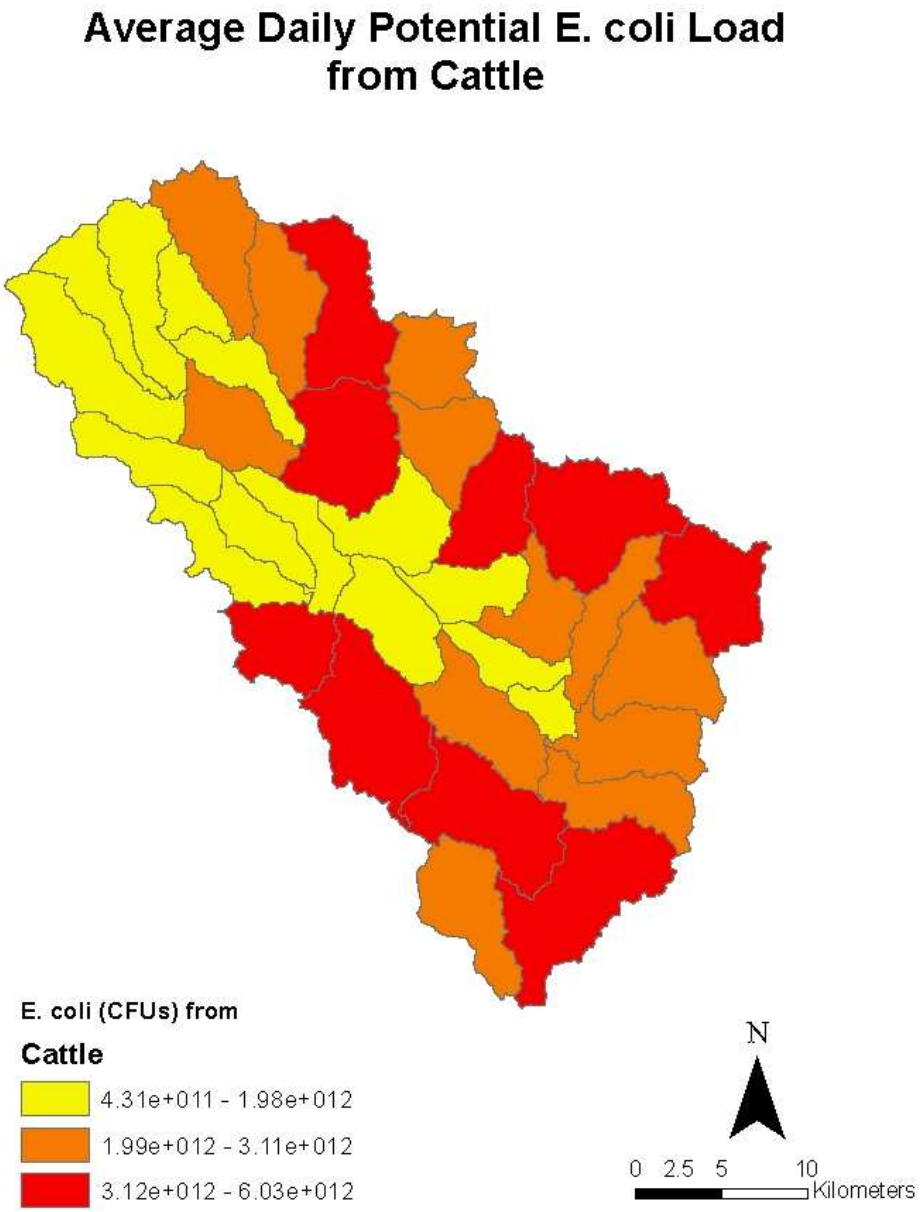

Figure 2.8. Average Daily Potential *E. coli* Load from Resulting from Cattle.

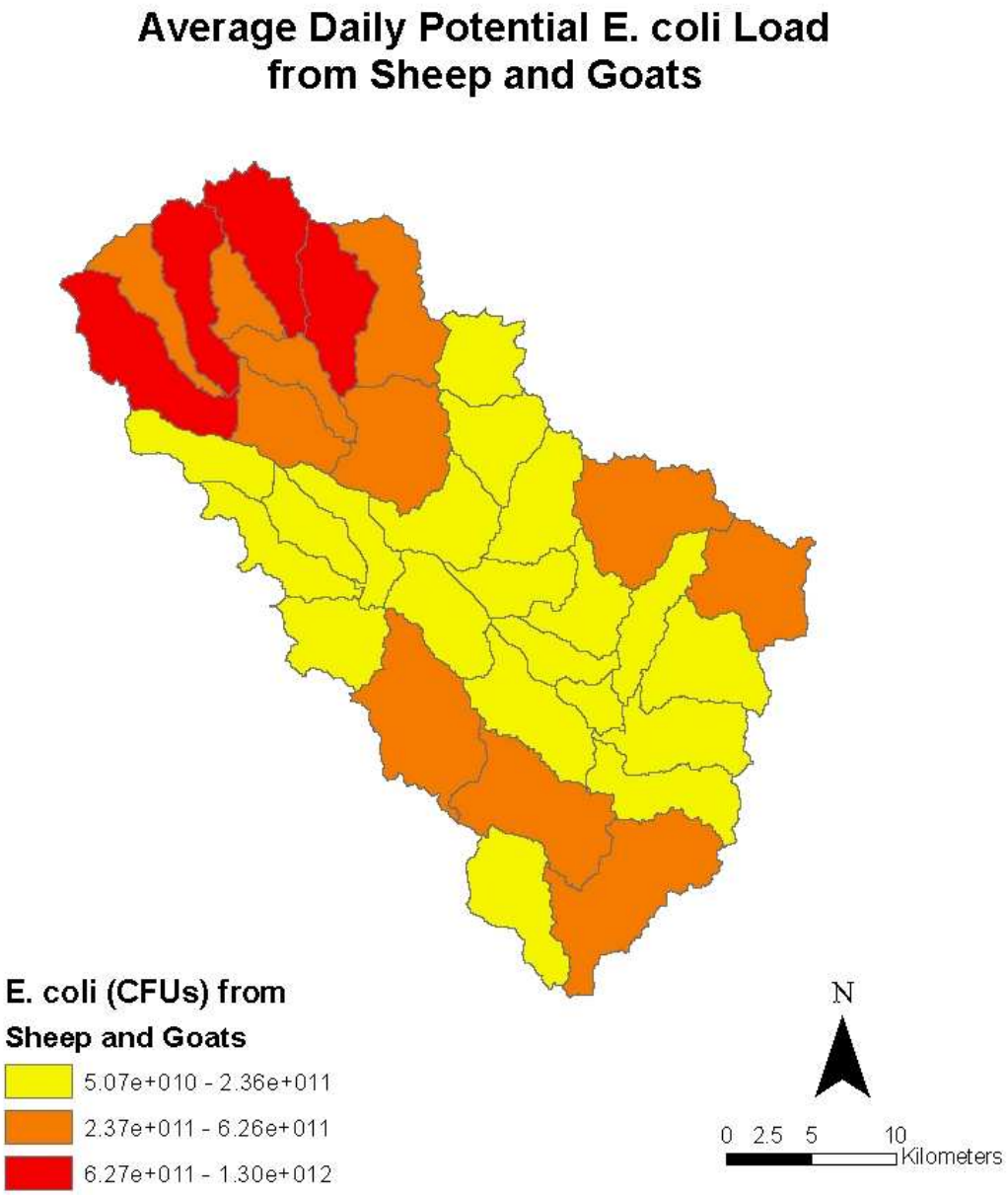

Figure 2.9. Average Daily Potential *E. coli* Load Resulting from Sheep and Goats.

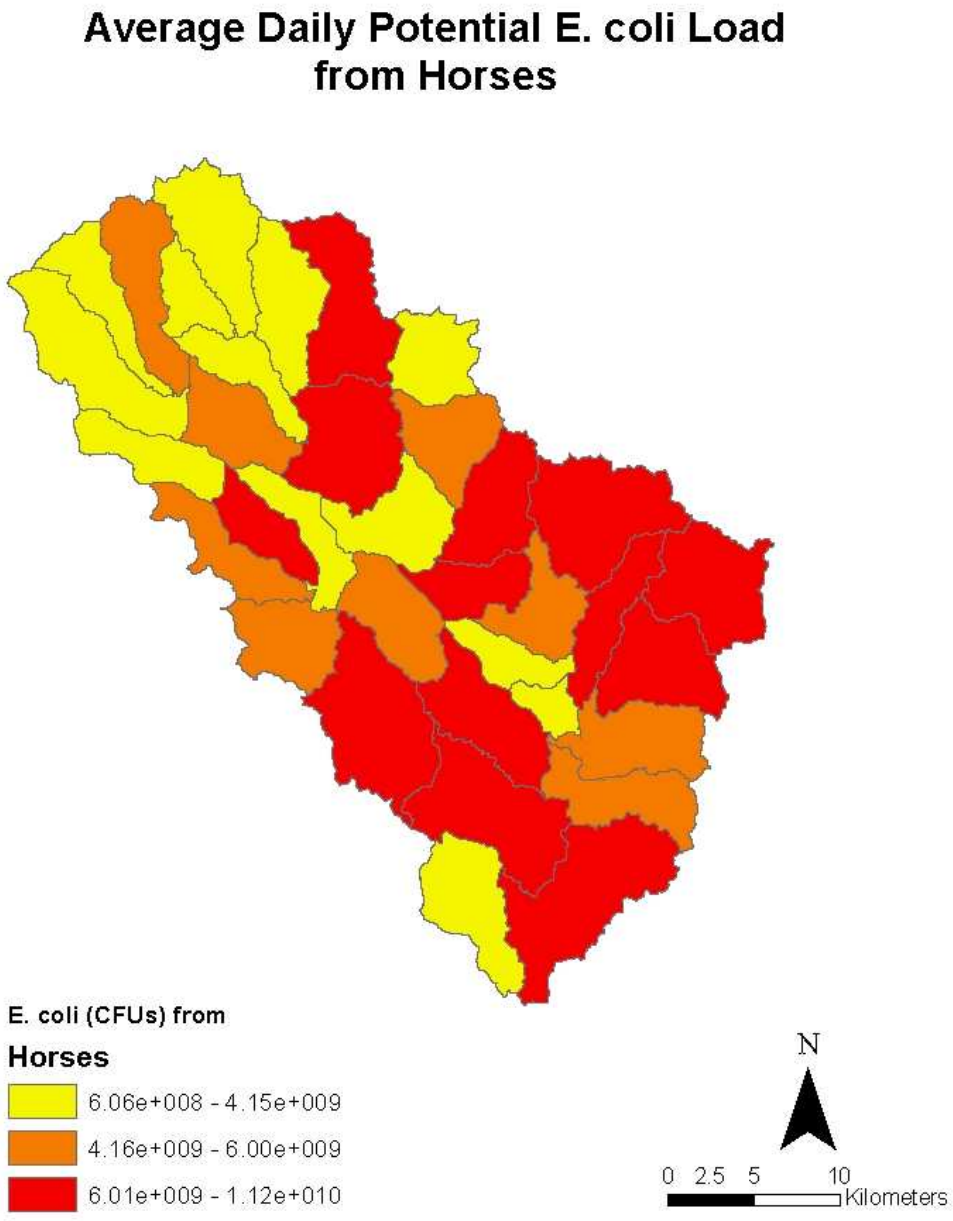

Figure 2.10. Daily Average Potential *E. coli* Load Resulting from Horses.

The potential *E. coli* estimated from feral hogs is in Figure 2.11. As stated in the methodology, the feral hogs were distributed to the riparian areas around streams. Each sub-watershed has an estimated potential contribution from feral hogs. The highest potential loads are in areas along the east and south of the watershed (Figure 2.11), where there is a larger area of undeveloped land adjacent to a stream. Feral hogs have an estimated potential load (Figure 2.11) that is the same magnitude as cattle (Figure 2.8) and sheep and goats (Figure 2.9). Unfortunately, the best management practices to address *E. coli* contamination from feral hogs are quite challenging because fencing and other traditional practice are not practical in addressing this source population. Feral hogs are highly invasive and destroy agricultural crops and riparian vegetation (Baron, 1982). Therefore landowner education, and population control are the most appropriate measures to implement in the southern portion of the watershed.

The potential *E. coli* load from deer is shown in Figure 2.12. The south-eastern portion of the watershed has the highest loads from deer where there are large sections of range and forested areas. The estimated potential load for deer (Figure 2.12) is two orders of magnitude smaller than the estimated load for feral hogs (Figure 2.11). Wildlife BMPs are more efficiently focused on addressing feral hogs than deer.

2.4.3 Potential *E. coli* Sources Throughout the Watershed

Two sources, septic and dogs are considered to be both urban and rural sources. However because these sources are associated with human populations, the larger estimated loads will correspond to population centers. In urban areas, the contributions will not only be larger in magnitude but also concentrated to a small area. In the rural areas, these sources are diffuse and smaller in magnitude (number of cfu in potential load from each sub-watershed). The WPP should address these sources across the entire watershed. In urban areas a total approach can be taken for dogs and septic. Large BMPs that are structural in nature, such as detention ponds that would collect runoff, are

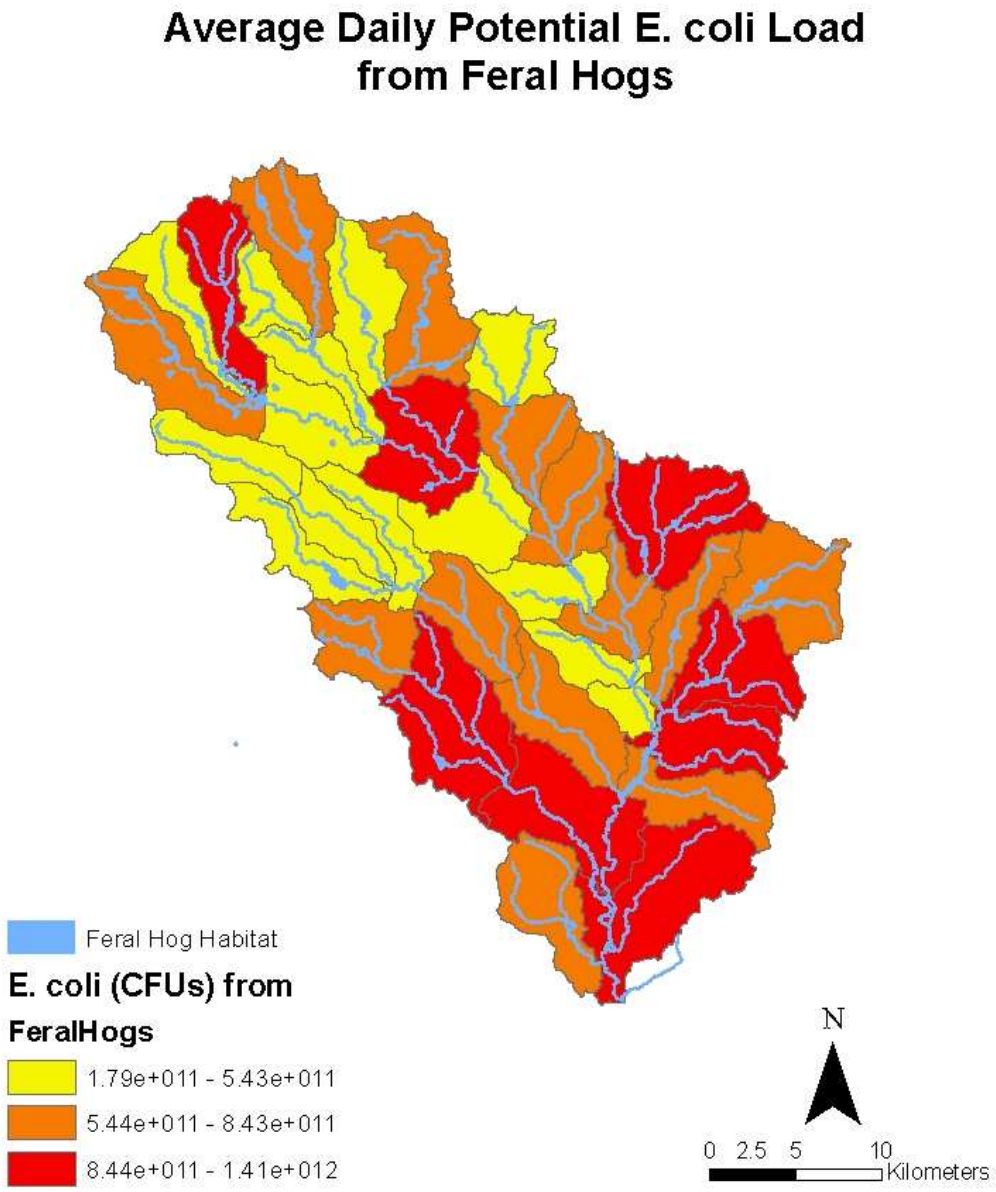

Figure 2.11. Daily Average Potential *E. coli* Load Resulting from Feral Hogs.

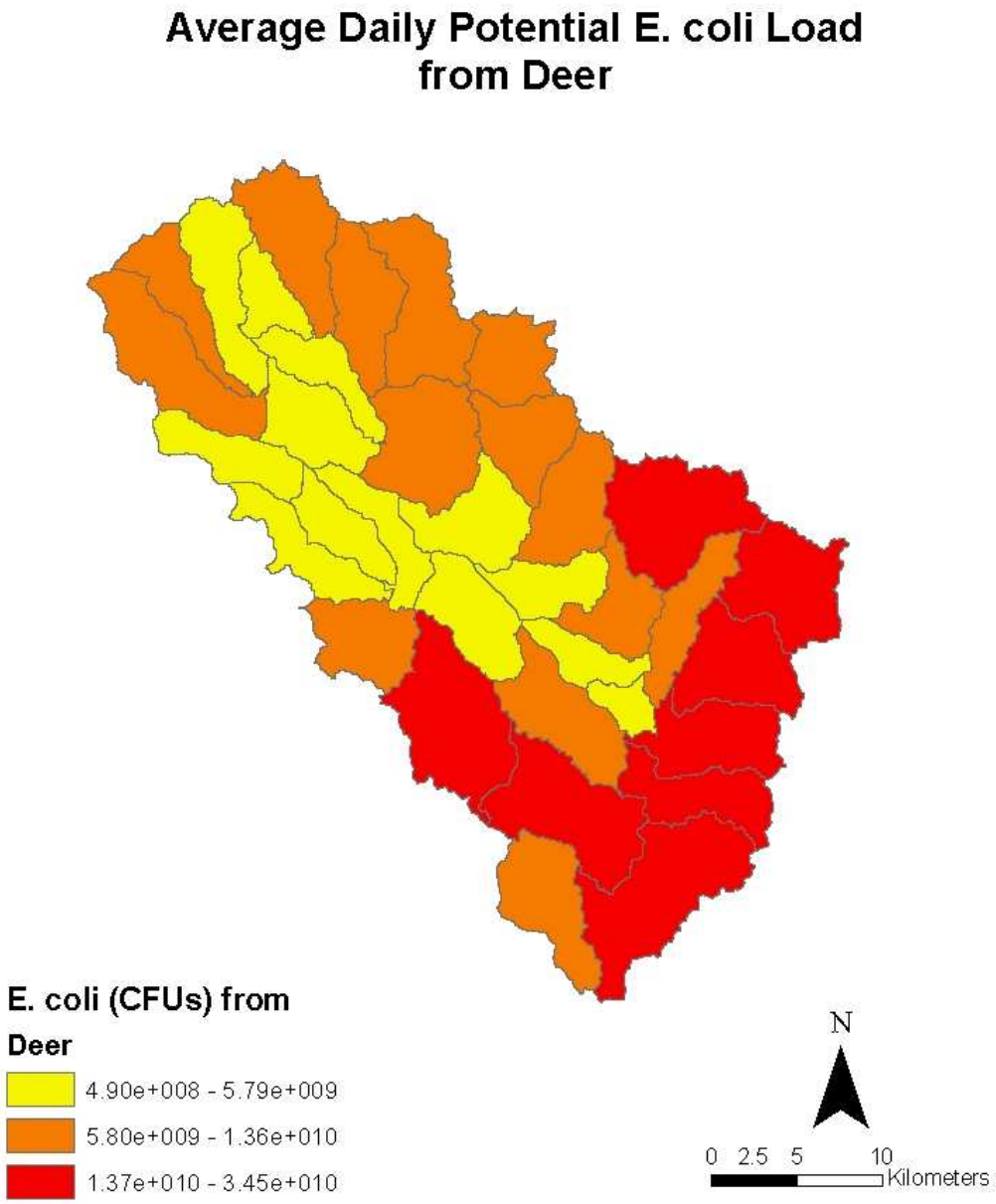

Figure 2.12. Daily Average Potential *E. coli* Load Resulting from Deer.

efficient in the urban areas due to the magnitude of the load. For rural areas, homeowner education should be implemented to increase septic maintenance, but should focus particularly on residences near to streams.

The total estimated sub-watershed loads are shown in Figure 2.13. The darker sub-watersheds (4.07×10^{11} to 1.87×10^{12}cfu) have the highest estimated potential load. These four sub-watersheds each correspond to urban areas, and have area incorporated into the cities Kyle, Lockhart, and Luling. The medium color, or mid range estimated loads (1.88×10^{12} to 4.06×10^{12}cfu), are highly influenced by regional effects (Figure 2.13). Figure 2.13 shows the relative contribution of each source to the total estimated load (Figure 2.13) for each sub-watershed. The mid range load sub-watersheds in the northern section of the watershed (Figure 2.13) show mixed influence of septic, dog, and agricultural animal sources (Figure 2.14). The mid range load sub-watersheds in the southern and eastern portions of the watershed (Figure 2.13) have a high percentage of the load estimated from agricultural animals and wildlife sources (Figure 2.14).

Table 2.4 displays the sub-watersheds with the highest potential *E. coli* and the high potential sources within each of these sub-watersheds. Table 2.5 displays the five sources with the highest total potential and the sub-watersheds that have the highest potentials for each of these sources. Overall, cattle have the highest potential contribution, with 41% of the total average potential *E. coli* load (Table 2.5). The second highest potential daily contributor is urban runoff with 27% of the total potential load. Dogs and feral hogs each have a potential of approximately 10.5% of the total potential load and failing septic systems comprise approximately 6.5% of the total. All other sources contribute less than five percent to the total potential load. It is notable that although SELECT did not indicate that WWTPs were a major source of *E. coli*, there is uncertainty as to whether the pathogens in the effluent are viable but non-culturable and will be reactivated once further downstream in a nutrient rich effluent

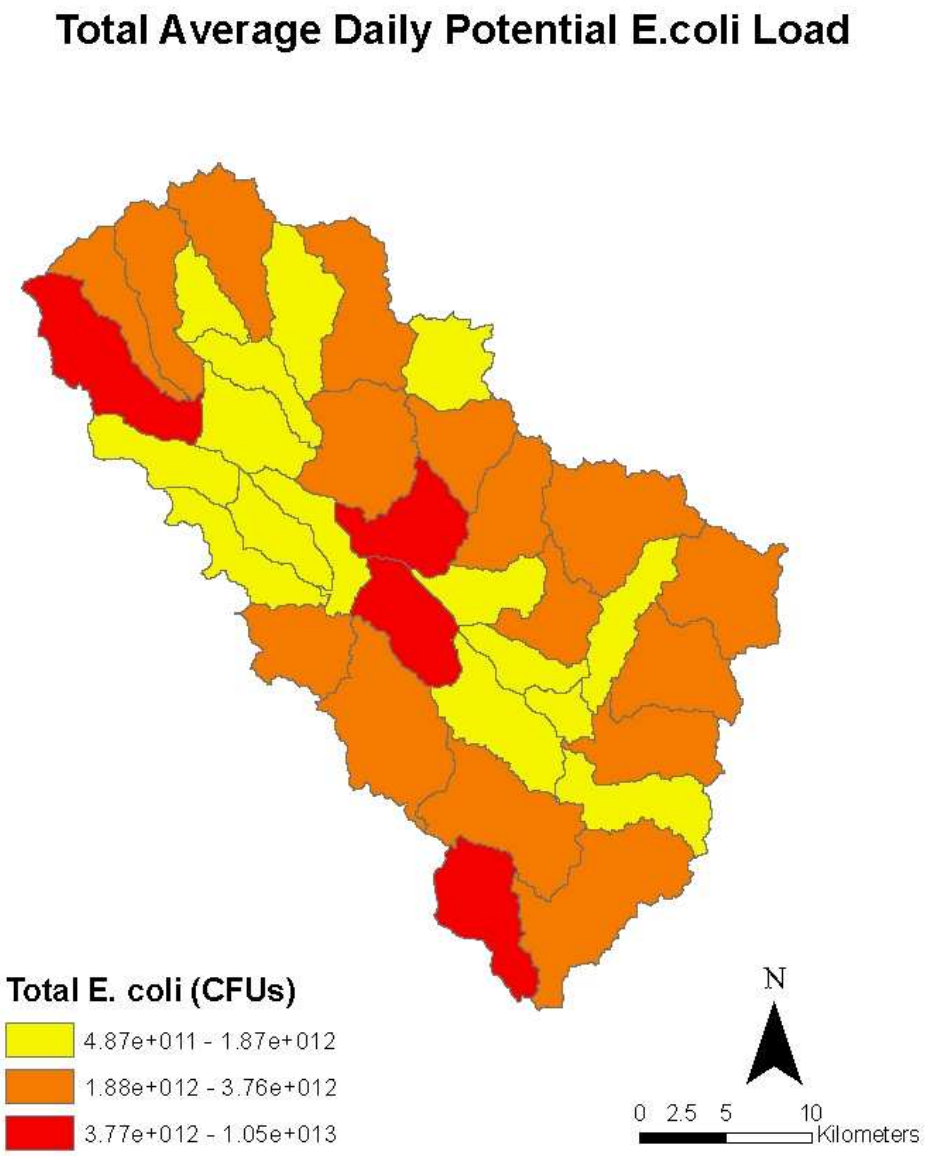

Figure 2.13. Total Potential Average Daily *E. coli* Load.

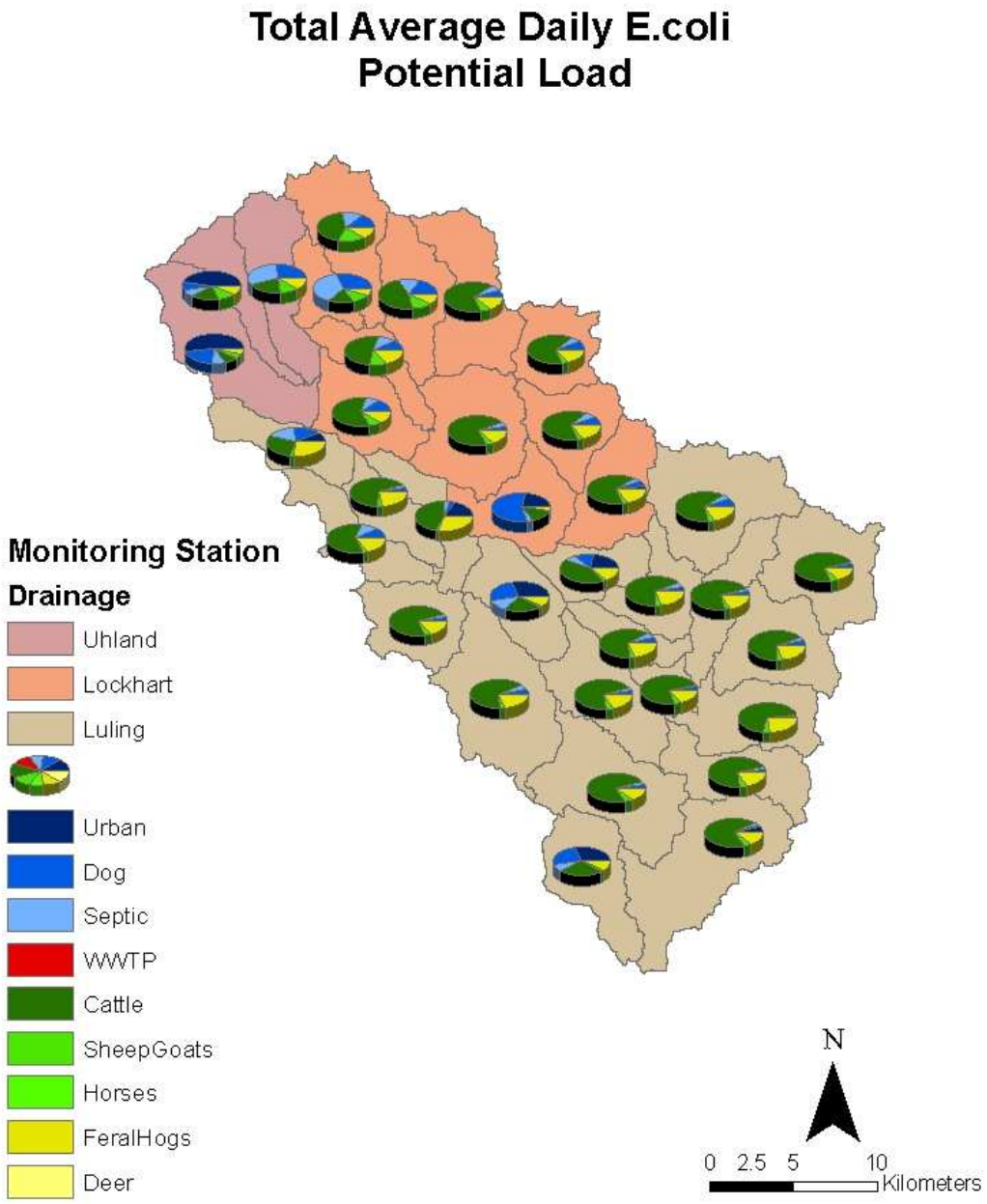

Figure 2.14. Comparison of Relative Percent Contributions from Potential Sources in Each Sub-Watershed.

(Petersen et al., 2005). A fate and transport model is needed to model these physical phenomena and understand *E. coli* population dynamics (Steets and Holden, 2003).

Table 2.4. High Potential Sources of High Contributing Sub-Watersheds.

		High Potential Contributors				
Rank	**Sub-Watershed**	**1**	**2**	**3**	**4**	**5**
1	34	Urban	Dogs	Septic	Cattle	Sheep & Goats
2	16	Urban	Dogs	Cattle	Septic	Feral Hogs
3	32	Urban	Cattle	Dog	Feral Hogs	Septic
5	18	Urban	Cattle	Dog	Septic	Feral Hogs
4	3	Urban	Cattle	Sheep & Goats	Feral Hogs	Dogs

Table 2.5. Sub-Watersheds of High Potential Sources.

	Sub-Watershed Contributors				
Source	**1**	**2**	**3**	**4**	**5**
Cattle	33	13	31	35	20
Urban	34	16	32	18	3
Dogs	16	34	4	32	18
Feral Hogs	35	20	33	27	13
Septic	4	34	1	2	18

Although the highest total potential *E. coli* load is estimated to be from cattle (Table 2.5), all the sub-watersheds with the greatest total potential have urban runoff as the greatest source of potential *E. coli* (Table 2.4). None of the top cattle sub-watersheds are high potential watersheds (Table 2.5). Sheep and goats are the other top agricultural sources in the high potential sub-watersheds. In contrast to cattle, the fourth and fifth high potential sub-watersheds for sheep and goats (Table 2.5) are also high overall potential sub-watersheds (Table 2.4). The top sub-watersheds for urban runoff and dogs (Table 2.5) are also the overall high potential sub-watersheds (Table 2.4). Of the high

potential sub-watersheds for septic failure (Table 2.5), two are also high total potential sub-watersheds (Table 2.4). Therefore it can be observed that where urban runoff is present it is the dominant potential source. Furthermore, although cattle are the overall largest contributor, it is a more diffuse source so the effects will not be concentrated at a single point. The sub-watersheds where cattle are the predominant source do not contribute similar total potential loads.

The sub-watersheds with high estimated potential *E. coli* loads are sporadically spatially placed throughout the watershed (Figure 2.13). However, based on the individual source analysis, groupings of high potential sub-watersheds can be seen. Thus BMPs can be devised appropriately for each source and targeted towards the spatial placement. Agricultural BMPs should be placed in the southern and eastern edges of the watershed, where cattle, horses, and sheep and goats are high contributors. Urban non-point BMPs should be instituted for the sub-watersheds around the Luling, Lockhart and Kyle areas to address *E. coli* from urban runoff, dogs, and septic failure. However, the sub-watersheds of 16 and 18 which include the city of Lockhart are not considered to have high load potential for septic failure, because the municipal sewer system serves many of the households. Best management practices that address feral hogs should be placed in the sub-watersheds on the southeastern and northwestern edges of the watershed, were there is a high degree of riparian corridor.

The comparison of the SELECT results to the actual monitoring can be seen in Figure 2.15. For a runoff event to occur the Uhland station had to have greater than 24.3 mm rainfall and Lockhart and Luling had to have greater than 22.5 and 20.9 mm of rainfall, respectively. There were four sampling events that occurred when there was enough rainfall to produce runoff. One of these samples was taken at the Uhland station, one at Lockhart station, and two samples were taken at the Luling station. When the results of SELECT are compared to the actual monitoring, SELECT overestimates the potential concentration at all four sampling events. This reinforces the known uncertainties of the

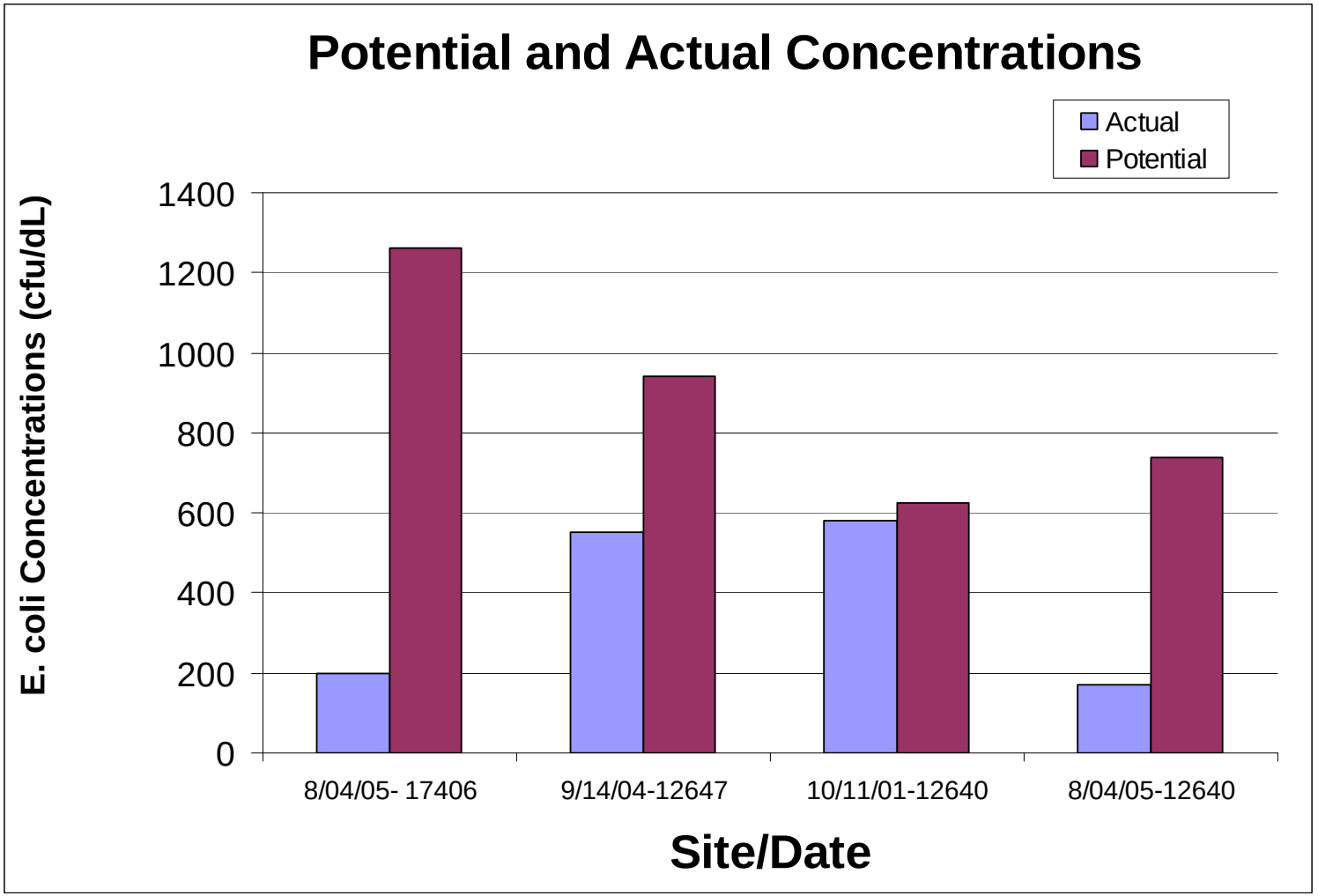

Figure 2.15. Comparison of the Potential *E. coli* Concentration with the Actual Monitored Concentration for Samples Occurring During a Runoff Event.

models, including the exact distribution of the source populations. The overestimation of the concentration is greatest at the Uhland station. The overestimation of SELECT in comparison to the actual is a result of incomplete knowledge of the transport processes. SELECT assumes that all *E. coli* will enter the stream. This does not account for settling, vegetative filtering, temperature or solar inactivation and other biological factors that will reduce the number of viable *E. coli* that will enter the stream. In order to get a more accurate model of the *E. coli* contamination, SELECT should be coupled with a watershed model that models the transport of the *E. coli*.

Another limitation of this study is that the analysis shows only a snapshot of the potential. It does not contribute to the temporal understanding of the *E. coli* survival or movement into the stream. In addition, the pathogen's environmental survival and replication is not modeled. Therefore additional understanding of pathogen fate and transport is required to further model the system with greater accuracy (Santhi et al., 2001).

Current methods such as LDC, bacterial source tracking, and spreadsheet methods lack explicit spatial referencing. The SELECT method fills this gap by estimating the load through spatial methods using ArcGIS. It allows for further watershed modeling using transport processes models that model fate and transport of the pathogen contamination. In addition it provides spatial understanding of the watershed.

2.5 Conclusions

The SELECT methodology estimates the daily average potential *E. coli* production from specified sources within the Plum Creek watershed. It aids in spatially characterizing the watershed. Both the source type and load quantity are characterized through identification of discrete units which have similar potential *E. coli* loads. It contributed to spatial understanding of the most appropriate placement of BMPs for efficient

allocation of resources. This allows for implementation of best management practices (BMPs) that are suitable for individual areas and ultimately results in the increased efficiency of resource allocation. Furthermore, this method helps in the identification of locations which benefit from added or increased monitoring, which in turn aids in the understanding of *E. coli* loads entering the stream. The analysis provides decision making assistance to watershed protection plan development and therefore is an important tool in the TMDL process.

CHAPTER III

STASTISTICAL CLUSTERING OF THE WATERSHED TO SUPPORT WATERSHED PROTECTION PLAN DEVELOPMENT

3.1 Introduction

The cost of a TMDL can range from thousands to over a million dollars per watershed (USEPA, 1996). Models are used as an alternative to intensive monitoring in order to save time, reduce cost and provide forecasting of TMDL implementation impacts (Shirmohammadi et al., 2006). However, the cost of modeling to support TMDL efforts averages 32% of the total costs (USEPA, 1996). This represents a considerable burden to the stakeholders. In order to reduce the cost and effort required to fulfill the goal of TMDL studies appropriate models must be chosen based on the characteristics of the watershed. By understanding influence of watershed characteristics to the contaminant load allocations and grouping discrete areas based on these characteristics, appropriate efforts can be directed towards targeted areas. Thus knowledge of the influencing factors through factor and principal component analysis allows for optimal modeling in future efforts. Furthermore the watershed can be spatially characterized, by cluster analysis, into groups allowing for targeted efforts as determined by the identified important factors. Discriminant analysis then is used to check the results of the cluster analysis so that further refinement of the selected variables can improve cluster analysis.

3.2 Statistical Methods

3.2.1 Factor/Principal Component Analysis

Factor and Principal Component Analysis (FAPCA) is conducted in order to redcuce the number of variables while at the same time retaining the variability of a dataset (Jolliffe,

2002, pp 111-119). It explores the structure of the data in order to classify the relationships between variables. By identifying the correlations between different variables, variables redundant to the end result can be eliminated thus reducing the cost of data analysis. Factors or principal components are derived variables that are uncorrelated and form the best linear approximations of the original variables while producing maximum variance (Helena, 2000).

Factors are found by first finding a matrix of covariance between the original variables. The eigenvalue of the matrix is equal to the variance of the factors of the variables. The sum of the eigenvalues should be equal to the number of variables. The process of extracting factors is also described as variance maximizing rotation of the original variable space (Alberto et al., 2001). When evaluating the factors to retain, there are two tests to determine the number of factors retained. The first, the Kaiser Criterion states that only those with an eigenvalue greater than one should be retained (Thyne et al., 2004). The eigenvalues are plotted in the second method, the Scree test. The neck of the plot or where the value of eigenvalues level off at one, reflects the number of factors to retain (Jackson, 1993). Thus the factors with the highest eigenvalues are retained. Each factor is then a linear combination of the rotated factor score multiplied by the original variable (Carlon et al., 2001). The first factor accounts for a majority of the variation in the original variance.

3.2.2 Cluster Analysis

Cluster analysis is performed using the factors in a K-Means clustering algorithm. The K-means algorithm iteratively computes a cluster center and reassigns the cluster membership based on the shortest Euclidean distance of each member to the cluster center (Soltani and Modarres, 2006). The number of clusters is set *a priori* and the algorithm terminates when the cluster membership no longer changes (Jain et al., 1999). The cluster centers are assigned to maximize the variance between the clusters and the

algorithm is designed to minimize the variance within the cluster. The clusters are evaluated using the pseudo F (PSF) statistic, cubic clustering criterion (CCC), and silhouette width. In each of these statistics a local maximum indicates an appropriate number of statistics (DeGaetano, 1996).

3.2.3 Discriminant Analysis

The effectiveness of the cluster analysis can be evaluated by discriminant analysis (DA). At the same time DA is used to identify the factors that distinguish between the clusters (Paul et al., 2006). The DA process is the stepwise addition of variable with testing of each variable to make sure it meets certain criteria. An F-test is performed at each step in order to test for the statistical significance. The variable with the highest F-value is added to the selected variables (Liao et al., 2006). Then Wilk's lambda is calculated and the variable that contributes the least to the discriminatory power is removed (SAS, 2003). Wilk's lambda is the likelihood ratio criterion that is the fractional amount with cluster variance relative to between cluster variance that remains unaccounted for after each variable selected in DA (Paul et al., 2006). The stepwise process stops when all variables meet the criteria to stop and no other variables meet the criteria to enter the selected set. When all the variables have been either accepted or rejected, then a discriminant function or a linear combination of the accepted variable is produced through linear regression (Liao and Chang, 2005). The linear discriminant function is then used to create a matrix for evaluation of the effectiveness of the cluster analysis. The average squared canonical correlation (ASCC) is the proportion of the variance accounted for by the accepted variables (Rencher, 1992).

The objective of statistical analysis was to identify similar clusters of the sub-watersheds of the Plum Creek watershed, Texas, based on the identification of distinguishing variables. The variables that contribute the greatest to the variability of the dataset are identified and used to identify clusters of sub-watersheds.

3.3 Methodology

3.3.1 Overview of Statistical Analysis

Plum Creek watershed was first divided into 35 sub-watersheds using Soil and Water Assessment Tool (SWAT) analysis of on land use and hydrology, by the Texas A&M Spatial Sciences Laboratory. The sub-watersheds are regions that drain into an ephemeral or perennial stream (Chapter II Figure 2.2).

The sub-watersheds are then characterized by twenty five variables that cover percent land use, average distance to land use, drainage factor, and source populations. The data for each of these variables were normalized to perform the following statistical analysis. Factor analysis was performed on all 25 variables. Factors that were linear combinations of the normalized variables were identified as contributing most to the variability of the data set. Scree test and Kaiser criterion were used to determine the number of factors to retain.

The factors of each sub-watershed were then used in K-means clustering. The K-means clustering algorithm was performed with one to 35 clusters. Then the PSF, CCC, and silhouette width were used to determine the appropriate number of clusters. With the cluster membership of each sub-watershed determined, stepwise DA was performed to check the clustering results. Based on these results, the discriminating variables identified by DA were used to re-perform factor analysis and then cluster analysis.

3.3.2 Characterization of Sub-Watersheds

Variables reflecting the percent land use were calculated using land use classification. The land use classification was performed by the Texas A&M University Spatial

Sciences Laboratory, by digitizing the 2004 NAIP aerial photograph (Chapter II Figure 2.1). This variable type contributed twelve variables to the original dataset. Then a straightline distance was calculated to each type of land use, and an average taken. The results for each land use were averaged for each sub-watershed.

The drainage factor was calculated by dividing the area of each sub-watershed by the length of the stream within the sub-watershed. The area of the sub-watershed was determined from the output of SWAT, and the length of the stream taken from the NHD dataset (USGS, 2002).

The population of each for each sub-watershed was calculated based on SELECT results (Chapter II). The source populations included number of households using sewers, number of failing septics, number of dogs, cattle, sheep and goats, horses, feral hogs, and deer. Data was used from the National Agriculture Statistics Survey (USDA, 2002), U.S. Government Census (USCB, 2000), County Subdivision Data, Texas Parks and Wildlife deer surveys (Lockwood, 2005), and literature estimates for feral hog densities (Hellgren, 1997). The populations were evenly distributed to appropriate land uses.

3.3.3 Normalization of Data

The data set for each variable was tested for normality using the Kolmogorv-Smirnov test (Haan, 2002, pp 213-219). Variables that were not distributed normally were then transformed using a Box-Cox transformation (Box and Cox, 1964), with R statistical software (WU Wien, 2007). Normality was again tested using the Kolmogorv-Smirnov test. Variables that were still not normally distributed, were then transformed with a rank-order transform (Juang et al., 2001).

3.3.4 Factor Analysis

Using Statistical Analysis Software (SAS), factor analysis was performed on the normalized data in order to identify the factors that would affect the load of *E.coli* from a sub-watershed. Both the Kaiser criterion and Scree test were used to determine the number of factors to retain.

3.3.5 Cluster Analysis

Several techniques were explored to cluster the sub-watersheds. K-means was determined to be the most appropriate because this algorithm produced the best clustering results. This method requires that the number of clusters be known *a priori*, so the K-means clustering algorithm was employed for one to 35 clusters and the pseudo F statistic (PSF), CCC, and silhouette width calculated for each algorithm output. Based on the first local maximum of each statistical test, the optimal number of clusters was determined.

3.3.6 Discriminant Analysis

Discriminant analysis was performed to evaluate the clustering done with cluster analysis. The cluster membership of the optimal number of clusters was used with the original normalized data set. Stepwise DA was performed. Discriminating variables were found and the agreement between DA and cluster analysis tested. Based on the percent error between the cluster analysis and DA, it was determined that factor analysis and cluster analysis should be preformed again using only the discriminating variables.

3.3.7 Duncan's Multiple Range Test

With the final cluster membership found by re-performing cluster analysis with the new factors from only the discriminating variables, Duncan's multiple range test evaluated the similarity of the clusters at an $\alpha = 0.05$, in order to determine the clusters that were statistically different from the other clusters in regard to each discriminating variable. Then the means of each variable for each cluster were plotted.

3.4 Results

The characteristics of each sub-watershed were determined through SELECT analysis (Chapter II). After testing each variable for normality, the variables were transformed using either the transform lambda determined by Box-Cox or rank order transform.

Factor analysis was performed on this tranformed dataset using SAS. Then the number of factors to retain was determined using the Scree Test and Kaiser Criterion. As shown in the Scree Plot, Figure 3.1, the neck of the curve is approximately at five factors. This was in agreement with the Kaiser Criterion (Table 3.1). The factors that were retained are shown in Table 3.1. By examining the cumulative eigenvalues, these factors reflect 82% of the variability of the dataset. Each factor is a linear combination of parameters for each variable. The parameters that contribute to each factor are retained if they are greater than 0.6. In Table 3.1 these values are underlined. The first factor has parameters for the percent of low density development, percent of medium density development, average density to residential land use, number of households using septic, and the population of dogs. This first factor reflects low density development. The second factor is a linear combination of the populations of cattle, horses, deer, and feral hogs, thus encompassing animal source populations. The third factor includes the

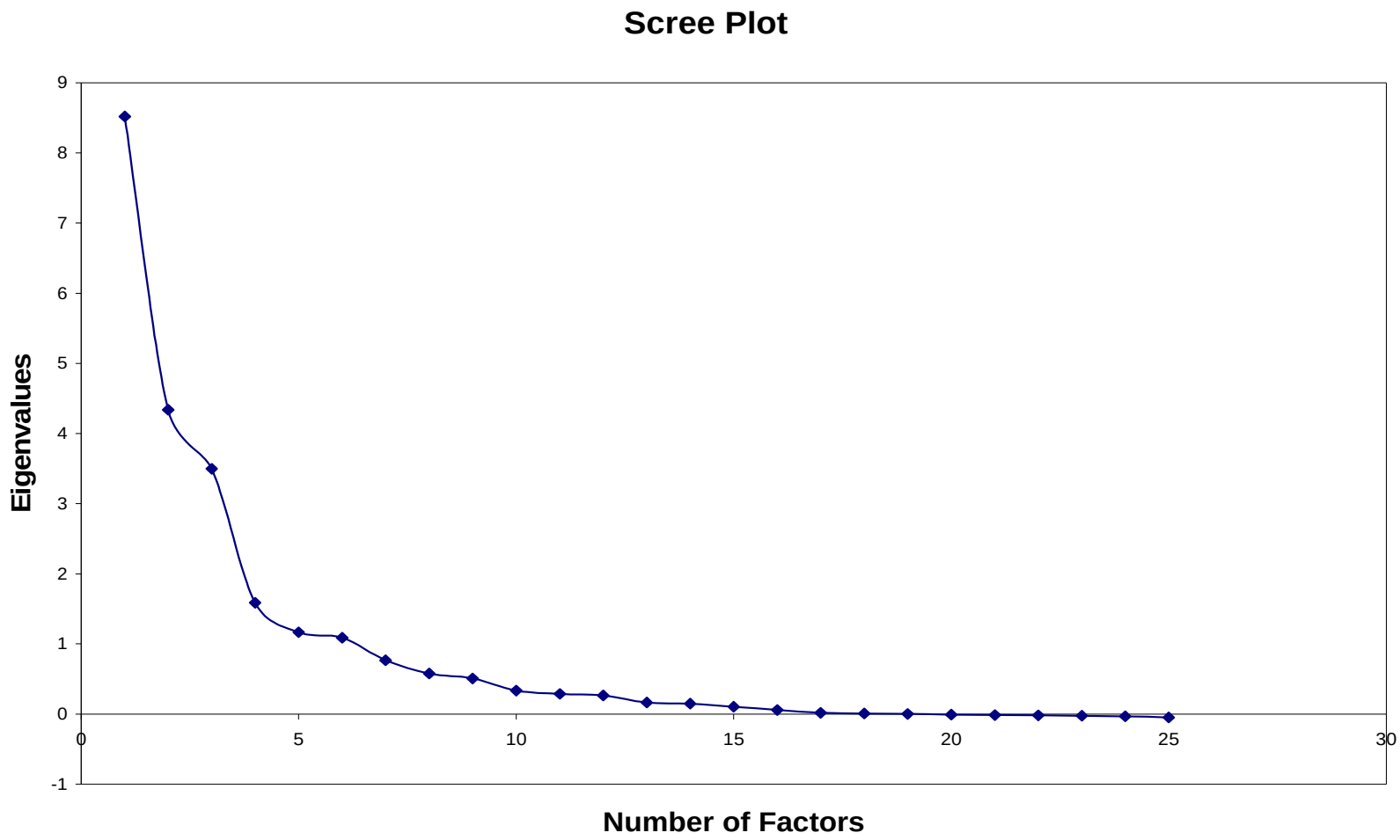

Figure 3.1. Five Factors Retained Based on Scree Plot Test.

percent rangeland and the average distance to pasture. Both of these parameters reflect agricultural land use. The fourth factor includes the percent riparian corridor and the average distance to wetlands. The fifth factor is a linear combination of percent high density development and the number of households using sewer, thus accounting for the variability due to high density development.

Table 3.1. Factors Retained by Factor Analysis.

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Percent Open Developed	-0.191	0.400	-0.160	0.274	0.431
Percent Low Intensity Developed	0.602	-0.228	0.235	-0.270	-0.240
Percent Medium Intensity Developed	0.844	-0.244	0.095	-0.054	0.141
Percent High Intensity Developed	0.276	0.122	-0.113	0.044	0.744
Percent Open Water	0.281	0.124	0.508	-0.185	-0.321
Percent Barren	0.039	0.009	-0.094	0.021	0.309
Percent Forest Land	-0.412	0.231	-0.087	0.213	0.076
Percent Near Riparian Corridor	-0.185	0.388	-0.039	0.800	0.171
Percent Mixed Forest	-0.204	0.242	-0.038	0.228	0.174
Percent Rangeland	0.144	0.208	0.833	-0.033	-0.187
Percent Pasture	-0.366	0.034	-0.133	0.150	0.081
Percent Cultivated Crops	-0.040	-0.467	-0.417	-0.189	0.015
Average Distance to Wetland	-0.045	0.376	0.052	0.766	0.139
Average Distance to Forest	-0.337	0.464	0.066	0.374	0.086
Average Distance to Residential	0.848	-0.130	0.099	-0.087	0.085
Average Distance to Pasture	-0.032	0.224	0.801	0.142	0.132
Drainage Factor	-0.102	0.014	-0.038	-0.040	-0.183
Households using Sewers	0.161	-0.022	-0.026	0.301	0.747
Failing Septic Systems	0.873	0.058	0.138	0.003	-0.014
Cattle	-0.136	0.874	0.342	0.205	-0.021
Sheep and Goats	0.394	0.147	0.565	-0.301	-0.205
Horses	-0.239	0.771	0.052	0.256	0.183
Dogs	0.740	-0.155	-0.123	-0.089	0.232
Deer	-0.146	0.772	0.280	0.161	0.055
Feral Hogs	-0.052	0.898	0.034	0.138	-0.022
Eigenvalues	8.52	4.34	3.49	1.59	1.17
Cumulative Percent of Variance	36.57	55.20	70.21	77.02	82.04

The values for each factor were calculated for each sub-watershed. These factors were then used in cluster analysis. Clustering using the K-means algorithm was performed for *k* number of clusters from one through thirty-five. The appropriate number of clusters was determined by looking for a local maximum in the pseudo F statistic, cubic clustering criterion, and the silhouette width. In each case, a local maximum was found at four clusters (Figure 3.2 and Figure 3.3). The clusters of sub-watersheds identified by cluster analysis are shown in Figure 3.4

The sub-watershed cluster membership was then used with the transformed original dataset in stepwise discriminant analysis. Discriminant analysis identified the discriminating variables shown in Table 3.2. For each step in the process, the variable with a high F-value was retained. The F-value reflects the statistical significance of the variable to the cluster membership. The Wilk's lambda is the unaccounted for intra-cluster variance in relation to the inter-cluster variance (Paul et al., 2006). The average squared canonical correlation (ASCC) is the amount of variation in the dataset that is attributed to the group of selected variables. As seen in Table 3.2, the eight discriminating variables, selected by the DA algorithm, account for 79% of the variability of the dataset. Using the discriminant function or a linear combination of the discriminating variables the results of the cluster analysis was evaluated. The results of the cluster analysis and DA are shown in Table 3.3. The highlighted diagonal elements show where the CA and DA agree (Table 3.3). For cluster one, DA assigned one of the two sub-watersheds to a different cluster. Three of the seven sub-watersheds in cluster two were reassigned. Two of the 17 sub-watersheds from cluster three and three of the nine sub-watersheds from cluster four were reassigned. Overall, nine of the 35 sub-watersheds were reassigned, accounting for 35% error (Table 3.3).

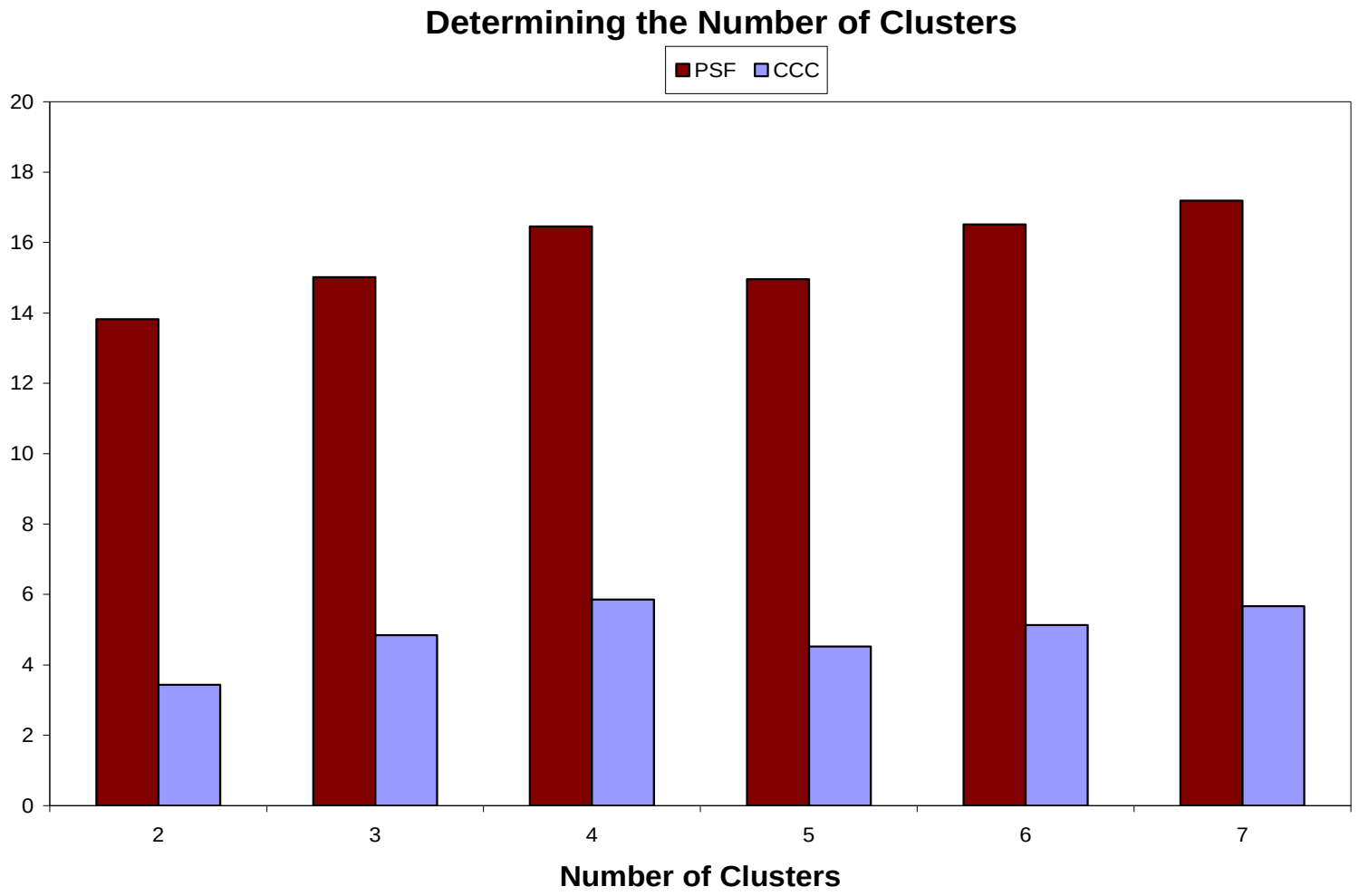

Figure 3.2. Division of Watershed into Four Clusters Based on Pseudo F (PSF) statistic and Cubic Clustering Criterion (CCC).

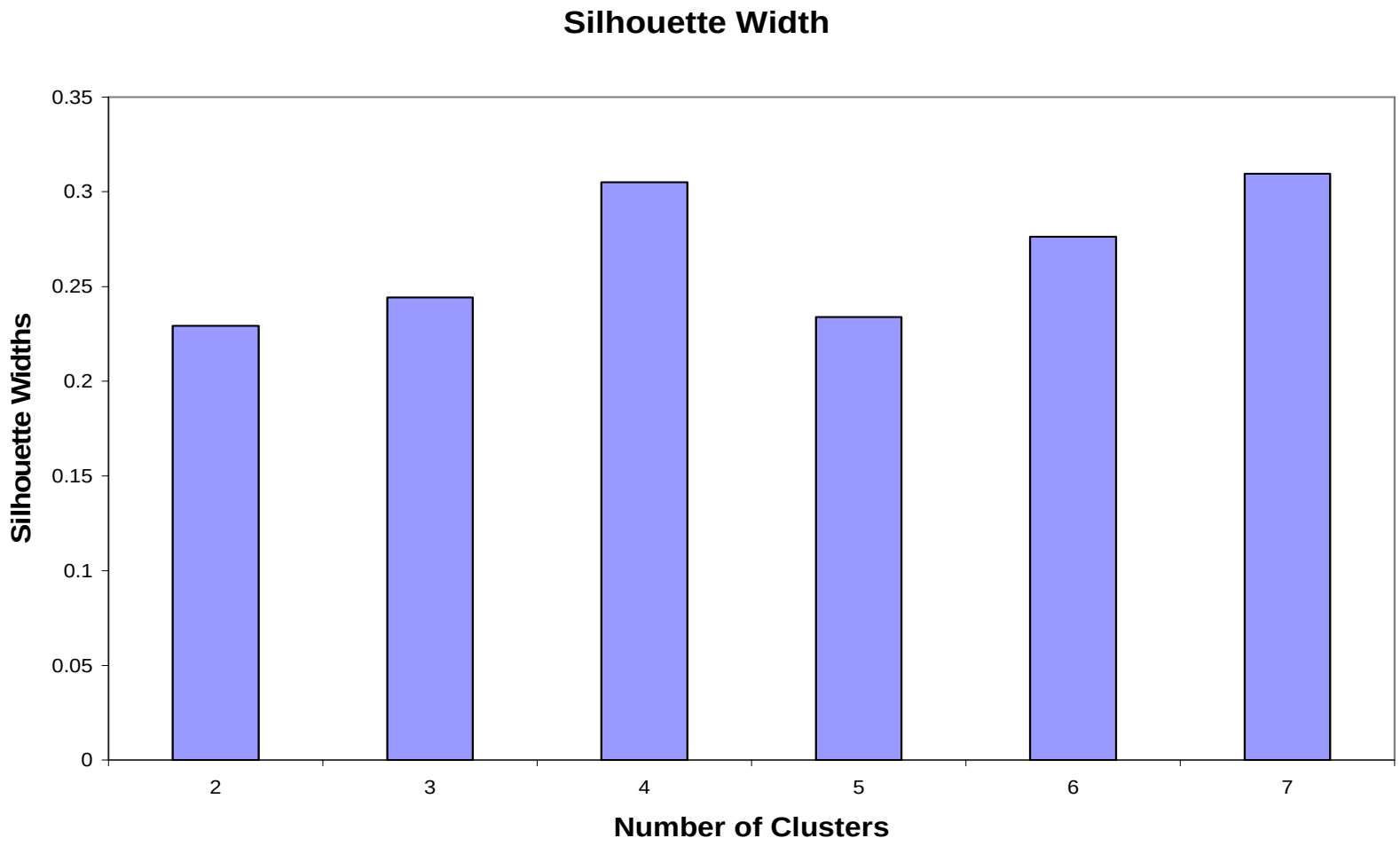

Figure 3.3. Division of Watershed into Four Clusters Based on Silhouette Width.

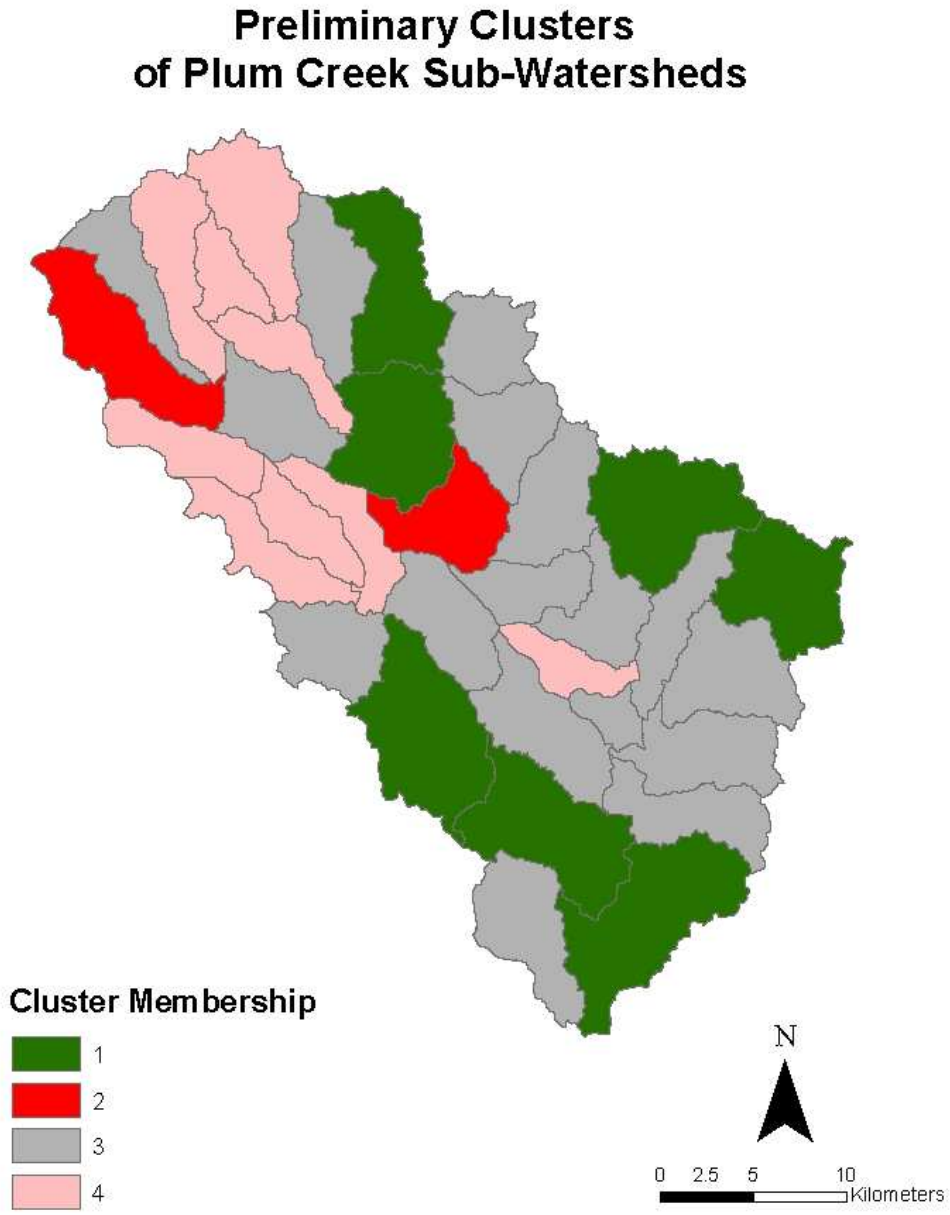

Figure 3.4. Preliminary Clusters of Sub-Watersheds in Plum Creek.

Table 3.2. Discriminating Variables Determined by Discriminant Analysis.

Step	Number Variables	Entered	Removed	Partial R^2	F value	Pr > F	Wilk's Lambda	ASCC
1	1	Number of Cows		0.7987	40.99	<.0001	0.2013	0.2662
2	2	Number of Dogs		0.6972	23.03	<.0001	0.0610	0.4928
3	3	Number of Sewers		0.5812	13.41	<.0001	0.0255	0.66
4	4	Percent of Mixed Forest		0.2383	2.92	0.0514	0.0194	0.7054
5	5	Percent of Open Developed Land		0.378	5.47	0.0045	0.0121	0.7484
6	6	Average Distance to Wetlands		0.2228	2.48	0.0831	0.0094	0.7591
7	7	Percent Cultivated Crops		0.2129	2.25	0.1069	0.0074	0.7714
8	6		Percent of Mixed Forest	0.1562	1.54	0.2281	0.0088	0.755
9	7	Percent High Intensity Development		0.2101	2.22	0.1111	0.0069	0.7615
10	8	Percent Med Intensity Development		0.2002	2	0.1405	0.0055	0.788
11	9	Percent Rangeland		0.2744	2.9	0.0568	0.0040	0.8004
12	8		Percent High Intensity Dev.	0.19	1.8	0.1757	0.0050	0.794

Note. Pr>Wilk's Lambda >0.001 and Pr> ASCC >0.001

Table 3.3. Errors in Cluster Assignment.

Cluster	Quantity	1	2	3	4	Total
1	Number	1	0	0	1	2
	Percentage	50	0	0	50	100
2	Number	1	4	2	0	7
	Percentage	14.29	57.14	28.57	0	100
3	Number	0	1	15	1	17
	Percentage	0	5.88	88.24	5.88	100
4	Number	1	0	2	6	9
	Percentage	11.11	0	22.22	66.67	100
Total	Number	3	5	19	8	35
	Percentage	8.57	14.29	54.29	22.86	100
Priors		0.25	0.25	0.25	0.25	
Error Rate		0.5	0.4286	0.1176	0.3333	0.3449

Based on the error rate of cluster assignment between DA and cluster analysis, the factor analysis was re-performed using the discriminating variables. Following the same procedure (Scree test Figure 3.5) three factors were retained for cluster analysis (Table 3.4). The cluster membership determined by the K-means algorithm is shown in Figure 3.6. Three sub-watersheds were reassigned by re-performing FA and CA. Then Duncan's multiple range test was performed to determined the similarity of the clusters for each discriminating variable. The results of the test are in Table 3.5. Clusters that are grouped together in parenthesis are similar. Clusters that are in a different group are dissimilar. Then each cluster was given a qualitative ranking of high, medium, or low based on the average mean for that variable within each cluster. The average mean of each cluster for each variables are plotted in Figure 3.7 and Figure 3.8.

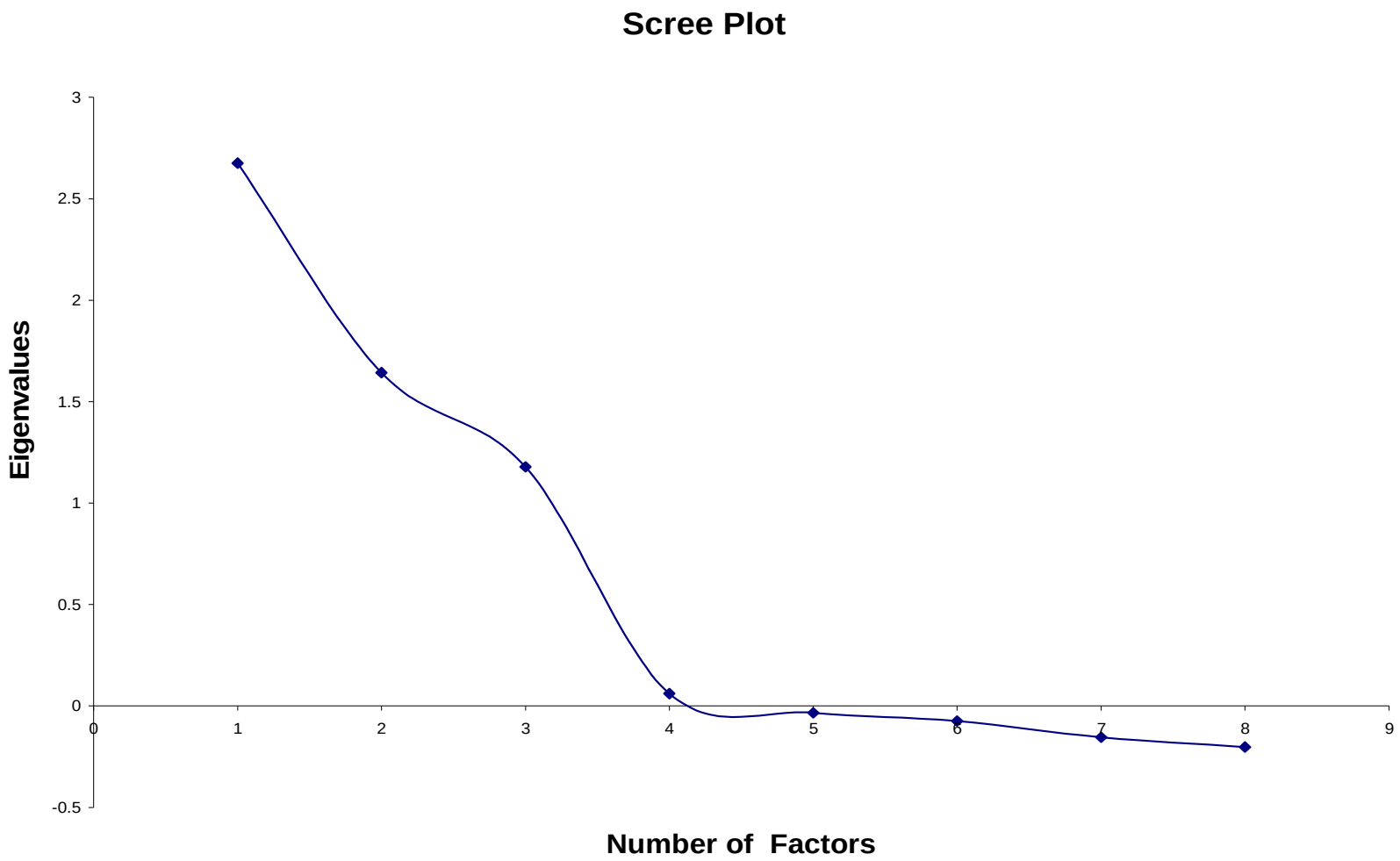

Figure 3.5. Three Factors Are Retained Based on the Scree Plot Test.

Table 3.4. Factors Retained of Discriminating Variables.

Variable	Factor 1	Factor 2	Factor 3
Percent Open Developed	0.895	-0.018	-0.140
Percent Medium Intensity Developed	-0.207	0.024	0.798
Percent Rangeland	-0.244	0.825	0.043
Percent Cultivated Crops	-0.503	-0.671	0.086
Average Distance to Wetland	0.700	0.276	-0.118
Households Using Sewer	0.661	-0.139	0.429
Cattle	0.434	0.648	-0.309
Dogs	0.089	-0.129	0.771
Eigenvalues	2.68	1.64	1.18
Cumulative Percent of Variance	52.53	84.77	107.91

Table 3.5. Cluster Comparison Using Duncan's Multiple Range Test.

Variable/Cluster	Duncan Results	1	2	3	4
Frequency		8	2	13	12
Percent Open Developed Land	(2)(1,3,4)	Low	High	Low	Low
Percent Medium Intensity	(2)(1,3,4)	Low	High	Low	Low
Pecent Rangeland	(1,3,4)(2,3,4)	High	Medium	Medium	Low
Percent Cultivated Crops	(2,3,4)(1,2,3)	Low	High	Medium	High
Average Distance to Wetland	(1,2,3)(2,3,4)	High	Medium	Medium	Low
Numbers of Sewers	(2)(1,3,4)	Low	High	Low	Low
Numbers of Cows	(1)(3,4)(2,4)	High	Low	Medium	Medium
Numbers of Dogs	(2)(1,3,4)	Low	High	Low	Low

3.5 Discussion

The eight discriminating variables identified by DA have an ASCC of 0.82 and thus accounts for 82% of the variability of the original dataset. The four clusters identified by cluster analysis based on the three factors that are a combination of these discriminating variables (Table 3.2).

3.5.1 Cluster One

Cluster one has eight sub-watersheds. These sub-watersheds are on the southwestern and eastern edges of the watershed (Figure 3.6). Cluster one had the greatest mean of percent open developed land, rangeland, cattle, and distance to mixed forest (Figure 3.7). Duncan's multiple range test identified cluster one's cattle population as being significantly different from other clusters' cattle populations (Table 3.5). It would be most effective for best management practices (BMPs) to focus on addressing loads from agriculture, such as cattle.

3.5.2 Cluster Two

Cluster two contains two sub-watersheds, 34 and 16 (Figure 3.6 and Figure 2.2). Both of these sub-watersheds are urban areas encompassing the cities of Kyle and Lockhart. Duncan's multiple range test identified cluster two as being distinctly different from the other clusters, with the characteristics of dogs and percent medium intensity development (Table 3.5). When the discriminating variable cluster means are examined, cluster two has high mean for medium intensity development, dogs, and sewers (Figure 3.8). Therefore BMPs should focus on reducing loads from urban runoff, dogs, and wastewater treatment plant effluent.

3.5.3 Cluster Three

Cluster three contained 13 sub-watersheds, with nine sub-watersheds in the center of the southern portion of the watershed (Figure 3.6). The other four sub-watersheds were separate and isolated with three placed in the northern portion and the fourth at the southern tip of the watershed. Duncan's multiple range test did not identify any variable for which cluster three was distinctive from all the other clusters (Table 3.5). In

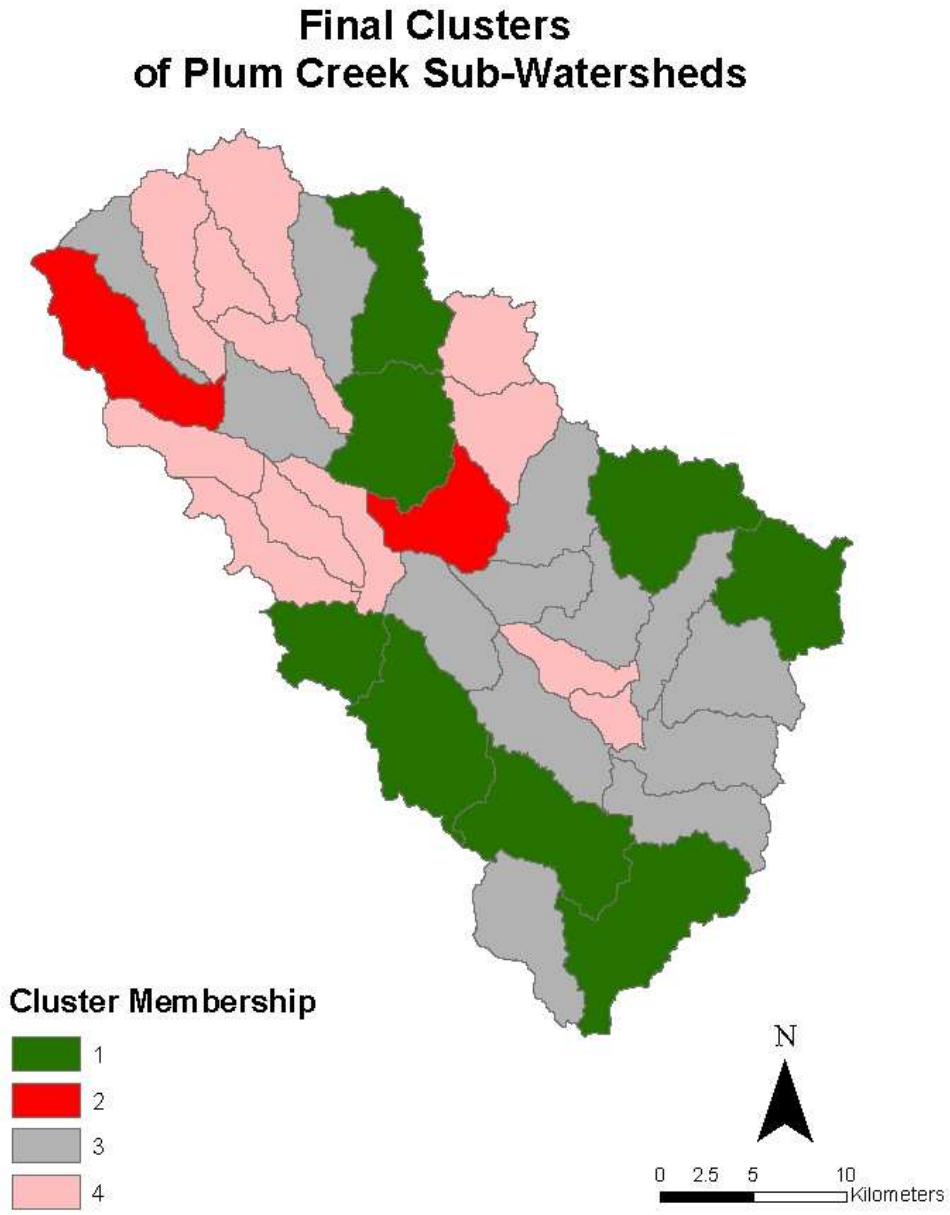

Figure 3.6. Final Clusters of Sub-Watersheds in Plum Creek.

addition, cluster three did not have any variable means that were the greatest or lowest of the four clusters (Figure 3.7 and Figure 3.8). The cluster means and Duncan's multiple range test do not identify any general distinctive characteristics that would assist in decision assistance for identification of BMPs.

3.5.4 Cluster Four

Cluster four has 4 groupings of sub-watersheds (Figure 3.6). Two groups of four sub-watersheds are in the northern portion of the watershed. Two groups of two sub-watersheds are located in the center and the north central edge of the watershed. Duncan's multiple range test only identified the number of households using sewers as a variable that cluster four was significantly different from all other clusters (Table 3.5). It was identified as having low numbers. Cluster four had the highest mean of percent cultivated crops (Figure 3.7). This distinguishing characteristic of cluster four, does not assist in decision making or placement of BMPs.

3.5.5 SELECT Validation

When the sub-watersheds were ranked in descending order by the SELECT output (Chapter II, See Figure 2.13) of average daily potential load and then compared to the cluster membership, the clusters and ranks matched up with the exception of 11 sub-watersheds (Figures 3.6 and 2.13). This means that the statistical analysis and SELECT matched up for 68.6% of the sub-watersheds. Therefore in approximately 69% of the predictions of potential loads from SELECT can be validated by statistical methods.

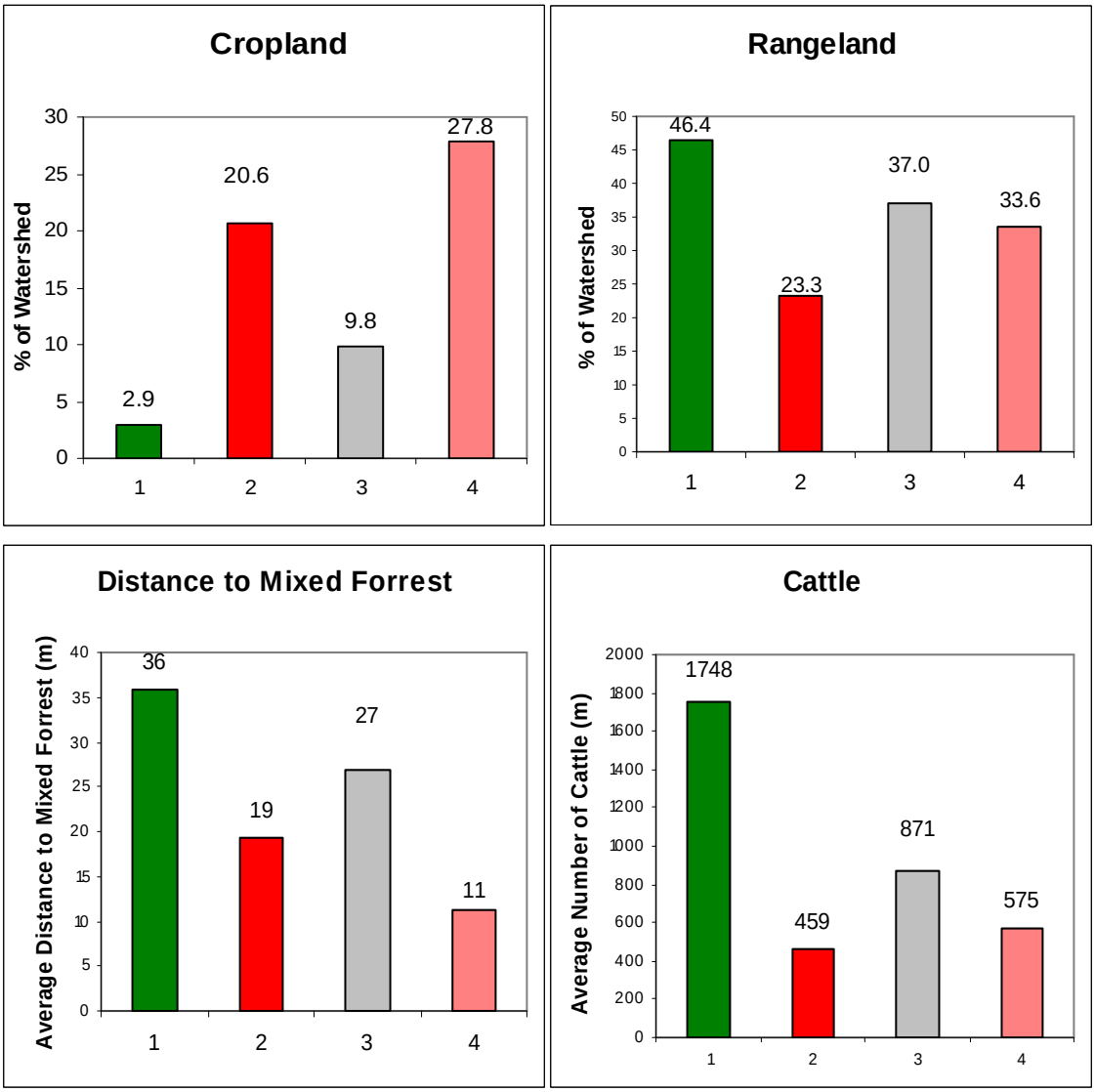

Figure 3.7. Cluster Means of Variables Distinguishing Rural Sub-Watersheds.

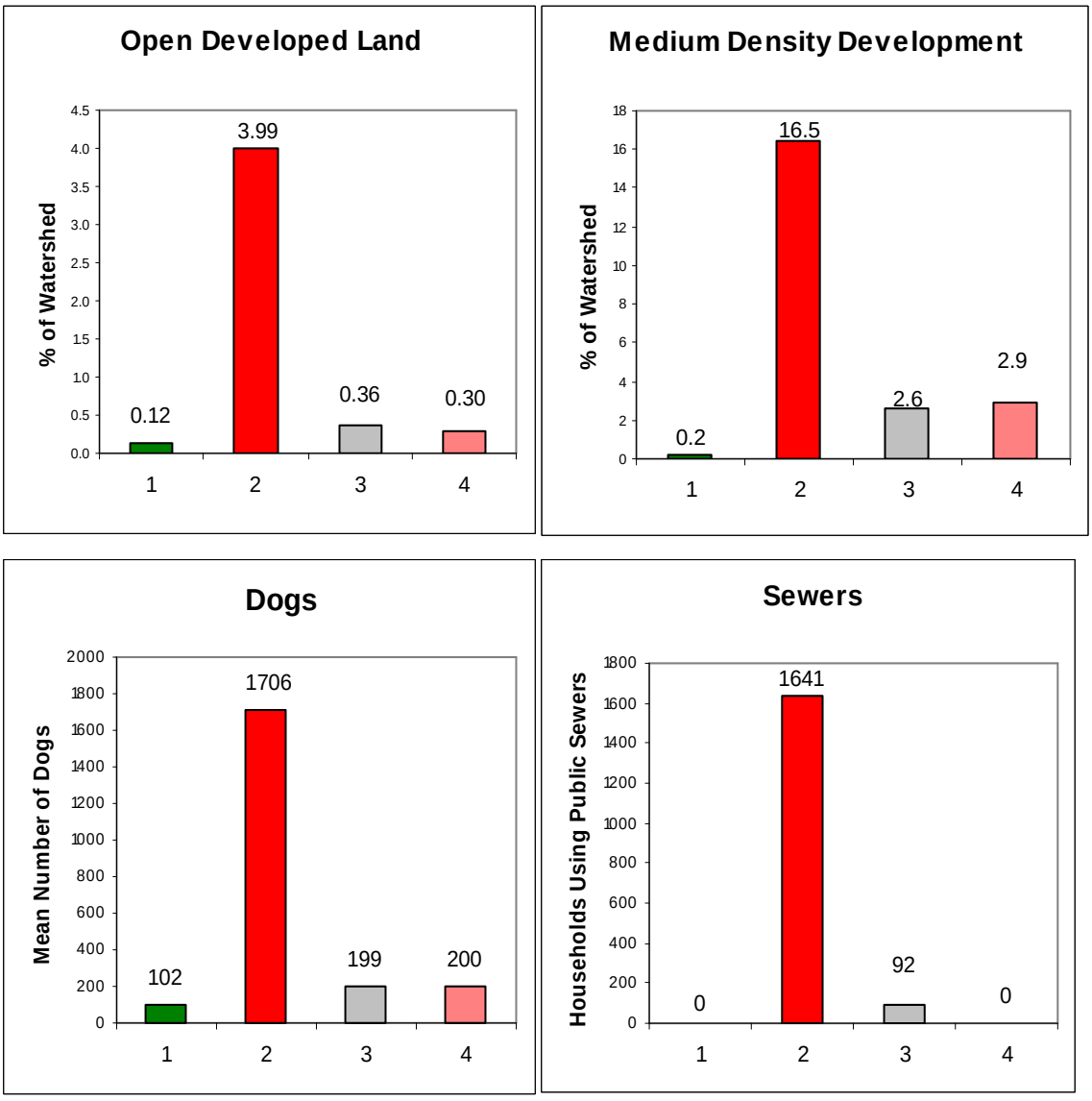

Figure 3.8. Cluster Means of Variables Distinguishing Urban Sub-Watersheds

3.6 Conclusions

Plum Creek was statistically characterized in order to cluster the sub-watersheds into groupings of management areas. Four clusters were identified. One cluster was high density urban, one was high in cultivated crops, another with range and forest lands, and a cluster with no distinguishing characteristics. The discriminating variables that distinguish the sub-watershed were identified. The variables of cattle and dog populations attribute a majority of the variability within the dataset. This information provides important support for selection of BMPs. In addition, it provides direction for future modeling efforts.

The SELECT method provides decision assistance for stakeholders participating in the TMDL process. It serves as an input for watershed models that couple the potential input from SELECT and transport processes. When coupled with statistical cluster analysis, resources for BMPs can be efficiently allocated.

CHAPTER IV

CONCLUSIONS

4.1 Conclusions

- Spatially Explicit Load Enrichment Calculation Tool (SELECT) was developed. This tool was designed to distribute point sources and non-point source populations then calculate the average daily potential *E. coli* load produced from each source.

- SELECT was applied to Plum Creek Watershed in Texas. The results of SELECT were used to support the development of the watershed protection plan. SELECT produced maps of the distribution of non-point sources and the load throughout the watershed. In addition, sub-watershed totals were calculated and the percentage contribution of each source determined.

- The sub-watersheds of Plum Creek watershed were statistically characterized and clustered. This was accomplished through factor analysis, cluster analysis, and discriminant analysis. As a result of these statistical techniques, the sub-watersheds were divided into four clusters. The set of variables used to characterize the sub-watershed was reduced to factors that captured 80% of the variability. Furthermore, variables describing dog and cattle population were found to account for the majority of the variability within the watershed.

4.2 Limitations

There are several limitations of SELECT that restrict the utility of its application. It does not account for fate and transport of the *E. coli* cells or temporal variability. It does not account for cell death, inactivation, or re-growth. The transport mechanisms that would carry the *E. coli* from deposition to the stream are also not considered. The present method assumes that all potential *E. coli* will enter the stream. Therefore SELECT is only applicable for high flows conditions with a runoff event. Other temporal variations that are not considered are the changes in the conditions of the stream that would affect the growth, survival, and transport of the *E. coli* to a monitoring station.

The distribution of non-point sources assumes uniform distribution to appropriate land uses. The unknown variability of the distribution limits the accuracy of the potential predictions. In addition, the population dynamics of the non-point sources are not considered. The seasonal changes in livestock stocking rates, wildlife population, and septic failure mechanisms are a few examples of variability which SELECT does not capture.

4.3 Recommendations

The output of SELECT, information regarding the distribution of the potential *E. coli* throughout the watershed should be coupled with a watershed model in order to account for the transport processes. Future efforts should focus on using the SELECT in a pathogen fate and transport model in an attempt to more accurately model the actual *E. coli* loading to the stream. Furthermore, SELECT should be applied to other watersheds and the output evaluated for its utility and accuracy. Based on these results improvements can be made. Improvements in data acquisition should focus on the

variables that were identified by the statistical analysis as accounting for the greatest percentage of variability, namely cattle and dog population distributions.

Overall, further refinement of SELECT should focus on improved data to increase accuracy and linking SELCT with a transport process model for a more in depth understanding of the physical system. These improvements would increase the efficacy within the TMDL process and the usefulness of the output for stakeholder decision support.

The strength of the combination of SELECT and the cluster analysis is that it is a tool that can guide the stakeholders in determining what further refinement of data is needed, where sampling should be implemented, and how the effectiveness of BMPs can be evaluated. It is a generic tool that can be applied to any watershed by proper selection of contamination sources. Furthermore, SELECT can be modified to also evaluate other water contaminants, such as nutrients, given there is sufficient information concerning application and production rates.

REFERENCES

Ahmed W., R. Neller, and M. Katouli. 2005. Host species-specific metabolic fingerprint database for Entercocci and *Escheria coli* and its application to identify sources of fecal contamination in surface waters. *Applied and Environmental Microbiology* 71(8): 4461-4468.

Alberto W.D., D.M. del Pilar, A.M. Valeria, P.S. Fabiana, H.A. Cecilia, and B.M. de los Angeles. 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. *Water Research* 35(12): 2881-2894.

Anderson, D. and E. Flaig. 1995. Agricultural best management practices and surface water improvement and management. *Water Science and Technology* 31(8): 109-121.

AVMA. 2002. *US Pet Ownership and Demographics Source Book.* Schaumburg, IL., Center for Information Management, American Veterinary Association.

Baron, J., 1982. Effects of feral hogs (*Sus scrofa*) on the vegetation of Horn Island, Mississippi. *American Midland Naturalist* 107(1): 202-205.

Benham, B., C. Baffaut, R. Zeckowski, K. Mankin, Y. Pachepsky, A. Sadeghi, K. Brannan, M. Soupir, and M. Habersack. 2006. Modeling bacteria fate and transport in watershed to support TMDLs. *Transactions of ASABE* 49(4): 987-1002.

Bonta, J., and B. Cleland. 2003. Incorporating natural variability, uncertainty, and risk into water quality evaluations using duration curves. *Journal of American Water Resources Association* 39(6): 1481-1496.

Box, G., and D. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society Series B* 26(2): 211-252.

Braune, M., and A. Wood. 1999. Best management practices applied to urban runoff quantity and quality control. *Water Science and Technology* 39(12): 117-121.

Carlon, C., A. Critto, A. Marcomini, and P. Nathanail. 2001. Risk based characterization of contaminated industrial site using multivariate and geostatistical tools. *Environmental Pollution* 111: 417-427.

Carson, C., B. Shear, M. Ellersieck, and A. Asfaw. 2001. Identification of fecal *Escheria coli* from humans and animals by ribotyping. *Applied and Environmental Microbiology* 67(4): 1503-1507.

Cleland, B. 2002. TMDL development from the "Bottom Up" – Part II: Using duration curves to connect the pieces. In *National TMDL Science and Policy 2002 – WEF Specialty Conference*. 7-14. Washington, DC: America's Clean Water Foundation.

Cleland, B. 2003. TMDL Development from the "Bottom Up" – Part III: Duration curves and wet-weather assessments. Washington, DC: America's Clean Water Foundation. Available at: http://www.tmdls.net/tipstools/docs/TMDLsCleland.pdf. Accessed May 15, 2007.

Danz, N., R. Regal, G. Niemi, V. Brady, T. Hollenhorst, L. Johnston, G. Host, J. Hanowski, C. Johnston, t. Brown, J. Kinston, and J. Kelly. 2005. Environmentally stratified sampling desiign for the development of Great Lakes environmental indicators. *Environmental Monitoring and Assessment* 102: 41-65.

DeGaetano, A. 1996. Delineation of mesoscale climate zones in the northeastern United States using a novel approach to cluster analysis. *American Meteorological Society* 9(August): 1765-1782.

Doyle, M. and M. Erikson. 2006. Closing the door on the Fecal Coliform Assay. *Microbe* 1(4): 162-163.

Fenwick, A. 2006. Waterborne infectious diseases – Could they be consigned to History ?. *Science* 313: 1077-1081

Fraser, R., P. Barten, and D. Pinney. 1998. Predicting stream pathogen loadings from livestock using a geographic information system-based delivery model. *Journal of Environmental Quality* 27(4): 935-945.

GBRA. 2006. Guadalupe River Basin: Basin highlights report – Spring 2006. Seguin, TX: Guadalupe-Blanco River Authority. Available at: http://www.gbra.org/Documents/CRP/BDA/2006BasinHighlightsReport.pdf. Accessed August 22, 2006.

Geldreich, E. 1996. Pathogenic agents in freshwater resources. *Hydrologic Processes*. 10: 315-333.

Godfree, A. and J. Farrell. 2005. Processes for managing pathogens. *Journal of Environmental Quality* 34: 105-113.

Haan, C. 2002. *Statistical Methods in Hydrology*. 2nd ed. Ames, IA: Iowa State Press.

Haan, C. B. Barfield, and J. Hayes. 1994. *Design Hydrology and Sedimentology for Small Catchments*. San Diego: Academic Press.

Helena, B., R. Pardo, M. Vega, E. Barrago, J. Fernandez and L. Fernandez. 2000. Temporal evolution of a groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Research* 34(3) :807-815.

Hellgren, E. 1997. Biology of feral hogs (*Sus scrofa*) in Texas. In *Feral Swine Symposium*. College Station, TX : Texas Cooperative Extension Service.

Jackson, D. 1993. Stopping rules in principal component analysis: A comparison of heuristical and statistical approaches. *Ecology* 74(8): 2204-2214.

Jain, A., M. Muty, and P. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3): 264-323.

Jolliffe, I. 2002. *Principal Component Analysis*. Secaucus, NJ: Springer-Verlag New York, Incorporated.

Juang, K., D. Lee, and T. Ellsworth. 2001 Using rank order geostatistics for spatial interpolation of highly skewed data in a heavy-metal contaminated site. *Journal of Environmental Quality* 30: 894-903.

Kemper, J. 2000. Septic systems for dogs? Nonpoint Source News-Notes 63. Pet Waste: Dealing with a Real Problem in Suburbia. New Jersey Department of Environmental Protection. Available at: http://www.state.nj.us/dep/watershedmgt/pet_waste_fredk.htm. Accessed May 21, 2007.

Lesikar, B. 2005. Chapter 1 Introduction to onsite wastewater treatment systems. *OWTS 101 : Basics of Onsite Wastewater Treatment Systems*. 4. College Station, TX.: Texas Cooperative Cooperative Extension.

Liao, S., and W. Chang. 2005. Interpretation and discrimination of marshy wetlands by soil factors in the Kuan-Tu Natural Park, Taiwan. *Environmental Monitoring and Assessment* 107: 181-202.

Liao, S., W. Lai, J. Chen, and C. Lee. 2006. Water quality during development and apportionment of pollution from rivers in Tapeng Lagoon, Taiwan. *Environmental Monitoring and Assessment*. 122: 81-100.

Lockwood. 2005. White-tailed deer population trends. Federal Aid in Fish and Wildlife Restoration. Project. W-127-R-14. Texas Parks and Wildlife Department. Austin TX.

Mintz, E., F. Reiff, and R. Tauxe. 1995. Safe water treatment and storage in the home: A practical new strategy to prevent waterborne disease. Centers for Disease Control. Atlanta, GA. Available at: http://www/cdc.gov/safewater/publications_pages/1995/mintz_1995.pdf. Accessed November 12, 2007.

NCDC. 2007. Historic Climate Records. New Braunfels, TX.: National Water Service Forecast Office. Available at: http://www.srh.noaa.gov/ewx/html/cli/monthdaily.htm. Accessed May 9, 2007.

Office of Texas State Demographer. 2006. Texas Population Estimates Program. Austin, TX. Texas State Data Center and Office of the State Demographer. Available at: http://txsdc.utsa.edu/tpepp/txpopest.php. Accessed May 15, 2007.

Paul, S., R. Srinivasan, J. Sanabria, P. Haan, S. Muktar, and K. Neimann. 2006. Groupwise modeling study of bacterially impaired watersheds in Texas : Clustering analysis. *Journal of the American Water Resource Association* August: 1017-1031.

Payment, P., and P. Hunter. 2001. Endemic and epidemic infectious intestinal disease and its relationship to drinking water. In *Water Quality: Guidelines, Standards, and Health*, 61-86. L. Fetrall and J. Bartram ed. London: IWA Publishing:

PBS&J. 2000. Final Report Predicting Effects of Urban Development on Water Quality in the Cities of New Braunfels, San Marcos, Seguin and Victoria. Chapter 2. 16. Austin, TX: Guadalupe-Blanco River Authority and Texas Natural Resource Conservation Commission.

Petersen, T., H. Rifai, M. Suarex, and R. Stein. 2005. Bacterial loads from point and nonpoint sources in an urban watershed. *Journal of Environmental Engineering* 131(10): 1414-1425.

Reed, Stowe, & Yanke LLC. 2001. Study to Determine the Magnitude of, and Reasons for Chronically Malfunctioning On-Site Sewage Facility Systems in Texas, pp vi and x.. Austin, TX: Texas On-Site Wastewater Treatment Research Council.

Rencher, A. 1992. Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician* 46(3): 217-225.

Santhi, C., J. Arnold, J. Williams, L. Huack, and W. Dugas. 2001. Application of a watershed model to evaluate the management effects on point and nonpoint source pollution. *Transactions of ASAE* 44(6) 1559-1570.

SAS. 2003. *SAS User's Guide: Statistics*. Ver. 8. Cary, NC.: SAS Institute, Inc.

Schueler. 1999. Microbes in urban watershed. *Watershed Protection Techniques* 3(1): 551-600.

Shirmohammadi, A., I. Chaubey, R. Marmel, D, Bosch, R, Munoz-Carpena, C. Dharmasri, A. Sexton, M. Arabi, M. Wolfe, J. Frankenberger, C. Graff, and T. Sohrabi. 2006. Uncertainty in TMDL models. *Transactions of ASABE* 49(4): 1033-1049.

Soltani, S., and R. Modarres. 2006. Classification of spatio-temporal pattern of rainfall in Iran using hierarchical and divisive cluster analysis. *Journal of Spatial Hydrology* 6(2): 1-12.

Steets, B., and P. Holden. 2003. A mechanistic model of runoff-associated fecal coliform fate and transport through a coastal lagoon. *Water Research* 37: 589–608.

TCEQ. 1997. Chapter 307 – Texas Surface Water Quality Standards. Austin, TX: TCEQ. Available at: http://www.tceq.state.tx.us/assets/public/permitting/waterquality/attachments/standards/97stand307.pdf. Accessed May 21, 2007.

TCEQ. 2002. 2002 Texas Water Quality Inventory. Austin, TX: TCEQ. Available at: http://pcwp.tamu.edu/docs/02_1810_data.pdf. Accessed Nov 12, 2006.

TCEQ. 2004. Guidance for Assessing Texas Surface and Finished Drinking Water Quality Data, 2004. Austin, TX: TCEQ Office of Compliance and Enforcement Monitoring Operations Division Surface Water Quality Monitoring Program. Available at: http://www.tceq.state.tx.us/assets/public/compliance/monops/water/04twqi/04_guidance.pdf. Accessed January 25, 2007.

TCEQ. 2005. Overview of Surface Water Quality in Texas : 2004 Water Quality Inventory and 303(d) List. Austin, TX: TCEQ. Available at http://www.tceq.state.tx.us/assets/public/compliance/monops/water/04twqi/04_overview.pdf. Accessed January 25, 2007.

TCEQ. 2007. Draft 2006 303(d) List. Austin, TX: TCEQ. Available at: http://www.tceq.state.tx.us/compliance/monitoring/water/quality/data/06twqi/twqi06.html. Accessed May 7, 2007.

Thyne, G., C. Guler, and E. Poeter. 2004. Sequential analysis of hydrochemical data for watershed characterization. *Ground Water* 42(5): 711-723.

Tian, Y., P. Gong, J. Radke, and J. Scarborough. 2002. Spatial and temporal modeling of microbial contaminants on grazing farmlands. *Journal of Environmental Quality* 31:860-869.

UN. 2005. World Population Prospectus: The 2004 Revision. United Nations Population Division, New York. Available at: http://www.un.org/esa/population/publications/WPP2004/World_Population_2004_chart.pdf. Acessed on November, 15, 2006.

USCB. 2000. Census 2000 TIGER/Line® Files. Washington, DC: US Census Bureau. Available at: http://www.census.gov/geo/www/tiger/index.html. Accessed on December, 14, 2006.

USDA-NASS. 2002. 2002 Census of Agriculture-County Data. pp 560-634, 716-718, 719-729, 730-732, 733-734.Washington, DC: USDA National Agricultural Statistics Survey.

USEPA. 1986. Ambient water quality criteria for bacteria. EPA440/5-84-002. pp. 15-16.Washington, DC: USEPA Office of Water Regulations and Standards.

USEPA. 1991. Guidance for water quality-based decisions: The TMDL process. EPA440/4-91-001. pp 3-1–3-17. Washington, DC: USEPA Office of Water.

USEPA. 1996. TMDL development cost estimated : Case studies of 14 TMDLs. EPA841-R/96/001. pp. I-6–I-17. Washington, DC: USEPA Office of Water.

USEPA. 1999. Draft guidance for water quality based decisions: The TMDL process. 2nd ed. EPA841-D-99-001. pp. 3-1-3-18. Washington, DC: USEPA Office of Water.

USEPA. 2001. Protocol for developing pathogen TMDLs: source assessment. 1st ed. EPA841-R-00-002. pp 5-1-5-18. Washington, DC:USEPA Office of Water.

USEPA. 2005. Handbook for developing watershed plans to restore and protect our waters: Chapter 2. Overview of watershed planning process. EPA841-B-05-005. pp 2-3 -2-9. Washington, DC: USEPA Office of Water Non-Point Source Control Branch.

USEPA. 2006a. Drinking water pathogens and their indicators: A reference resource. USEPA Office of Water, Washington, DC, Available at: http://www.epa.gov/enviro/html/icr/gloss_path.html. Accessed on: November 17, 2006.

USEPA.2006b. National section 303(d) list fact sheet. USEPA Office of Water, Washington DC, Available at http://iaspub.epa.gov/waters/national_rept.control. Accessed on: November 17, 2006.

USGS. 2002. National Hydrography Dataset. USGS, Reston, VA. Available at http://nhd.usgs.gov/index.html. Accessed on August 4, 2006.

Weiskel, P., B. Howes, and G. Heufelder. 1996. Coliform contamination and transport pathways. *Environmental Science and Technology* 30(6): 1872-1881.

WU Wien (Vienna University of Economics and Business Administration). 2007. The R Project for Statistical Computing. WU Wien Department of Statistics and Math, Vienna, Austria. Available at: http://www.r-project.org/. Acessed on: February 19, 2007.

Zeckoski, R., B. Benham, S. Shah, M. Wolfe, K. Brannan, M. Al-Smadi, T. Dillaha, S. Mostaghimi, and D. Heatwole. 2005. BLSC: A tool for bacteria source characterization for watershed management. *Applied Engineering in Agriculture* 21(5): 879-889.

Zhang, X., and M. Lulla. 2006. Evaluation of pathogenic indicator bacteria in structural best management practices. *Journal of Environmental Science and Health Part A* 41: 2483-2493.

VITA

Name:	Aarin Elizabeth Teague
Address:	2117 TAMU College Station, TX 77840
E-mail Address:	AesTeague@tamu.edu
Education:	B.S., Biological Systems Engineering, Texas A&M University, 2005 M.S., Biological and Agricultural Engineering, Texas A&M University, 2007