Spatially Explicit Load Enrichment Calculation Tool to Identify Potential *E. coli* Sources in Watersheds

A. Teague, R. Karthikeyan, M. Babbar-Sebens, R. Srinivasan, R. A. Persyn

Abstract. *In 2006, bacterial pathogens were the leading cause of water quality concerns in the U.S. With more than 300 water bodies in the state of Texas failing to meet water quality standards because of bacteria, managing bacteria pollution commanded the attention of regulatory agencies, researchers, and stakeholders across Texas. In order to assess, monitor, and manage water quality, it was necessary to characterize the sources of pathogens within the watershed. The objective of this study was to develop a spatially explicit method to estimate potential* E. coli *loads in Plum Creek watershed in east central Texas. 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 quantified by applying source specific bacterial production rates, and ranking of each contributing source was assessed for the entire watershed. Cluster and discriminant analyses were used to identify similar regions within the watershed for selecting appropriate best management practices. Based on the statistical analysis and the spatially explicit method, four clusters of subwatersheds were found and characterized. The analysis provided a basis for development of spatially explicit identification of best management practices (BMPs) to be applied within the Watershed Protection Plan (WPP).*

Keywords. *Cluster analysis, Fecal bacteria, Impaired streams, Spatial statistics, TMDL.*

The Clean Water Act authorized the U.S. Environmental Protection Agency (USEPA) to set water quality standards. To ensure compliance with the standards set by the EPA, the Total Maximum Daily Load (TMDL) process was developed. The TMDL process 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 water quality standards (USEPA, 1999). If a stream segment did not support its designated use it was listed as impaired on the 303(d) list. In Texas, 41.7% of the stream segments listed on the 303(d) list were impaired due to pathogens (TCEQ, 2006). *Escherichia coli* (*E. coli*) was used as the indicator organism for pathogens from fecal contamination (USEPA, 1986). The Texas Commission on Environmental Quality (TCEQ) set an *E. coli* limit of a geometric mean of 126 cfu dL^{-1} or a single grab sample of 394 cfu dL^{-1} (TCEQ, 2004). For the TMDL process addressing pathogen contamination, the EPA published recommendations to assess *E. coli* source contribution and identification, characterize the sources, and estimate the *E. coli* load produced by each source (USEPA, 2001). The EPA document recommended identifying the location and densities of *E. coli* contributing source populations to characterize the loads in a watershed.

The EPA recommended characterizing non-point sources by multiplying an individual species' excretion rate by the corresponding species' population (USEPA, 2001). The total estimated bacterial pollution is then calculated by combining estimated non-point and 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 subwatersheds (Zeckoski et al., 2005). Direct stream monitoring methods such as ribotyping use genetic testing to find the sources of 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). Genetic fingerprinting and the load duration curve method do not spatially reference the sources, and thus their application within the Watershed Protection Plan (WPP) is limited because they do not provide information regarding the optimal placement of BMPs. The cost of a TMDL has ranged from thousands to over a million dollars per watershed (USEPA, 1996). Models have been 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 has

Submitted for review in October 2007 as manuscript number SW 7244; approved for publication by the Soil & Water Division of ASABE in July 2009.

The authors are **Aarin Teague,** former Graduate Research Assistant, **Raghupathy Karthikeyan, ASABE Member Engineer,** Assistant Professor, and **Meghna Babbar-Sebens, ASABE Member Engineer,** former Post-Doctoral Research Associate, Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas; **Raghavan Srinivasan, ASABE Member Engineer,** Professor and Director, Spatial Sciences Laboratory, Texas A&M University, College Station, Texas; and **Russell A. Persyn, ASABE Member Engineer,** Engineer III, San Antonio River Authority, San Antonio, Texas. **Corresponding author:** R. Karthikeyan, 2117 TAMU, Texas A&M University, College Station, TX; phone: 979-845-7951; fax: 979-845-3932; e-mail: karthi@tamu.edu.

ISSN 0001-2351

averaged 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 implementation, appropriate models must be chosen based on the characteristics of the watershed. By understanding the influence of watershed characteristics on the contaminant load allocations and grouping discrete areas based on these characteristics, appropriate management efforts can be directed towards targeted areas.

The major objective of this study was to develop a Spatially Explicit Load Enrichment Calculation Tool (SELECT) for the characterization of *E. coli* sources and to apply this standalone tool to Plum Creek watershed in Texas for the WPP development process. The secondary objective of this research was to identify similar clusters of subwatersheds of the Plum Creek watershed based on the identification of distinguishing variables with the most significant contribution to bacterial loads. Knowledge of the influencing factors through factor and principal component analysis would allow for optimal watershed modeling. Furthermore, the watershed can be spatially characterized by cluster analysis into groups allowing for implementing BMPs. Discriminant analysis then was used to check the results of the cluster analysis to further refine the selected variables.

Study Area: 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 (fig. 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 29° 38′ 33.94″ N to 30° 5′ 20.11″ N and in longitude from 97° 54′ 36.29″ W to 97° 27′ 13.60″ W. Within the watershed are several rapidly growing towns, including Lockhart, Kyle, and Luling. As of 2006, the populations of Kyle, Lockhart, and Luling were 19,335, 12,978, and 5,704, respectively (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 (fig. 2). The watershed is 38% rangeland, 17% pasture, 11% cultivated cropland, 18% forest, 8% developed land, 6% near riparian forest, and 2% open water and barren land. The landscape is characterized as rolling hills of pasture and cropland surrounded by scrub oak forest (GBRA, 2006).

Methodology

Spatially Explicit Methodology

The SELECT methodology was developed using ArcGIS 9.2 with the Spatial Analyst extension available from ESRI. This spatially explicit method divided 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 were estimated from the sources that were 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. Delineating subwatersheds within Plum Creek using the Soil and Water Assessment Tool (SWAT) model resulted in 35 subwatersheds (fig. 1) (SWAT, 2005). Table 1 lists the spatial database files and formats used as SELECT input.

The SELECT method identified point and non-point sources throughout the watershed. The identified point sources are active wastewater treatment plants, and non-point sources include livestock, dogs (Schueler 1999),

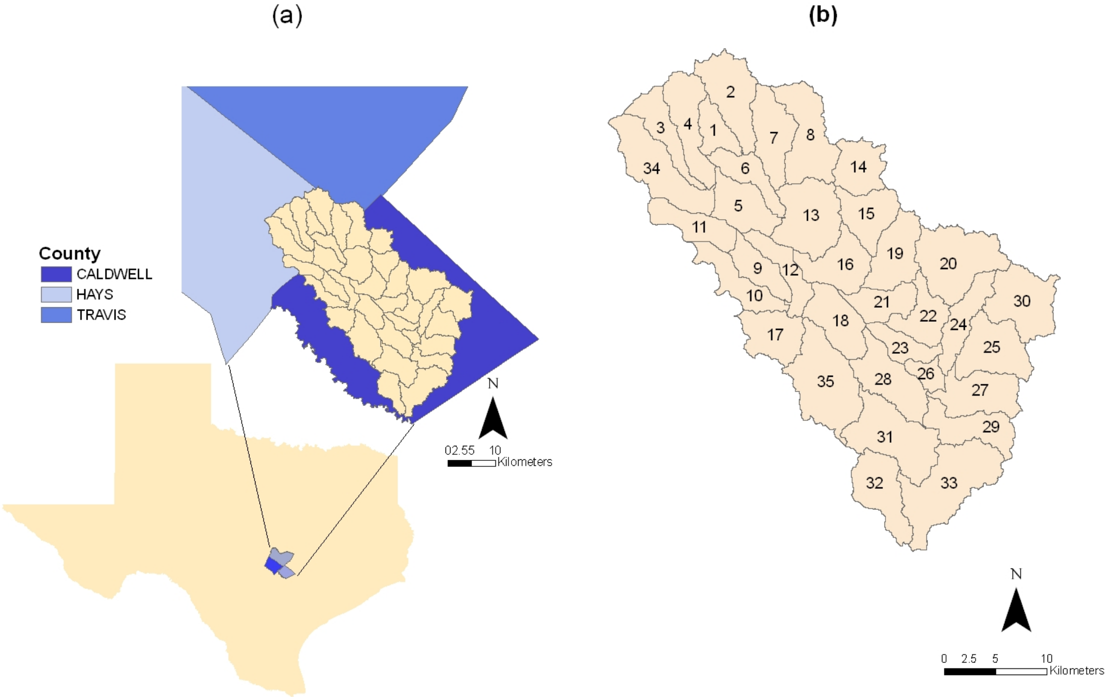

Figure 1. (a) Location of Plum Creek watershed with (b) subwatersheds delineated using SWAT.

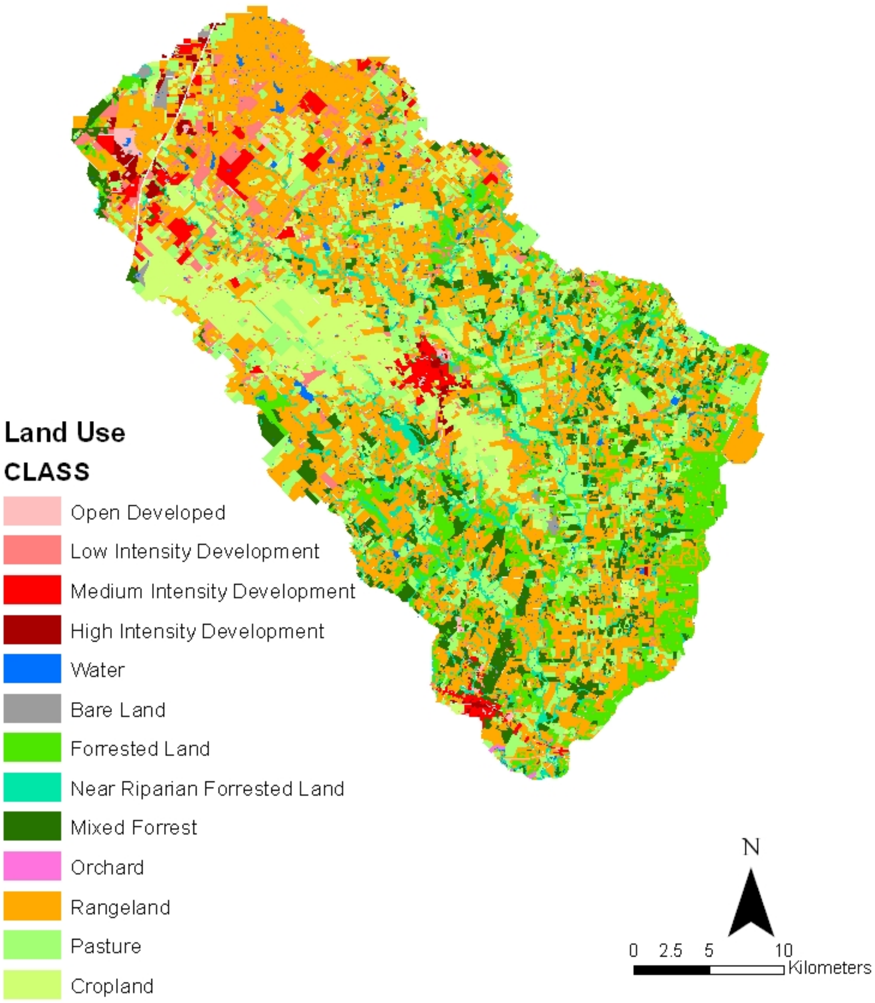

Figure 2. Land use classification of Plum Creek watershed (source: Texas A&M University Spatial Sciences Laboratory, 2006).

wildlife (Weiskel et al., 1996), and failing on-site wastewater treatment systems (OWTS) (Reed, Stowe, and Yanke LLC, 2001). Wildlife sources included many types of wild animals and birds. In this study area, the known wildlife included feral hogs, whitetail deer, raccoons, rodents, opossums, and migratory birds. Feral hogs and deer were the only wildlife sources included within SELECT because they were the only populations of concern with available data. Livestock production within the study area was primarily cattle, horses, sheep, and goats. Generally, dogs were the primary pets allowed to defecate outside the home, and most often the defecated waste was not cleaned up. Cats and other pets were primarily kept in homes, and their waste was disposed of directly to solid waste management, so these contributions were neglected.

Table 1. Data sources and format used in SELECT to predict potential *E. coli* load in Plum Creek watershed.

Source	Spatial Data File	Format	Data Source[a]
Livestock	Counties Ag inventory	Shapefile Tabular	USDA-NASS USDA-NASS
Wildlife	Suitable habitat Urban areas Streams Wildlife inventory	Shapefile Shapefile Shapefile Shapefile	Local wildlife census TIGER census NHD Plus Texas parks and wildlife
OWTS	Subdivisions Census blocks Demographics	Shapefile Shapefile Tabular	Appraisal district TIGER census TIGER census
Pets	Census blocks Demographics	Shapefile Tabular	TIGER census TIGER census
WWTP	Outfall locations Permitted discharge	Shapefile Field in shapefile	State regulatory agency EPA Envirofacts warehouse

[a] TIGER = Topologically Integrated Geographic Encoding and Referencing system; NHD = National Hydrography Dataset (USGS, 2002).

Potential *E. coli* Load Estimation

Each *E. coli* source was first distributed to the appropriate locations within the watershed, and then the load was calculated. The average daily potential load was calculated according to EPA 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% conversion is a rule of thumb that estimates that 50% of fecal coliform (FC) are *E. coli* (Doyle and Erikson, 2006).

Point Sources

Wastewater Treatment Plants

Wastewater treatment plants (WWTPs) were point sources permitted to discharge treated effluent into Plum Creek. There were 13 permitted WWTPs in the watershed, but only five release effluent into the streams. Each WWTP was permitted to release effluent at the water quality standard of 126 cfu dL^{-1}. The load from each WWTP was calculated by multiplying the permitted concentration by the permitted effluent outflow.

Non-Point Sources

Livestock

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 were primarily kept in pasture and on rangeland. Horses were principally confined to pasture areas. Livestock populations within city limits were not included in this study. Watershed areas that were classified as pasture and rangelands were selected from digitized land use data, and areas within the city limits were eliminated. The animal populations obtained from the 2002 Census of Agriculture were aggregated per county (USDA-NASS, 2002). These data were uniformly distributed across the non-urban rangeland and pasture of each county. Based on this distribution, a density of animals per 900 m^2 was calculated. The non-urban rangeland and pasture 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). Then *E. coli* loads were aggregated to the subwatershed level.

Pets

Dog waste was 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 this number by the number of households in Texas, the average number of dogs per household was found to be 0.8. This average was multiplied by the number of households in each census block as published by the U.S. Census Bureau. This provided an estimated number of dogs per census block. Because the census blocks overlap multiple subwatersheds, the density of dogs per area should be calculated in order to account for the spatial variability of high to low density areas. Using the area of each census block, a density of dogs per 900 m^2 was 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 was 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. The potential *E. coli* load contribution from dogs was aggregated for each subwatershed.

Table 2. Calculation of potential *E. coli* loads from various sources in the watershed.

Source	Calculation
Cattle	$EC = \#\text{cattle} \cdot 2.7 \cdot 10^9 \text{cfu d}^{-1} \text{head}^{-1}$
Horses	$EC = \#\text{horses} \cdot 2.1 \cdot 10^8 \text{cfu d}^{-1} \text{head}^{-1}$
Sheep and goats	$EC = \#\text{sheep} \cdot 9 \cdot 10^9 \text{cfu d}^{-1} \text{head}^{-1}$
Deer	$EC = \#\text{deer} \cdot 1.75 \cdot 10^8 \text{cfu d}^{-1} \text{head}^{-1}$
Feral hogs	$EC = \#\text{hogs} \cdot 4.45 \cdot 10^9 \text{cfu d}^{-1} \text{head}^{-1}$
Dogs	$EC = \#\text{households} \cdot \frac{0.8 \text{ dogs}}{\text{household}} \cdot 2.5 \cdot 10^9 \text{cfu d}^{-1} \text{head}^{-1}$
Failing septic systems	$EC = \#\text{ failing systems} \cdot \frac{5 \cdot 10^5 \text{cfu}}{100 \text{ mL}} \cdot \frac{2.65 \cdot 10^5 \text{mL}}{\text{person/day}} \cdot \frac{\text{Avg \# persons}}{\text{household}}$
WWTP	$EC = \text{permitted MGD} \cdot \frac{126 \text{ cfu}}{100 \text{ mL}} \cdot \frac{10^6 \text{gal}}{\text{MGD}} \cdot \frac{3758.2 \text{ mL}}{\text{gal}}$

Wildlife

Wildlife also contributed to the *E. coli* within Plum Creek watershed. Within the watershed, data were available only for two major wildlife contributors: deer and feral hogs. Deer habitat included shrubland and forest areas. Feral hogs primarily used riparian corridors of undeveloped land. To distribute the deer population within Plum Creek watershed, appropriate land use areas with a contiguous 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). 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 was calculated. A fecal coliform excretion rate of 3.5×10^8 cfu day^{-1} $animal^{-1}$ (Zeckoski et al., 2005) was multiplied by the deer per unit area in order to find the *E. coli* load (see the equation in table 2). Then the potential *E. coli* load was aggregated to the subwatershed level.

Feral hog population densities and distribution data were scarce for Plum Creek watershed. Estimates of feral hog densities for the Rio Grande Plains and lower coastal prairie of Texas range from 3.2 to 6 hogs km^{-2} (Hellgren, 1997). Feral hogs utilize nearly all types of landscape, but primarily use forested and shrublands adjacent to river bottomlands. Plum Creek habitat was comparable to the landscape of the Rio Grande Plains and lower coastal prairies. A landscape wide density of 5 hogs km^{-2} was 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 land within 100 m of a stream. 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 was calculated according to the equation found in table 2, where 4.45×10^9 cfu animal^{-1} day^{-1} was the fecal coliform excretion rate multiplied by the 50% rule of thumb. Then the distributed *E. coli* load was aggregated to the subwatershed level.

On-Site Wastewater Treatment Systems

On-site wastewater treatment systems (OWTSs) could contribute pathogens to a water body due to system failure and surface or subsurface malfunction (USEPA, 2001). According to stakeholder input, there were a number of older failing systems within the study area. However, there were 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 have a failure rate of 12%, and unregulated systems have a 50% failure rate (Reed, Stowe, and Yanke LLC, 2001). On-site wastewater treatment systems were regulated starting in 1989, while systems installed prior to 1989 remained unregulated (Lesikar, 2005).

First, the number of households that utilize OWTSs was 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 were obtained from Caldwell and Hays counties. The number of houses both inside and outside of a subdivision was 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 exceeded 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 265 L person^{-1} day^{-1} (70 gal person^{-1} day^{-1}) discharge and a 5×10^6 cfu dL^{-1} concentration in this discharge, the *E. coli* load was calculated according to the equation in table 2 and units were appropriately converted. The average number per household was the average number of people in each household as reported by the 2000 U.S. Census (USCB, 2000). Then potential *E. coli* load was aggregated for each subwatershed.

Statistical Clustering of Subwatersheds

Using SELECT methodology, total potential *E. coli* load resulting from point and non-point sources and the potential *E. coli* load resulting from each source could be estimated. In reality, not all potential *E. coli* will eventually reach the stream. The actual *E. coli* amount in streams will depend on various fate and transport processes in the watershed. To estimate the actual *E. coli* concentrations in streams, a process-based model simulating watershed processes should be incorporated using the potential *E. coli* outputs generated by SELECT as inputs to the model.

Here, statistical clustering techniques were implemented to weigh the influence of the populations and sources of *E. coli* contamination using total potential load and watershed characteristics. The statistical clustering identified areas within the watershed vulnerable to contributing *E. coli* to waterbodies. The 35 subwatersheds were grouped into "clusters" with statistically similar characteristics to recommend and implement BMPs effectively. This process provides stakeholders and decision makers with useful information for implementing mitigation efforts in areas of greatest concern for *E. coli* impairment.

In our analysis, each subwatershed was characterized by 25 variables (table 3). Variables reflecting the percent land use, the average straight-line distance from the NHD Plus defined stream to particular land use types, length of the stream within each subwatershed, drainage factor, and potential *E. coli* source population for each subwatershed, as calculated based on SELECT results, were included in the analysis. Each of the percent land use, average distance from land use to stream, length of stream, and drainage factor variables for each subwatershed were calculated using ArcGIS functions. The drainage factor was calculated by dividing the area of the subwatershed by the length of the stream within the subwatershed. 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 (Box and Cox, 1964; Juang et al., 2001) to ensure normality as required by factor and principal component analysis.

Table 3. Variables used to characterize subwatersheds.

Percent land use variables	Percent open developed Percent low-intensity developed Percent medium-intensity developed Percent high-intensity developed Percent open water Percent barren Percent forest land Percent near riparian corridor Percent mixed forest Percent rangeland Percent pasture Percent cultivated crops
Source population variables	Households using sewers Failing septic systems Cattle Sheep and goats Horses Dogs Deer Feral hogs
Average distance from stream and other variables	Average distance to wetland Average distance to forest Average distance to residential Average distance to pasture Drainage factor

Factor and principal component analysis (FAPCA) was conducted in order to reduce the number of variables while at the same time retaining the variability of the dataset (Jolliffe, 2002, pp. 111-119) using SAS (SAS, 2003). Factor analysis was performed on the normalized data in order to identify the factors that would affect the load of *E. coli* from a subwatershed. Both the Kaiser criterion and Scree test (Thyne et al., 2004; Jackson, 1993) were used to determine the number of factors to retain.

Cluster analysis was performed using the factors in a K-means clustering algorithm. The K-means clustering algorithm was employed to group the subwatersheds into clusters ranging from one to 35 clusters. The clusters were evaluated using the pseudo F (PSF) statistic, cubic clustering criterion (CCC), and silhouette width. In each of these statistics, the local maximum indicates an appropriate number of clusters (DeGaetano, 1996). Discriminant analysis was then conducted to identify discriminating variables. Based on the identified discriminating variables, the factor and cluster analyses were performed again. The final clusters were then further characterized using Duncan's multiple range test.

Results and Discussion

The results from SELECT for all sources are shown in figures 3 through 6. The larger potential *E. coli* loads are found in the darker shaded (red) subwatersheds. The mid-range loads are in the medium shaded (orange) subwatersheds, and the lowest loads are in the lightest shaded (white and yellow) subwatersheds.

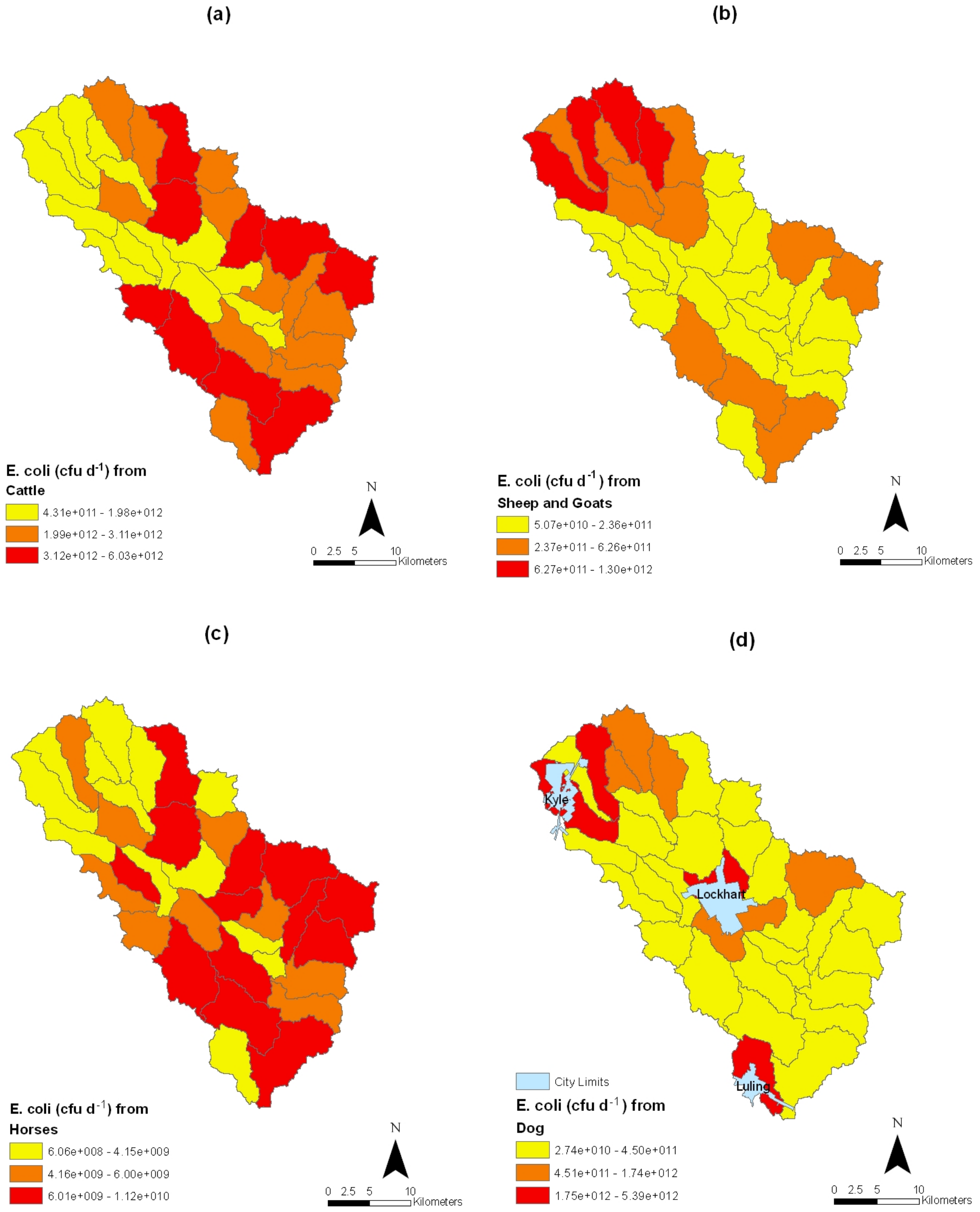

Figure 3. Average daily potential *E. coli* load in Plum Creek watershed resulting from various non-point domesticated animal sources: (a) cattle, (b) sheep and goats, (c) horses, and (d) dogs.

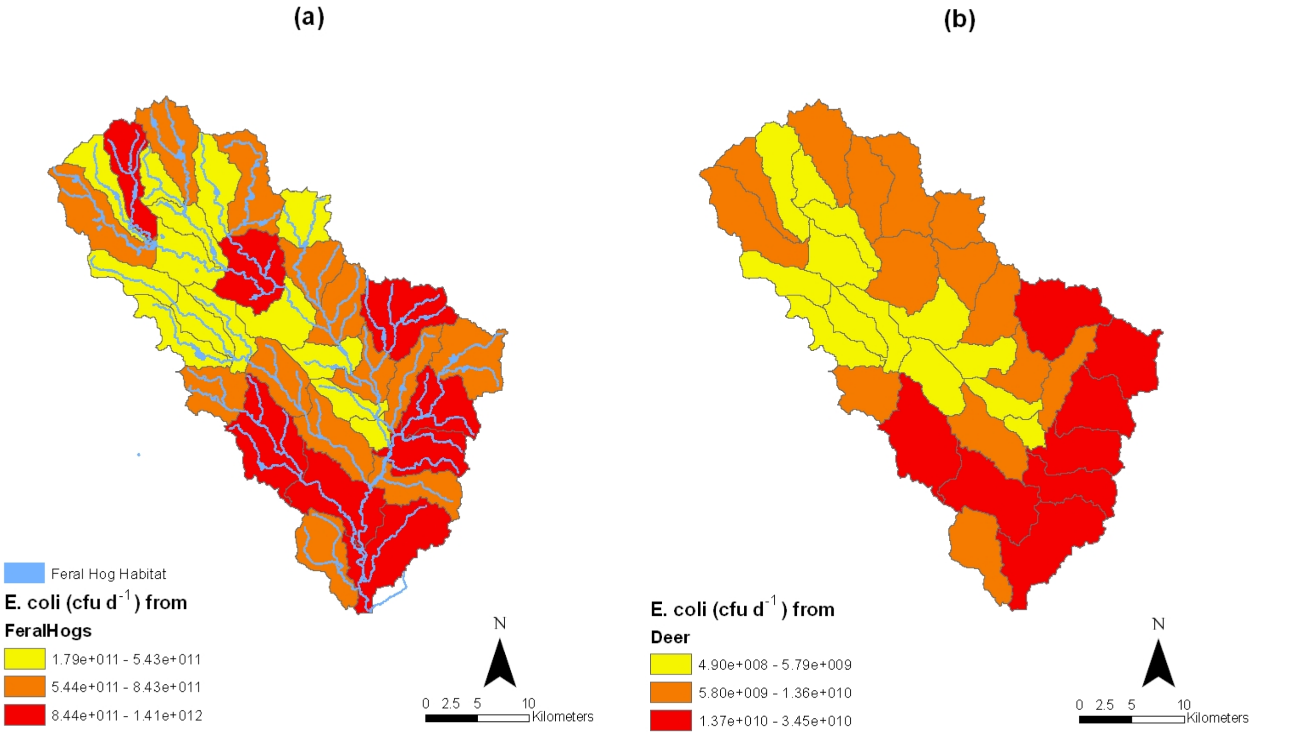

Figure 4. Average daily potential *E. coli* load in Plum Creek watershed resulting from various non-point wildlife sources: (a) feral hogs and (b) deer.

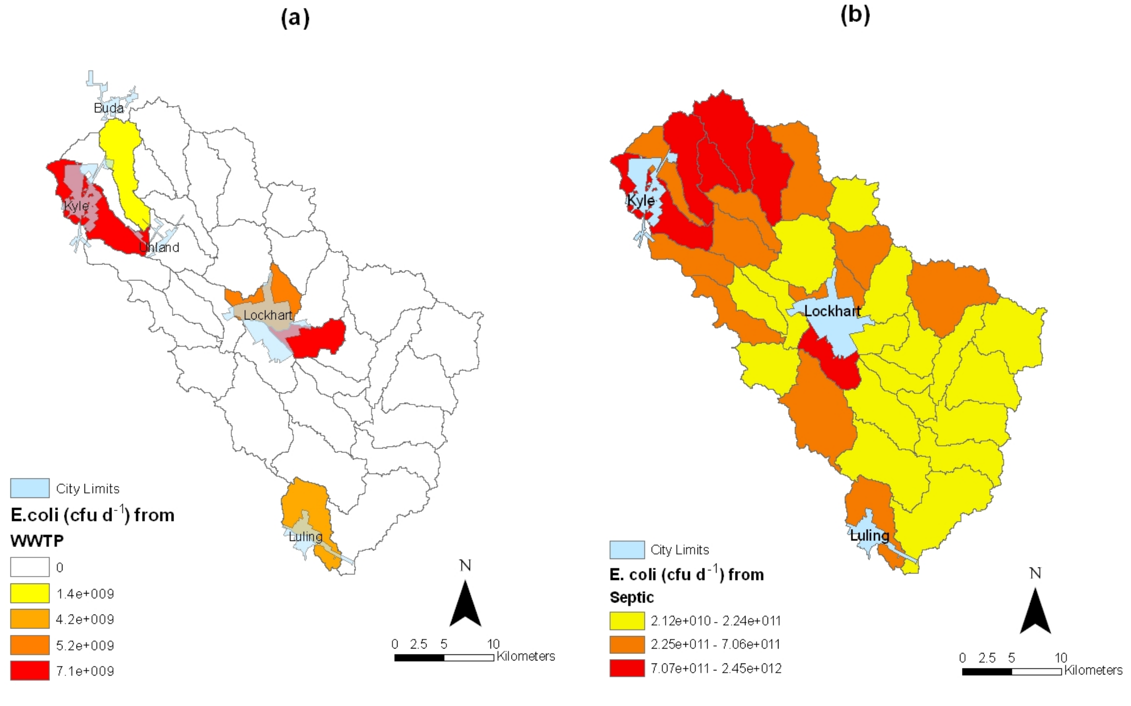

Figure 5. Average daily potential *E. coli* loading from (a) wastewater treatment plants (point source) and (b) failing OWTSs (non-point source).

Non-Point Sources

Livestock

Non-point sources include livestock, dogs, wildlife, and failing OWTSs. The load allocations from cattle, sheep and goats, and horses are shown in figures 3a, 3b, and 3c. Greater *E. coli* loads from cattle were estimated for subwatersheds on the southwestern portion of the watershed and along the southeastern edge (fig. 3). The subwatersheds that have larger estimated loads of *E. coli* from cattle have the highest amounts of land used for pasture and rangeland and range from first to 19th in percent land used for pasture and rangeland (fig. 2). In contrast, the high estimated *E. coli* potential subwatersheds for sheep and goats were in the north of the watershed (fig. 3). These subwatersheds have percentages of pasture and rangeland from second to 31st. Similar to subwatersheds that have the highest potential loads from cattle, these subwatersheds are characterized by a greater than 50% pasture and rangeland. Subwatershed 34

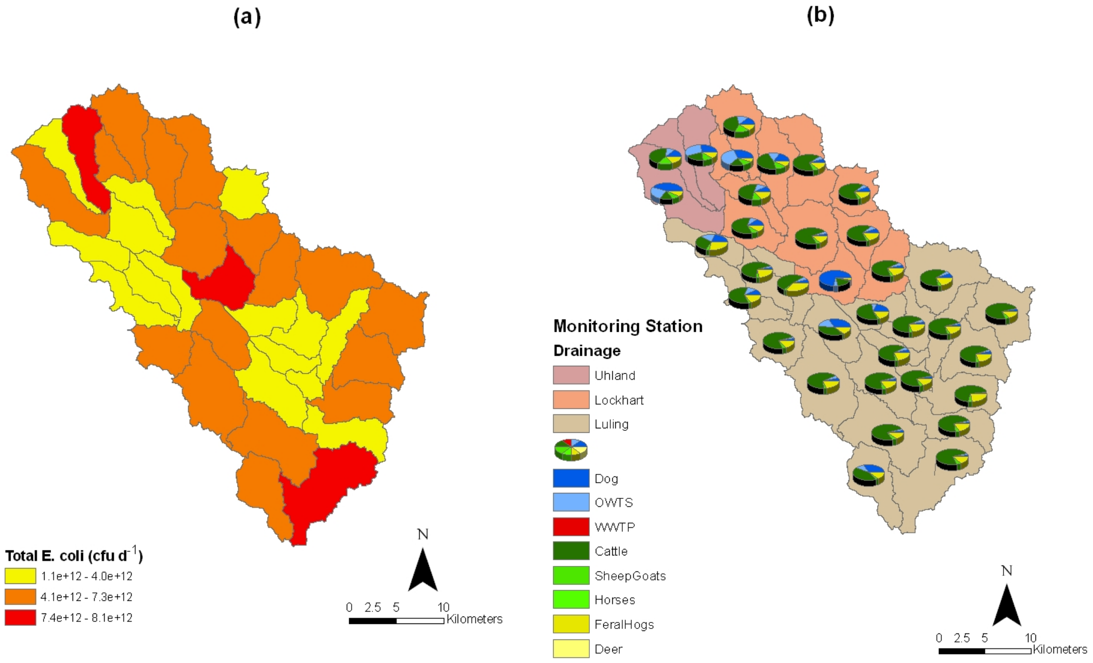

Figure 6. (a) Average daily total potential daily *E. coli* load and (b) relative percent contributions resulting from all sources in Plum Creek watershed.

was the exception, with a low percentage (38% and rank of 31st) of pasture and rangeland (fig. 2). Subwatershed 34's large load was due to its large area (fig. 1). The *E. coli* loads from sheep were estimated to be primarily in the northern part of the watershed, whereas the *E. coli* loads from cattle were estimated to be primarily in the southern portion of the watershed because, according to the USDA census, there was 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 were primarily found in the southern and middle section of the watershed (fig. 3), and these subwatersheds had large areas of pasture lands (fig. 2). When the total loads allocated to cattle, sheep and goats, and horses were compared (fig. 3), the magnitudes were quite different. The total estimated potential loads for cattle and sheep and goats were two orders of magnitude larger than the estimated load for horses (fig. 3). Because of the higher population of cattle, cattle had a larger potential load than sheep and goats. This suggests that agricultural BMPs such as riparian fencing, vegetative filter strips, and alternative watering (Anderson and Flaig, 1995) should be prioritized for cattle producers in the southern section of the watershed and for sheep and goat producers in the northern section of the watershed.

Pets

The potential *E. coli* load estimated from pets (dogs) is shown in figure 3. Subwatersheds with large allocations were associated with the cities of Kyle, Lockhart, and Luling. This could be attributed to the large number of households in the urban areas. In addition, as with OWTS failure, subwatersheds 1, 2, 4, and 7 (fig. 1) were estimated to have higher potential loads of *E. coli*. This area had higher population in comparison to the other subwatersheds, despite the lack of urban centers. The higher population of this area was attributed to urban sprawl from the nearby metropolitan area of Austin. This conclusion implies that 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.

Wildlife

The potential *E. coli* estimated from feral hogs is shown in figure 4. As stated in the methodology, feral hogs are distributed in the riparian areas around streams. Each subwatershed had an estimated potential contribution from feral hogs. The highest potential loads were in areas along the east and south of the watershed (fig. 4), where there was a large area of undeveloped land adjacent to a stream ranging in subwatershed rank from first to 18th. Feral hogs had an estimated potential load (fig. 4) that was of the same magnitude as cattle and sheep and goats (fig. 3). Unfortunately, best management practices to address *E. coli* contamination from feral hogs are quite challenging because fencing and other traditional practices 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 could be the most appropriate measures to implement in the southern portion of the watershed.

The potential *E. coli* load from deer is shown in figure 4. The southeastern portion of the watershed had the highest loads from deer, where there were large sections of range and forested areas that account for greater than 63% of the subwatershed. The estimated potential load for deer was two orders of magnitude smaller than the estimated load for feral hogs (fig. 4). This suggests that wildlife BMPs could be more efficiently focused on addressing feral hogs than deer.

OWTS Failure

The estimated potential *E. coli* load from OWTS failure is shown in figure 5. The darker subwatersheds indicate the larger estimated potential *E. coli* load. Larger loads (2.13 ×

10^{10} to 2.34×10^{12} cfu) were associated with subwatersheds that correspond to the cities of Lockhart and Kyle. However, large loads were also associated with subwatersheds 1, 2, 4, and 7 (figs. 5 and 1). These subwatersheds had the first, second, fourth, and fifth highest percentages of low-intensity development (fig. 2). The area in subwatersheds 1, 2, 4, and 7 (fig. 1) in the north of the watershed had a large population reported in the 2000 census, which was not yet incorporated into a city and thus not provided with sewer service. In addition, the average age of the subdivisions in subwatersheds 1, 4, and 7 were all pre-1988. As a result, the septic systems in these subwatersheds were unregulated. Therefore, the results suggest that BMPs within this region should address regulation of septic systems, focusing on proper operation and owner maintenance of the system (Lesikar, 2005).

Point Sources

Wastewater Treatment Plants

The estimation of potential *E. coli* loads from WWTPs is shown in figure 5. The five subwatersheds in which WWTPs are located are highlighted; the higher the permitted effluent discharge, the higher the estimated potential load and the darker subwatersheds are in figure 5. This suggests that best management practices such as tertiary treatment (Godfree and Farrell, 2005) or overflow monitoring would be most efficient for the subwatersheds that fall near the cities of Lockhart and Kyle.

Potential *E. coli* Sources Throughout the Watershed

In general, two sources, OWTSs and dogs, were considered to be both urban and rural sources. However, because these sources were estimated based on an even distribution corresponding to human populations, the larger estimated loads correspond to population centers. Thus, because of the underlying assumptions, the populations for rural areas may be underestimated. Based on the analysis of estimations, the contributions from urban areas would not only be larger in magnitude but also concentrated in a small area. In the rural areas, these sources were estimated to be diffuse and smaller in magnitude. The WPP should address these sources across the entire watershed. In urban areas, a total approach can be taken for dogs and OWTSs. Large BMPs that are in structural in nature, such as detention ponds that collect runoff, could be efficient in urban areas due to the magnitude of the load. For rural areas, homeowner education could be implemented to increase septic maintenance, but should focus particularly on residences near streams.

The total estimated potential *E. coli* loads in different subwatersheds are shown in figure 6. The darker subwatersheds (4.87×10^{11} to 3.95×10^{12} cfu) have the highest estimated potential loads. These subwatersheds correspond to urban areas, including the cities of Kyle, Lockhart, and Luling. Mid-range estimated loads (3.96×10^{12} to 1.02×10^{13} cfu) were highly influenced by regional effects (fig. 6). Figure 6 also shows the relative contribution of each source to the total estimated load for each subwatershed. The mid-range load subwatersheds in the northern section of the watershed show mixed influence of OWTSs, dogs, and agricultural animal sources (fig. 6). In the mid-range load subwatersheds in the southern and eastern portions of the watershed, a high percentage of the load was estimated from agricultural animals and wildlife sources (fig. 6).

Table 4 displays the subwatersheds with the highest potential *E. coli* contribution and the highest potential sources within each of these subwatersheds. Overall, cattle had the highest estimated potential contribution, with 41% of the total average potential *E. coli* load (table 3). The second highest potential daily contributor was urban runoff, with 27% of the total potential load. Dogs and feral hogs each had a potential of approximately 10.5% of the total potential load, and failing OWTSs comprise approximately 6.5% of the total. All other sources contributed less than 5% to the total potential load. Even though our analysis did not indicate that WWTPs were a major source of *E. coli*, regrowth of *E. coli* further downstream of the wastewater outfall should be addressed (Petersen et al., 2005). This can be achieved by combining SELECT with a fate and transport model to simulate *E. coli* population dynamics in streams (Steets and Holden, 2003). Statistical clustering techniques were implemented in this research to determine the influence of the variables describing the populations of sources of *E. coli* contamination in the watershed.

Statistical Clustering of Subwatersheds

Initial factor analysis was performed, and five factors (out of 25) were retained based on the results of the Scree test and Kaiser criterion. The identified factors were calculated from the subwatershed characterization variables for use in cluster analysis. After the subwatersheds were assigned membership to four clusters by means of the K-means clustering algorithm, discriminant analysis (DA) was used to identify the discriminating variables and test the cluster membership. Based on a 35% error rate between discriminant analysis and cluster analysis, factor analysis was repeated using the discriminating variables. The eight discriminating variables identified by DA have an average squared canonical correlation (ASCC) of 0.82 and thus account for 82% of the variability of the original dataset (Rencher, 1992). The discriminating variables included the population of cattle, estimated dogs, the estimated number of households using sewers, percent of open developed land, the average distance to wetlands, percent cultivated cropland, percent of medium developed land, and percent rangeland. These variables accounted for the greatest variability between the clusters.

Table 4. Ranking of Plum Creek subwatersheds receiving *E. coli* from high-contributing potential sources.

		Highest Potential Sources of *E. coli*				
Rank	Subwatershed	First Highest	Second Highest	Third Highest	Fourth Highest	Fifth Highest
1	34	Urban	Dogs	Septic systems	Cattle	Sheep and goats
2	16	Urban	Dogs	Cattle	Septic systems	Feral hogs
3	32	Urban	Cattle	Dog	Feral hogs	Septic systems
4	18	Urban	Cattle	Dog	Septic systems	Feral hogs
5	3	Urban	Cattle	Sheep and goats	Feral hogs	Dogs

Table 5. Cluster comparison using Duncan's multiple range test.

Variable	Duncan Results[a]	Cluster 1 (8 subwatersheds)	Cluster 2 (2 subwatersheds)	Cluster 3 (13 subwatersheds)	Cluster 4 (12 subwatersheds)
Percent open developed land	(2)(1,3,4)	Low[b]	High	Low	Low
Percent medium intensity	(2)(1,3,4)	Low	High	Low	Low
Percent rangeland	(1,3,4)(2,3,4)	High	Low	Medium	Medium
Percent cultivated crops	(2,3,4)(1,3)	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	(1)(2,3,4)	Low	High	Low	Low

[a] Similar clusters are grouped in parentheses. Dissimilar clusters are in different parenthetical groups.

[b] Low, medium, and high are qualitative descriptions of the cluster mean in relation to other cluster means.

These variables were used to calculate three factors as determined by the repeated FAPCA. Three factors were retained for use in cluster analysis. The first factor included the percent open developed land, the average distance to wetlands, and the number of households using sewers. This factor accounted for the variables describing urban development near streams. The second factor included the variables of percent rangeland and the population of cattle, accounting for ranching activities. The third factor included the variables percent medium intensity development and estimated dog populations, describing suburban development.

These new factors were calculated from the variables selected by discriminant analysis and used to reassign cluster membership, which changed the cluster membership of three subwatersheds. Then Duncan's multiple range test was performed to determine the similarity of the clusters for each discriminating variable. The results of the test are shown in table 5. Clusters that are grouped together in parentheses are similar, and clusters in different parenthetical groups are dissimilar. Each cluster was then given a qualitative ranking of high, medium, or low based on the average mean for that variable within each cluster.

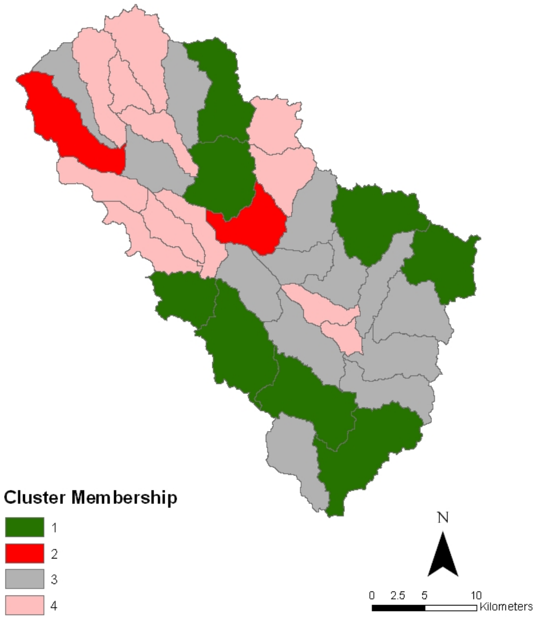

Figure 7. Final clusters of subwatersheds in Plum Creek watershed based on statistical clustering analysis.

Cluster 1

Cluster 1 had eight subwatersheds. These subwatersheds were on the southwestern and eastern edges of the watershed (fig. 7). Cluster 1 had the greatest mean of percent open developed land, rangeland, cattle, and distance to mixed forest. Duncan's multiple range test identified the cattle population of this Cluster as being significantly different from the other clusters' cattle populations (table 6). It would be most efficient for best management practices (BMPs) to focus on addressing loads from agriculture, such as cattle.

Cluster 2

Cluster 2 contained two subwatersheds, 34 and 16 (figs. 3 and 7). Both of these subwatersheds were urban areas encompassing the cities of Kyle and Lockhart. Duncan's multiple range test identified Cluster 2 as being distinctly different from the other clusters, with respect to the characteristics of dogs and percent medium-intensity development (table 6). When the discriminating variable cluster means were examined, Cluster 2 had a high mean for medium-intensity development, dogs, and sewers. Therefore, BMPs should focus on reducing loads from urban runoff, dogs, and wastewater treatment plant effluent.

Cluster 3

Cluster 3 contained 13 subwatersheds, with nine subwatersheds in the center of the southern portion of the watershed (fig. 7). The other four subwatersheds 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 3 was distinctive from all the other clusters (table 6). In addition, Cluster 3 did not have any variable means that were the highest or lowest of the four clusters. The cluster means and Duncan's multiple range test did not identify any general distinctive characteristics that would assist in identification and targeting of BMPs.

Table 6. Ranking of *E. coli* contributing potential sources in Plum Creek subwatersheds.

	Rank (1 to 5) of Subwatersheds (1 to 35)				
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

Cluster 4

Cluster 4 had four groupings of subwatersheds (fig. 7). Two groups of four subwatersheds were in the northern portion of the watershed, and two groups of two subwatersheds were located in the center and on the north central edge of the watershed. Duncan's multiple range test identified only the number of households using sewers as a variable that caused Cluster 4 to be significantly different from the other clusters (table 6); Cluster 4 was identified as having low numbers. Cluster 4 also had the highest mean of percent cultivated crops, but this distinguishing characteristic of Cluster 4 did not assist in decision making or placement of BMPs.

Conclusions

The SELECT methodology estimated the daily average potential *E. coli* production from specified sources within the Plum Creek watershed. It contributed to spatial understanding of the most appropriate placement of BMPs for efficient allocation of resources.

Plum Creek was statistically characterized in order to cluster the subwatersheds into groupings of management areas. Four clusters were identified: one cluster was high-density urban, one was high in cultivated crops, another was high in range and forest lands, and the fourth cluster had no distinguishing characteristics. The discriminating variables that distinguish the subwatersheds were identified. The variables of cattle population and dog population contributed most 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, engineers, and other specialists participating in technical water assessments as part of the TMDL process. It could provide input for watershed models that couple the potential input from SELECT and transport processes. When coupled with statistical cluster analysis, resources for BMPs could be efficiently allocated. The strength of the combination of SELECT and cluster analysis is that this method can guide stakeholders in determining what further refinements of the data are 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. It can also be applied to watersheds for which lack of data prohibits using other modeling techniques. Furthermore, SELECT can be modified to evaluate other water contaminants, such as nutrients, given sufficient information concerning application and production rates.

Currently, SELECT is being developed in Visual Basic for Applications (VBA) within ArcGIS 9.X to provide a graphical user interface (GUI). This automation will help users adjust project parameters for various pollutant loading scenarios, use the visual outputs to identify areas of greatest concern for contamination contribution, and incorporate that information while developing the WPP or the TMDL.

Acknowledgements

The authors would like to thank Dr. Mark McFarland, Nikkoal Dictson, Matt Berg, and Jennifer Jacobs for their insightful discussions and for providing data for this project. The authors would also like to acknowledge Plum Creek Watershed Partnership for their cooperation and assistance with this project.

References

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

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

AVMA. 2002. *U.S. Pet Ownership and Demographics Source Book*. Schaumburg, Ill.: American Veterinary Medical Association, Center for Information Management.

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. *Trans. ASABE* 49(4): 987-1002.

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

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

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

Cleland, B. 2002. TMDL development from the "bottom up": Part II. Using duration curves to connect the pieces. In *Proc. National TMDL Science and Policy 2002—WEF Specialty Conf.*, 7-14. Washington, D.C.: America's Clean Water Foundation.

DeGaetano, A. 1996. Delineation of mesoscale climate zones in the northeastern United States using a novel approach to cluster analysis. *J. Climate* 9(8): 1765-1782.

Doyle, M., and M. Erikson. 2006. Closing the door on the fecal coliform assay. *Microbe* 1(4): 162-163.

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

Geldreich, E. 1996. Pathogenic agents in freshwater resources. *Hydrologic Proc.* 10(2): 315-333.

Godfree, A., and J. Farrell. 2005. Processes for managing pathogens. *J. Environ. Qual.* 34(1): 105-113.

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

Hellgren, E. 1997. Biology of feral hogs (*Sus scrofa*) in Texas. In *Proc. Feral Swine Symposium*, 50-58. College Station, Tex.: 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.

Jolliffe, I. 2002. *Principal Component Analysis*. New York, N.Y.: Springer-Verlag.

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. *J. Environ. Qual.* 30(3): 894-903.

Kemper, J. 2000. Septic systems for dogs? *Nonpoint Source News-Notes* 63(Dec.): 14-15. Available at: www.state.nj.us/

dep/watershedmgt/pet_waste_fredk.htm. Accessed 21 May 2007.

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

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

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

Reed, Stowe, and 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, Tex.: 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.

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

Schueler, T. 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. *Trans. ASABE* 49(4): 1033-1049.

Steets, B., and P. Holden. 2003. A mechanistic model of runoff-associated fecal coliform fate and transport through a coastal lagoon. *Water Res.* 37(3): 589-608.

SWAT. 2005. *Soil and Water Assessment Tool.* Version SWAT2005. Temple, Tex: USDA-ARS Grassland, Soil and Water Research Laboratory.

TCEQ. 2004. Guidance for assessing Texas surface and finished drinking water quality data, 2004. Austin, Tex: Texas Commission on Environmental Quality, Surface Water Quality Monitoring Program. Available at: www.tceq.state.tx.us/assets/public/compliance/monops/water/04twqi/04_guidance.pdf. Accessed 25 January 2007.

TCEQ. 2006. Water quality inventory and 303(d) list. Austin, Tex.: Texas Commission on Environmental Quality. Available at: www.tceq.state.tx.us/compliance/monitoring/water/quality/data/wqm/305_303.html#y2006. Accessed 24 August 2008.

Texas State Demographer. 2006. Texas population estimates program. Austin, Tex.: Texas State Data Center and Office of the State Demographer. Available at: http://txsdc.utsa.edu/tpepp/txpopest.php. Accessed 15 May 2007.

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

USCB. 2000. Census 2000 TIGER/Line files. Washington, D.C.: U.S. Census Bureau. Available at: www.census.gov/geo/www/tiger/index.html. Accessed 14 December 2006.

USDA-NASS. 2002. 2002 Census of agriculture: County data, 560-634, 716-718, 719-729, 730-732, 733-734. Washington, D.C.: USDA National Agricultural Statistics Survey.

USEPA. 1986. Ambient water quality criteria for bacteria, 15-16. EPA440/5-84-002. Washington, D.C.: U.S. Environmental Protection Agency, Office of Water Regulations and Standards.

USEPA. 1991. Guidance for water quality-based decisions: The TMDL process, 3-1 - 3-17. EPA440/4-91-001. Washington, D.C.: U.S. Environmental Protection Agency, Office of Water.

USEPA. 1996. TMDL development cost estimated: Case studies of 14 TMDLs, I-6 - I-17. EPA841-R/96/001. Washington, D.C.: U.S. Environmental Protection Agency, Office of Water.

USEPA. 1999. Draft guidance for water quality based decisions: The TMDL process, 3-1 - 3-18. 2nd ed. EPA841-D-99-001. Washington, D.C.: U.S. Environmental Protection Agency, Office of Water.

USEPA. 2001. Protocol for developing pathogen TMDLs: Source assessment, 5-1 - 5-18. 1st ed. EPA841-R-00-002. Washington, D.C.: U.S. Environmental Protection Agency, Office of Water.

USGS. 2002. National hydrography dataset. Reston, Va.: U.S. Geologic Survey. Available at: http://nhd.usgs.gov/index.html. Accessed 4 August 2006.

Weiskel, P., B. Howes, and G. Heufelder. 1996. Coliform contamination and transport pathways. *Environ. Sci. and Tech.* 30(6): 1872-1881.

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 Eng. in Agric.* 21(5): 879-889.