GROUPWISE MODELING STUDY OF BACTERIALLY IMPAIRED WATERSHEDS IN TEXAS: CLUSTERING ANALYSIS[1]

Sabu Paul, Raghavan Srinivasan, Joaquin Sanabria, Patricia K. Haan Saqib Mukhtar, and Kerry Neimann[2]

ABSTRACT: Under the Clean Water Act (CWA) program, the Texas Commission on Environmental Quality (TCEQ) listed 110 stream segments in the year 2000 with pathogenic bacteria impairment. A study was conducted to evaluate the probable sources of pollution and characterize the watersheds associated with these impaired water bodies. The primary aim of the study was to group the water bodies into clusters having similar watershed characteristics and to examine the possibility of studying them as a group by choosing models for total maximum daily load (TMDL) development based on their characteristics. This approach will help to identify possible sources and determine appropriate models and hence reduce the number of required TMDL studies. This in turn will help in reducing the effort required to restore the health of the impaired water bodies in Texas. The main characteristics considered for the classification of water bodies were land use distribution within the watershed, density of stream network, average distance of land of a particular use to the closest stream, household population, density of on-site sewage facilities (OSSFs), bacterial loading from different types of farm animals and wildlife, and average climatic conditions. The climatic data and observed instream fecal coliform bacteria concentrations were analyzed to evaluate seasonal variability of instream water quality. The grouping of water bodies was carried out using the multivariate statistical techniques of factor analysis/principal component analysis, cluster analysis, and discriminant analysis. The multivariate statistical analysis resulted in six clusters of water bodies. The main factors that differentiated the clusters were found to be bacterial contribution from farm animals and wildlife, density of OSSFs, density of households connected to public sewers, and land use distribution.

(KEY TERMS: nonpoint source pollution; water quality; statistical analysis; total maximum daily load (TMDL); coliform bacteria; cluster analysis.)

Paul, Sabu, Raghavan Srinivasan, Joaquin Sanabria, Patricia K. Haan, Saqib Mukhtar, and Kerry Neimann, 2006. Groupwise Modeling Study of Bacterially Impaired Watersheds in Texas: Clustering Analysis. Journal of the American Water Resources Association (JAWRA) 42(4):1017-1031.

INTRODUCTION

According to the Code of Federal Regulations (CFR) Title 40, Part 131, all states, territories, and authorized tribes of the United States must update the list of impaired water bodies – the CWA §303(d) list – and submit it to the U.S. Environmental Protection Agency (USEPA) for its approval (USEPA, 1998) and for development of the total maximum daily load (TMDL) for each of these water bodies. Among the many pollutants that require the development of TMDLs, fecal coliform is included because it indicates a serious potential health risk. Fecal coliforms are a group of bacteria that primarily live in the lower intestines of warm blooded animals, including humans. The presence of a high concentration of fecal coliform bacteria may indicate the presence of dangerous pathogens. Under the Clean Water Act (CWA) program, the Texas Commission on Environmental

[1]Paper No. 04216 of the *Journal of the American Water Resources Association* (JAWRA) (Copyright © 2006). **Discussions are open until February 1, 2007.**

[2]Respectively, Environmental Engineer, Tetra Tech, Inc., 10306 Eaton Place, Suite 340, Fairfax, Virginia 22030; Director, Spatial Sciences Laboratory, Texas Agricultural Experiment Station, 1500 Research Parkway, Suite 221E, College Station, Texas 77845; Blackland Research Center, Texas A&M University System, 720 East Blackland Road, Temple, Texas 76502; (Haan and Mukhtar) Assistant Professor and Associate Professor, Biological and Agricultural Engineering Department, Texas A&M University, College Station Texas 77843; and Project Manager,Texas Commission on Environmental Quality, 12100 Park 35 Circle, Bldg. F, Austin, Texas 78753 (E-Mail/Paul: Sabu.Paul@tetratech-ffx.com).

Quality (TCEQ) listed 110 water bodies (Figure 1) in 2000 with indicator bacteria concentrations greater than those permitted for the respective water bodies' designated uses (TNRCC, 2000). Once the impairment is verified, the development of TMDLs for these water bodies seems to be the best solution for the problem. However, developing a TMDL for every one of these stream segments will require an enormous amount of input, in terms of both capital and human labor. A case study conducted by the U.S. Environmental Protection Agency (USEPA) showed that the cost of a single TMDL study varied between $4,039 and $1,023,531 (USEPA, 1996). It was noted that on average 32 percent of the total expense was allotted for the modeling component of the TMDL studies. Many of the water bodies considered for TMDL development listed under the current CWA §303(d) for Texas may have similar characteristics and hence may be grouped based on their watershed characteristics and possible sources of pollution. Such a grouping scheme would be helpful in reducing the cost of restoration of water quality by restricting the development of the TMDL to one or two representative water bodies under a single group and applying the knowledge to other water bodies in the same group.

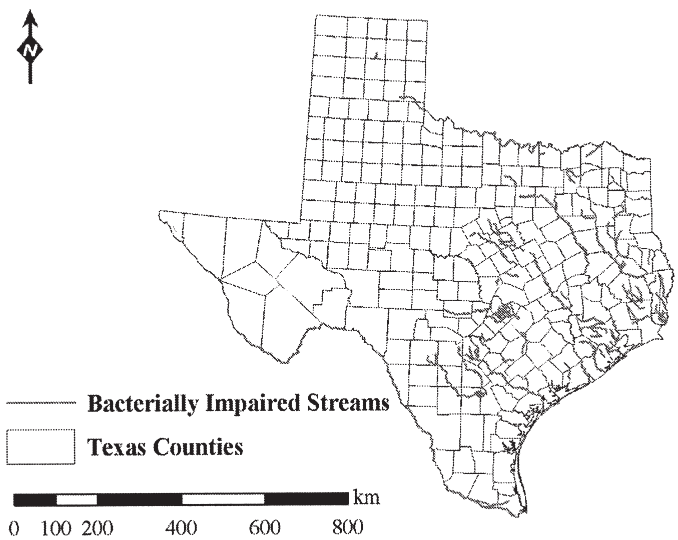

Figure 1. Network of Stream Segments Listed for Bacterial Impairment in Texas in 2000 (TNRCC, 2000).

This paper focuses on the development of a method for classifying the Texas water bodies listed for bacterial quality violation under CWA §303(d) into groups having similar watershed characteristics by using multivariate statistical techniques. The classification of the impaired water bodies was done using the various multivariate analysis techniques of factor analysis/principal component analysis (FA/PCA), cluster analysis (CA), and discriminant analysis (DA). A brief description of FA/PCA, CA, and DA follows.

Factor analysis is one of the most common multivariate statistical techniques used to reduce the dimensionality of large sets of variables (Karson, 1982). Factor analysis is used to analyze the interrelationships among different variables and to find common factors, thus to condense the information contained in a large number of variables into a smaller set of factors without sacrificing much information. The two main types of FA are PCA and common factor analysis.

Principal component analysis is used to create linear combinations of the original variables into a smaller set of new variables, the principal components, which explain the maximum amount of variance possible (Karson, 1982). These principal components are orthogonal to each other, and thus they are uncorrelated. Successive principal components account for decreasing proportions of the total variances of the original variables. The FA/PCA has been used in many water quality assessment studies (Vega *et al.,* 1998; Helena *et al.,* 2000). Alberto *et al.* (2001) used the FA/PCA technique in a study to evaluate the spatial and temporal changes of water quality in Suquía River Basin, Argentina.

Cluster analysis, also known as unsupervised pattern recognition, is a set of statistical techniques for exploratory data analysis and is used to classify a set of observations into multiple groups based on multivariate properties (Karson, 1982). The groups are formed in such a way that the observations are highly internally homogeneous and highly externally heterogeneous, meaning the members of a group are similar to one another and are different from members of other groups. Different methods of CA produce entirely different results, and they are interpreted based on the particular need.

Discriminant analysis is used to determine the variables that can discriminate among groups. A DA is used when membership of different observations to a given group is known *a priori*. A discriminant function is constructed for each group and is used to divide the observations into different regions in the data space (Karson, 1982). Discriminant analysis has been used in many studies to identify the sources of fecal pollution in aquatic systems using antibiotic-resistance patterns (Parveen *et al.,* 1999; Wiggins *et al.,* 1999; Whitlock *et al.,* 2002; Choi *et al.,* 2003). It also has been used to verify the efficiency of grouping schemes produced by CA (Clucas, 1997; Alberto *et al.,* 2001).

Jenerette *et al.* (2002) used a combination of PCA, CA, and DA to evaluate the effectiveness of delineating aquatic systems based on the ecoregion approach. Alberto *et al.* (2001) used a similar approach in a study to evaluate the spatial and temporal changes of water quality in Suquía River Basin, Argentina, using a combination of FA/PCA, CA, and DA.

METHODOLOGY

Overview

The sources of fecal coliform bacteria are divided into point and nonpoint sources. The main nonpoint sources of fecal coliform bacteria are wildlife, livestock not confined within a feedlot, poultry, humans, domestic pets, OSSFs, leaky sewer lines, and migratory birds. The point sources of fecal coliform bacteria are confined animal feeding operations, discharges from wastewater treatment plants, and fish and shellfish processing facilities (USEPA, 2001b). In this study the possible sources of fecal coliform bacteria within each impaired watershed were assessed, and based on the available information, groups of watersheds with similar characteristics were developed. The specific steps used in this study were to identify the contributing area for each impaired watershed; collect data regarding the bacterial sources and watershed characteristics for impaired watersheds; normalize the data to be used in multivariate statistical analysis; conduct the FA/PCA to find the principal components; group the watersheds with CA using the principal components; and conduct descriptive DA to find variables that discriminate the clusters and the accuracy of CA results.

The procedure starts with delineating the watershed that contributes flow to each water body listed for bacterial impairment. The watersheds delineated were analyzed using a geographic information system (GIS) to obtain characteristics such as land use distribution, distance of land of a particular use to the nearest stream, stream density within the contributing area, number of potentially failing OSSFs, number of wildlife, number of livestock, average precipitation, and average atmospheric temperature. A data matrix containing the characteristics of all the watersheds was compiled. This matrix was analyzed using FA/PCA, CA, and DA, and the water bodies were clustered based on their characteristics. The seasonal variability in the observed instream bacterial concentration was analyzed to identify the type of bacterial source, point or nonpoint, within each cluster. The GIS analysis was done using ArcView GIS software (ESRI, 1999). The statistical analysis was done with SAS software (SAS, 1999).

GIS Data Requirement

The important GIS datasets used for this study included the land use distribution, soil distribution, elevation, stream network, and locations of flow gauging stations, water quality monitoring stations, and precipitation gauging stations. The data layers for land use, soil, elevation, and stream network were obtained from National Land Cover Data (NLCD) from the U.S. Geological Survey (USGS, 2002b), STASTGO from USEPA (2002), Digital Elevation Model (DEM) from TNRIS (2002), and National Hydrography Dataset (NHD) from USGS (2002a), respectively. The water quality data and locations of the stations were from TCEQ (J. Allen, personal communication, 2002). Flow data and locations were obtained from USGS (2002c). Rainfall data and station locations were from National Climatic Data Center (NCDC)-National Oceanic and Atmospheric Administration (NOAA) (NCDC, 2002).

Watershed Delineation, Drainage Density, and Distance Factor

The watershed delineation was done either using the GIS data layers such as the 14-digit hydrologic unit code (HUC) boundaries and the NHD stream network or based on the DEM with the help of the automatic watershed delineation tool available within BASINS (USEPA, 2001a).

The transport rate of pathogens from different types of land use is a function of the loading rate of the pathogens on the land surface and the runoff to the stream. The density of the stream network and the distance of the pollutant source from the stream influence runoff rate and hence the contribution of pollutants from a nonpoint source to the stream. The drainage density was determined by dividing the total length of the NHD stream network within the watershed by the total watershed area. The average distance between source areas based on their land use to the nearest stream was determined using the NLCD land use layer and NHD stream network.

Water Quality Analysis

High concentrations of bacteria during low flow, dry weather conditions usually indicate continuous loading from point sources (USEPA, 2001b). When the concentration of bacteria is higher during storm events, there is a high probability that impairment is the result of nonpoint sources. Water quality observations from grab samples collected by the TCEQ from 1985 to 2000 for each water quality monitoring station were separated into base flow periods and rainfall periods, based on the streamflow data at USGS gauge stations close to the water quality stations or the precipitation data at neighboring weather stations maintained by NCDC. The data were compiled for all the water quality stations that fell within a single watershed. A pooled t-test (Milton and Arnold, 1995) was used to test the hypothesis that there was a significant difference between mean instream fecal coliform concentration during periods with rainfall and periods without rainfall. The details of the water quality flow/rainfall analyses are given in Paul (2003).

Human Source Assessment

The contribution of fecal coliform bacteria from humans through the discharge from failed septic systems is undisputable (USEPA, 2001b). Although an accurate calculation of bacterial loading from human sources is not possible, it can be assumed that the rate of loading would be proportional to the population and number of households within the contributing watershed of an impaired stream segment. Though other factors such as soils, depth to water table, and age of systems may affect the bacterial contribution, the current study considers only the population and number of households.

The population and number of households within a watershed were calculated by summing data from U.S. Census blocks intersecting the watershed under consideration. The population data for each Census block were obtained from the U.S. Census Bureau (2000). The number of households connected to public sewer systems and the number of OSSFs installed before 1990 were obtained from the U.S. Census Bureau (1990). The number of OSSFs installed after 1990 was obtained from the TCEQ, and GIS layers were used to identify the geographical locations, population, and number of households that utilize OSSFs or public sewer systems. It was assumed that the OSSFs were present only outside major cities. The assumption was that households located within major cities or Metropolitan Statistical Areas (MSAs) were connected to the public sewer systems. The number of OSSFs within the watershed was estimated based on the county population and proportion of households within a county that are located within the watershed boundary.

Animal Source Assessment

The animal sources of fecal coliform bacteria assessed for this study were wildlife and livestock. Availability of information regarding wildlife was limited to the number of white-tailed deer by county in Texas. Based on the deer population in Texas counties obtained from the Texas Parks and Wildlife Department, watershed level deer populations were estimated based on the percentage of area under forestland, barren land, and pastureland. The data on cattle, swine, sheep, and goats were obtained from the U.S. Department of Agriculture (USDA) Agricultural Statistics Database (USDA, 2002) and from concentrated animal feeding operations (CAFOs). The cattle data were available for each county. Goat, swine, and sheep data were available partially on the county level and partially on the agricultural district level. The locations and numbers of CAFOs were obtained from the TCEQ. The watershed level livestock population was estimated based on the percentage of area under pastureland. The details of the watershed level livestock and wildlife population calculations can be found in Paul (2002).

Normalization of Data

Many statistical tests are based on the assumption of normality. Hence, the distributions of the data for each of the variables were tested to see if they fit a normal distribution using a Kolmogorov-Smirnov goodness-of-fit test (Haan, 2002). Since the data for many of the variables were not found to be normally distributed, the data were transformed using the Box-Cox family of transformations given by Box and Cox (1964)

$$T(X) = (X^{\lambda} - 1)/\lambda \qquad (1)$$

where X is the original variable and λ is the transformation parameter. For $\lambda = 0$, the data are transformed using the natural log. In this study different values of λ were tried for each variable, and the transformed data were tested using Kolmogorov-Smirnov goodness-of-fit test. A λ was selected for which the transformed data were found to be normally distributed.

Factor Analysis/Principal Component Analysis

An FA/PCA (Srivastava and Carter, 1983) was used to identify the factors most important for clustering the watersheds from the group of factors known to affect instream fecal coliform concentrations. Each of these selected principal components is a linear combination of the original variables. The number of principal components considered for CA is based on the percentage of variance explained by the factors. The criteria used to select the number of factors retained were the Kaiser criterion (Kaiser, 1960) and the Scree test (Cattell, 1966). According to the Kaiser criterion, a factor is retained only if the eigenvalue is greater than 1. Essentially this means a factor is selected only if it extracts at least as much variance as the equivalent of one original variable. The Scree test is visual test where the eigenvalues are plotted against the number of factors. The general rule is to select the number of factors corresponding to a point beyond which the curve becomes approximately horizontal.

Cluster Analysis

The watersheds were grouped using hierarchical CA (Nathan and McMahon, 1990; SAS, 1999). The hierarchical method uses a sequential method for forming clusters, starting with the most similar pair of objects, and then forming higher clusters in a stepwise fashion. The similarity measure generally used is the Euclidean distance, which is given as

$$d_{ij} = \left(\sum_{k=1}^{p} \left(X_{ik} - X_{jk} \right)^2 \right)^{1/2} \tag{2}$$

where d_{ij} is the distance between the i^{th} and j^{th} observations, X_{ik} is the value of ith observation for the kth variable of p variables.

A CA was performed with the factors obtained during the FA/PCA step using the most frequently used Ward's minimum variance method (Kalkstein *et al.,* 1987). This method was selected after several trials using different methods. Ward's method has been used in many CA studies (Vega *et al.,* 1998; Helena *et al.,* 2000; Alberto *et al.,* 2001). The clustering is carried out by minimizing the within-cluster sum of squares, W, which is given as

$$W = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{N_k} \left(X_{ijk} - X_{.jk} \right)^2 \tag{3}$$

where K is the number of clusters, X_{ijk} is the value of the j^{th} variable for the i^{th} observation in the k^{th} cluster, J is the total number of variables, N_k is the number of observations in k^{th} cluster, and $X_{.jk}$ is the k^{th} cluster sample mean of j^{th} variable.

The determination of the number of clusters is difficult and has no perfect solution. Criteria used to decide the number of clusters were pseudo t^2 statistic, pseudo F statistic, and cubic clustering criterion (CCC), the number of clusters. The idea is to select the number of clusters corresponding to the local peak value in the pseudo F statistic and the CCC combined with a small value of the pseudo t^2 statistic and a larger pseudo t^2 for the next cluster fusion.

Cluster Mean Comparisons

The means of the variables were compared across the clusters using Duncan's multiple range test. Results were compared at the 95 percent confidence interval (α = 0.05). Based on the results from the mean comparisons, the clusters were labeled as high, medium, or low for the individual variables. This information along with the graphical plots of the means were used to determine the characteristics of the clusters.

Discriminant Analysis

To test the effectiveness of the clustering method and to determine the important parameters that discriminate among the clusters, the results from the CA were analyzed using a DA technique. The stepwise DA adds one variable at each step, starting with no variable, and examines the model to check for variables failing to meet the criterion to remain part of the model. If all variables in the model meet the criterion, a new variable that contributes the most to the discriminatory power of the model is entered. When all variables in the model meet the criterion to remain and none of the other variables meet the criterion to enter, the stepwise selection process stops. Discriminant analysis generates a function called the discriminant function, similar to multiple regression, to determine the group membership criteria. The discriminant function is created as a linear combination of discriminating (independent) variables as given by Johnson and Wichern (2002),

$$f(G_i) = k_i + \sum_{j=1}^{n} w_{ij} \cdot p_{ij} \tag{4}$$

where i is the number of groups (G), k_i is the constant inherent to each group (clusters in this case), n is the number of discriminating variables in group i, w_{ij} is the weight coefficient assigned by the DA to a j^{th} variable, and p_{ij} is the analytical value of the j^{th} variable. Based on the discriminant function, DA produces a classification matrix that can be used to determine the effectiveness of a given classification scheme. In the current study, based on the discriminant function, the DA analyzed the differences between groups and helped to reassign water bodies that have been wrongly assigned to a cluster by the CA to the appropriate cluster.

The DA ranks the variables used based on their F-values. The F-value indicates the statistical significance in the discrimination of clusters, that is, the contribution by a variable in prediction of the cluster membership. In each step of the DA the variable with the highest F-value will be selected. The Wilk's λ value is the fractional amount of within-cluster variance, relative to the between cluster variance, that remains unaccounted for after each variable is entered or selected in each step of the discriminant analysis. As more variables are selected, the value of Wilk's λ decreases. The average squared canonical correlation (ASCC) is the proportion of the variance accounted for by the selected independent variables. The larger the value of ASCC, the better is the ability of the variables to capture the overall variance in the data matrix. As new variables are added, the value of ASCC is also increased. The results from stepwise DA were used to generate a classification matrix based on a discriminant criterion.

RESULTS AND DISCUSSION

Water Quality Analysis

The list of water quality sampling points and the observed instream bacterial concentrations were obtained from the TCEQ (J. Allen, personal communication, 2002). The USGS flow data for all the USGS gauging stations within the impaired watersheds were obtained from the USGS (2002c). The precipitation data corresponding to the NCDC weather stations were obtained from NCDC (2002). First, the bacterial stations closest to the USGS stations were identified, and the observed bacterial concentrations were separated for periods of base flow driven streamflow and rainfall event driven streamflow. Then the means of observed instream fecal coliform bacteria concentrations for the two groups were compared to find out whether they were statistically significantly different. This process was repeated for all the watersheds. Of 110 water bodies, 67 showed higher mean bacterial concentrations during rainfall event driven streamflow periods, and 21 showed higher mean bacterial concentrations during base flow driven streamflow periods. Eleven water bodies showed no significant difference in the means of bacterial concentrations during the two periods. There were not enough data for analysis for these 11 water bodies. The detailed results are given in Paul (2003).

Normalization of Data

The Box-Cox parameter, λ, used for transforming each variable is listed in Table 1. The temperature and rainfall data were not transformed.

Principal Component Analysis

Based on the initial cluster analysis, the parameters related to the distance factors were excluded from the FA/PCA. The results of the PCA are reported in Table 1. Based on the Kaiser criterion and the Scree plot (Figure 2), six factors were retained for the cluster analysis. The cumulative variance explained by the six factors was 97 percent. Different authors report different values to distinguish major loading. Alberto *et al*. (2001) consider 0.7 as significant loading, while Carlon *et al*. (2001) consider this as 0.8. However, Yung *et al*. (2001) consider a much lower value (< 0.45). In the current study, all the parameters that have a magnitude of 0.6 or more were considered to be contributing significantly to a particular factor. Thus Factor 1 had five parameters with magnitudes greater than 0.6. These parameters are related to the human population. The second factor had two parameters with loadings greater than 0.6 in magnitude – percent wetland and average precipitation. In Factor 3, OSSF density and density of other septic systems had a high positive magnitude. The rate of bacterial loading from livestock was highly correlated to Factor 4. Factor 5 had high magnitudes of forestland and cropland. The main component of the sixth factor was the average temperature. Fourteen parameters were included to account for 97 percent of the overall variance. Other parameters had relatively low magnitudes on any of the factors retained for the analysis.

Cluster Analysis

In the current study there was no clear guidance from any of the criteria for the number of clusters. After some initial analysis based on criteria such as pseudo t^2 statistic, pseudo F statistic, and CCC, it was decided to obtain six clusters of watersheds.

TABLE 1. Varimax Rotated Factor Loading for the First Six Factors.*

Variable	λ	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Density of Households	0.005	<u>0.98</u>	0.07	0.14	-0.01	-0.05	0.07
Population Density	-0.05	<u>0.97</u>	0.10	0.14	0.00	0.01	0.12
Density of Public Sewers	0.01	<u>0.97</u>	0.06	0.13	-0.09	-0.03	0.04
Percent Residential	0.175	<u>0.90</u>	0.11	0.10	-0.22	-0.14	0.07
Percent Commercial	-0.1	<u>0.86</u>	0.19	0.14	-0.03	-0.08	0.16
Percent Wetland	0.25	-0.05	<u>0.83</u>	-0.01	0.11	0.02	0.07
Average Precipitation	NA**	0.11	<u>0.78</u>	0.13	0.30	-0.25	0.09
Bacterial Loading from Deer	0.5	-0.30	-0.51	-0.10	0.46	0.05	0.01
Percent Barren Land	0.1	-0.36	-0.76	-0.12	0.03	-0.08	0.22
Density of Other Septic Systems	-0.1	0.26	0.10	<u>0.92</u>	0.15	-0.05	0.19
Density of OSSFs	-0.02	0.32	0.17	<u>0.85</u>	0.28	0.02	-0.02
Average Age of Households		0.04	-0.04	<u>0.70</u>	-0.30	-0.12	-0.10
Bacterial Loading From Farm Animals	0.35	-0.14	0.13	0.10	<u>0.67</u>	-0.20	-0.05
Percent Forestland	0.44	-0.03	0.26	0.01	0.58	-0.47	0.17
Percent Pastureland	0.5	-0.10	0.10	0.03	-0.05	<u>0.69</u>	-0.28
Percent Cropland	0.12	-0.15	-0.24	-0.14	-0.24	<u>0.61</u>	0.03
Percent Water	Log Normal	0.13	0.29	-0.21	-0.12	0.32	0.20
Average Temperature	NA**	0.18	0.00	0.04	0.01	-0.14	<u>0.60</u>
Eigenvalues		4.91	2.48	2.26	1.42	1.36	0.66
Cumulative Percent of Variance		36.50	54.97	71.77	82.33	92.44	97.37

*Underlined values are considered to have significant loading for a magnitude greater than 0.6.
**No transformation is used.

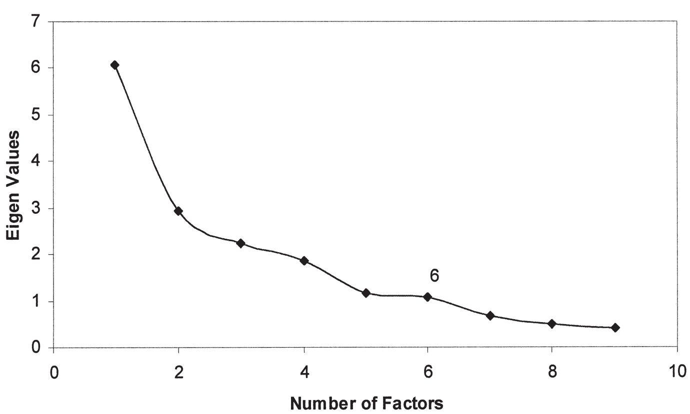

Figure 2. Scree Plot for Determining the Number of Factors to be Retained.

Discriminant Analysis

The CA results were analyzed using a DA technique. The CA produced six clusters, and the cluster membership information was added to the original data matrix. Initially, a stepwise DA was performed to obtain the discriminating variables. The summary result from the stepwise DA is given in Table 2. Some of the variables used in the CA procedure were excluded by the DA. The classification matrix and the corresponding error statistics are shown in Table 3, which shows the number of watersheds placed by DA into a specified cluster compared to the number of watersheds assigned in that cluster during CA. The larger the numbers in the diagonal elements in the matrix, the better is the grouping scheme based on the discriminating variables. It can be seen from the two tables that assignment of three water bodies in the third cluster by CA was not in agreement with the assignment by DA, accounting for a 2 percent error rate. Also the discriminating parameters were found to be different from those selected for the CA. Hence, it was decided to rerun the CA using only the discriminating parameters. Priors indicate the probabilities of a particular item being grouped under a given cluster if the assignment was done randomly.

TABLE 2. Stepwise Selection Summary of DA.

Step	Parameter Entered	Partial R^2	F Value	Pr > F	Wilk's λ	Pr > λ	ASCC*	Pr > ASCC
1	Age of Households	0.90	188.6	< 0.0001	0.099	< 0.0001	0.18	< 0.0001
2	Average Precipitation	0.76	65.36	< 0.0001	0.024	< 0.0001	0.33	< 0.0001
3	Loading Rate From Farm Animal	0.63	35.35	< 0.0001	0.009	< 0.0001	0.45	< 0.0001
4	Households	0.47	18.14	< 0.0001	0.005	< 0.0001	0.53	< 0.0001
5	Forestland	0.60	30.48	< 0.0001	0.002	< 0.0001	0.62	< 0.0001
6	Wetland	0.39	12.65	< 0.0001	0.001	< 0.0001	0.65	< 0.0001
7	Distance Factor Pasture	0.26	6.79	< 0.0001	0.001	< 0.0001	0.69	< 0.0001
8	Alternate Septic Systems	0.26	6.76	< 0.0001	0.001	< 0.0001	0.70	< 0.0001
9	Pastureland	0.28	7.31	< 0.0001	0.000	< 0.0001	0.72	< 0.0001
10	Average Temperature	0.20	4.70	0.0007	0.000	< 0.0001	0.73	< 0.0001
11	Loading Rate From Deer	0.19	4.42	0.0012	0.000	< 0.0001	0.74	< 0.0001
12	Cropland	0.14	3.05	0.0135	0.000	< 0.0001	0.75	< 0.0001
13	Distance Factor – Water	0.15	3.12	0.012	0.000	< 0.0001	0.75	< 0.0001
14	Distance Factor – Forest	0.12	2.54	0.0337	0.000	< 0.0001	0.76	< 0.0001
15	Distance Factor – Residential	0.11	2.24	0.0472	0.000	< 0.0001	0.76	< 0.0001

*Average Squared Canonical Correlation.

TABLE 3. Number of Observations and Percent Classified Into Cluster.

Cluster	Quantity	1	2	3	4	5	6	Total
1	Number	32	0	0	0	0	0	32
	Percentage	100	0	0	0	0	0	100
2	Number	0	19	0	0	0	0	19
	Percentage	0	100	0	0	0	0	100
3	Number	1	1	28	0	0	1	31
	Percentage	3.23	3.23	90.32	0	0	3.23	100
4	Number	0	0	0	6	0	0	6
	Percentage	0	0	0	100	0	0	100
5	Number	0	0	0	0	10	0	10
	Percentage	0	0	0	0	100	0	100
6	Number	0	0	0	0	0	12	12
	Percentage	0	0	0	0	0	100	100
Total	Number	33	20	28	6	10	13	110
	Percentage	30	18.18	25.45	5.45	9.09	11.82	100
Priors		0.17	0.17	0.17	0.17	0.17	0.17	
Error Rate		0	0	0.10	0	0	0	0.02

Cluster Analysis With Discriminating Variables

A CA using Ward's minimum variance method was performed with the factors obtained during the FA/PCA after selecting the variables retained by DA. The number of factors to be retained by the FA/PCA procedure was set to six. The number of clusters to be formed was also selected to be six. Once the six clusters were formed, the means of different watershed parameters were statistically compared among the clusters using Duncan's multiple range test with α =

0.05. The results of the statistical test are given in Table 4 (column labeled 'Duncan Results'). In Table 4, the values within the same parenthesis show the cluster numbers with means not significantly different from each other. The clusters shown in different parentheses are significantly different from each other if they do not appear together in any of the other parentheses. For example, based on the Duncan test, the mean of the percent forestland for Clusters 3 and 2 were significantly different from the means of all other clusters. At the same time, the mean of the percent forestland for Cluster 1 was not significantly different from that of Clusters 5 or 4, while it was different from Clusters 3, 2, and 6. The means of the percent forestland for Cluster 4 and Cluster 6 were not significantly different from each other, but the mean of the percent forestland for Cluster 6 was different from all other cluster means. Thus, based on Duncan's multiple range test, the mean of percent forestland was not significantly different among Clusters 1, 5, and 4 or between Clusters 4 and 6. The cluster number corresponding to the highest mean for a particular variable appears at the left-most position in the row, and means decrease from left to right.

The means of different watershed parameters were plotted graphically for comparison. Figures 3 through 5 show the mean plots for the important watershed parameters. Analyzing the results from the mean comparison tests and the graphical plots, some general conclusions were derived for different clusters and are given in Table 4. Table 4 explains the relative rankings of clusters when compared using individual variables. For example, consider the means of percent forestland. Cluster 3 had the highest mean compared to other clusters; the mean of Cluster 2 followed that; and the means of Clusters 1, 4, 5, and 6 were not significantly different from one another but were low compared to the means of Clusters 3 and 2.

Formation of Clusters

Based on the multivariate statistical analyses, six clusters of water bodies were formed. The locations of the water bodies falling under different clusters is shown in Figure 6. A brief summary of the water quality analysis results are given in Table 5.

The first cluster contains 39 impaired water bodies with relatively high densities of OSSFs. The term "relative" here and in the discussion of the clusters throughout indicates relative in comparison to the other clusters. The major land use within these watersheds is pastureland. This cluster of water bodies shows low bacterial loading from both farm animals and wildlife and relatively low public sewer use. Based on the statistical comparison of instream bacterial concentrations, three stream segments showed higher means during base flow periods, and 16 stream segments showed higher means during storm flow periods. Although there was no significant difference in the means for 19 of the stream segments, the means during both storm flow and base flow periods were higher than the water quality criteria value of 400 colony forming units (cfu)/100 ml. A few of the streams – stream segment 0805, for example – showed higher means during the storm flow period

TABLE 4. Comparison of Important Watershed Characteristics Among Clusters.

Variable/Cluster	**Duncan Results**	**1**	**2**	**3**	**4**	**5**	**7**
Frequency		**39**	**18**	**24**	**12**	**6**	**11**
Percent Forest	(3)(2)(1,5,4)(4,6)	Low	Medium	High	Low	Low	Low
Percent Cropland	(6)(4,1,2,3)(1,2,3,5)	Medium	Medium	Low	Medium	Low	High
Percent Urban Land	(5)(1,2)(3,6,4)	Medium	Medium	Low	Low	High	Low
Percent Pastureland	(6,1,4)(2,5,3)	High	Low	Low	High	Low	High
Population Density	(5)(2,1)(3,6,4)	Medium	Medium	Low	Low	High	Low
Density of Households	(5)(2,1)(1,3)(3,6,4)	Medium	Medium	Low	Low	High	Low
Density of OSSFs	(1,3,2,6,4,5)	High	Low	Medium	Low	None	Low
Density of Public Sewers	(5)(2,1)(3,6,4)	Medium	Medium	Low	Low	High	Low
Density of Other Septic	(3,1)(2,4,5,6)	High	Low	High	Low	Low	Low
Age of Households	(5)(6,4)(4,3,1)(3,1,2)	Medium	Low	Medium	High	No OSSFs	High
Loading From Deer	(4,2)(3,1,6,5)	Low	High	Low	High	Low	Low
Loading From Farm Animals	(4)(3,1)(1,5,2)(6)	Low	Low	Medium	High	Low	Low

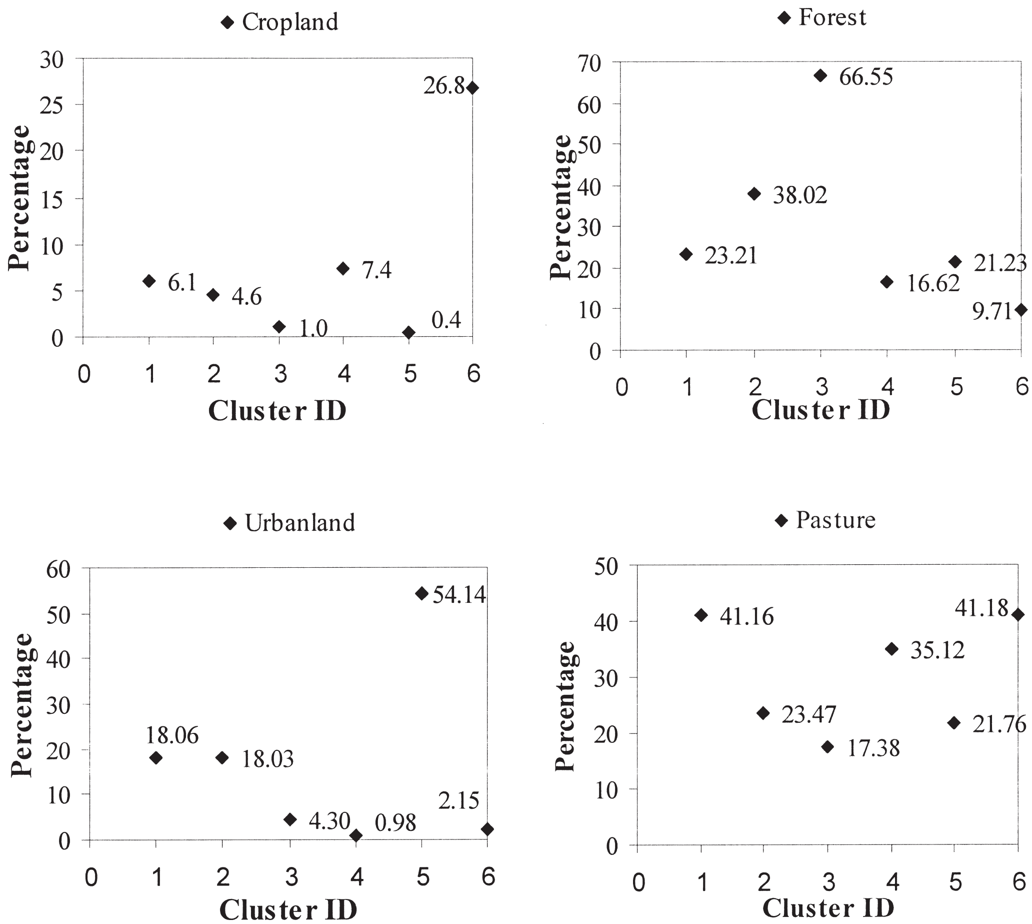

Figure 3. Mean Percentages of Different Land Uses Within Each Cluster.

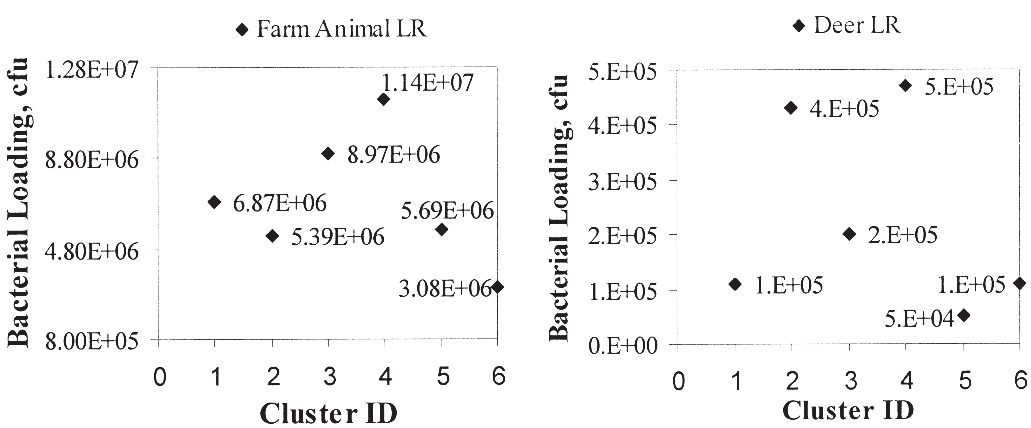

Figure 4. Mean Bacterial Loading Rates From Different Animals for Each Cluster.

but also showed high mean bacterial concentrations during the base flow period. This may indicate a relatively high contribution of bacteria from nonpoint sources, together with a noticeable contribution from point sources. No known facility was permitted to discharge fecal coliform bacteria into stream segments. However, there is possibility of accidental discharge from wastewater treatment plants (WWTPs). Based on the available information, except for three stream segments the total discharge capacity of WWTPs was relatively low or negligible. This fact, along with a high mean concentration during base flow periods,

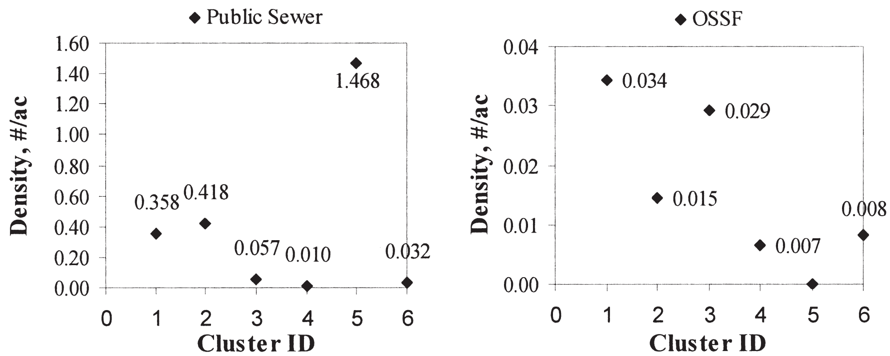

Figure 5. Mean Densities of Households Under Public Sewer Systems and OSSFs for Each Cluster.

can be interpreted as an indicator of a high background bacterial concentration in the streams, chronic failure of OSSF systems close to the streams, or a high density of animals within accessible distance of the water body. For one of the stream segments, 1108, the data were not adequate to test the difference in means between the two flow conditions.

Cluster 2 is a group of 18 impaired water bodies with well mixed land use distribution within the watersheds. The contribution of bacteria loading from wildlife is high compared to other clusters. These water bodies had low bacteria loading from farm animals and relatively low densities of OSSFs and public sewers. One of the water bodies showed a higher mean concentration of bacteria during base flow periods than during storm flow periods, and six water bodies showed higher mean bacteria concentrations during storm flow periods. Although stream segments such as 1427, 1428C, 1903, and 1906 showed no significant difference between the means during storm flow and base flow periods, they did show a reasonably high mean concentration during both storm flow and base flow periods. Overall, the water bodies in Cluster 2 showed lower mean bacterial concentrations during both storm flow and base flow periods than the other clusters. Higher concentrations during storm flow periods may be attributed to the high contribution of bacterial load from the wildlife in this cluster.

Cluster 3 contains 24 impaired water bodies with a high density of OSSFs and high bacterial loading from farm animals. The major land use within these watersheds is forestland. This cluster of water bodies had low bacterial loading from wildlife and relatively low public sewer use. One of the water bodies had a higher mean concentration of bacteria during base flow periods compared to storm flow periods, and 12 water bodies had higher mean concentrations during storm flow periods. The higher concentrations during storm flow periods may be attributed to the high contribution of bacteria from failed OSSFs and from farm animals. The data were not adequate to test for a significant difference of the mean concentrations between the two flow conditions for three of the water bodies. Seven water bodies had no significant difference between the means during storm flow and base flow periods and showed relatively low mean concentrations during both periods.

Twelve impaired water bodies fall into Cluster 4. These water bodies had high bacterial loading from both farm animals and wildlife. The main land uses in these watersheds are pastureland and cropland. The densities of OSSFs and public sewers are low compared to water bodies in the other clusters. Two stream segments showed higher mean concentrations during base flow periods, and seven stream segments showed higher mean concentrations during storm flow periods. Although stream segment 1255 showed relatively higher mean concentrations during storm flow periods, it also exhibited high mean concentrations during base flow periods. This may indicate the presence of farm animals within accessible reach of the water body. The total discharge capacity of WWTPs was relatively low or negligible for the water bodies in Cluster 4. One of the stream segments had insufficient data to test for a significant difference of mean bacterial concentrations between the two flow conditions.

Cluster 5 consists of six impaired water bodies in highly urbanized watersheds. The density of households connected to public sewers is the highest in this cluster compared to the other clusters. One characteristic of this cluster that separates it from other clusters is the absence of OSSFs within the watersheds. These watersheds lie completely within major urban areas. Three of the stream segments (1016, 1013, and 1113A) showed very high bacteria concentrations during both storm flow and base flow periods. Two stream segments had significantly higher mean instream fecal coliform concentrations during base flow periods. However, all the stream segments in this cluster had relatively high mean concentrations during base flow periods. Since the watersheds of these

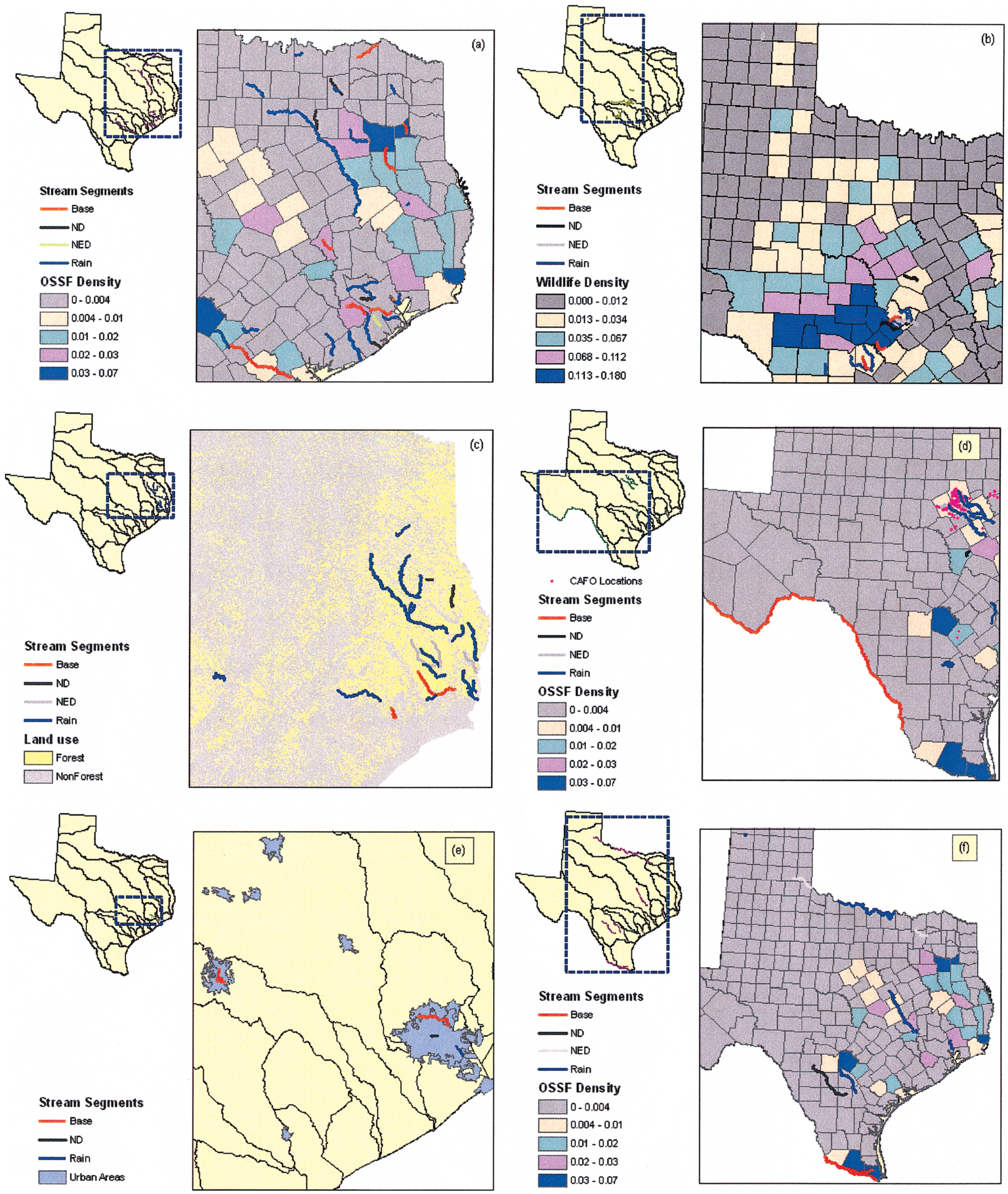

Figure 6. Locations of Stream Segments Belonging to Different Clusters
(Figures 6a through 6f correspond to Clusters 1 through 6).

stream segments fall completely within major urban areas of Texas, the higher concentrations of bacteria during base flow periods can indicate point sources from public sewage systems. These watersheds also may have been influenced by the presence of domestic pets.

Cluster 6 contains 11 impaired water bodies with low contributions of bacteria from any source. The watersheds are predominantly pastureland and cropland. Four stream segments showed reasonably high mean bacteria concentrations during both storm flow

TABLE 5. Summary of Water Quality Analysis Results.

Cluster ID	Number of Stream Segments				Not Enough Data
	Total	B_{conc} >> S_{conc}	S_{conc} >> B_{conc}	$B_{conc} \approx S_{conc}$	
1	39	3	16	19	1
2	18	1	6	7	4
3	24	1	13	7	3
4	12	2	7	2	1
5	6	2	0	4	0
6	11	1	3	6	2

Notes: B_{conc} means concentration during base flow period, S_{conc} means concentration during storm events, >> means significantly higher, and ≈ means not significantly different.

and base flow periods, though the contribution of bacteria from any source is not evident.

It can be assumed that the stream segments in a given cluster had similar hydrologic characteristics such that a hydrologic water quality model could be developed that would represent all watersheds in the cluster. To test this hypothesis, the impaired stream segments within five river basins – the Brazos, Neches, Nueces-Rio Grande Coastal, Sabine, and San Antonio – were selected based on their association with expected point and nonpoint sources. Using the same multivariate statistical techniques presented in this study, the water bodies were clustered into five groups. To test the possibility of applying the same model for a group of watersheds, two watersheds were selected from each of two clusters formed during the multivariate statistical analysis. The Hydrological Simulation Program-FORTRAN (HSPF) model was calibrated for one watershed within each group and validated for the other watershed in the same group. The study showed considerable similarities in the optimal parameter sets due to the similarities in watershed characteristics.

CONCLUSIONS

The Texas water bodies listed for bacterial water quality violations under CWA §303(d) were clustered based on their watershed characteristics. A pooled t-test was used to test for a significant difference in base flow and storm flow bacterial concentrations. The results of the t-test were used to help identify the potential sources of bacterial pollution in each watershed. The impaired water bodies were grouped into six homogeneous clusters based on their watershed characteristics using the multivariate statistical techniques of factor analysis/principal component analysis, cluster analysis, and discriminant analysis. The conclusions derived from the current study are summarized below.

The primary watershed characteristics that differentiate the clusters are bacterial contribution from farm animals and wildlife, density of OSSFs, density of households connected to public sewers, and land use distribution. The presence of point and nonpoint sources within the watershed boundaries was apparent for many watersheds regardless of their membership in a particular cluster. Higher concentrations during both storm flow and base flow periods may indicate both point and nonpoint sources of bacteria. However, in a few of these water bodies there were no known point sources, and the density of livestock and/or wildlife was high in the contributing watersheds. It may be appropriate to conclude that the livestock have direct access to many of the bacterially impaired stream segments.

A few of the watersheds were found to share a border with other states or with Mexico. The data collection was done only within the boundaries of the State of Texas. This may have left some potential sources of bacteria out of the analysis. Hence, the effect of potential sources across the boundaries on the instream water quality in these watersheds should be studied further.

The sample size of bacterial observations for 11 of the stream segments was not enough to carry out the statistical comparison between storm flow and base flow conditions. Also, the currently available information on domestic pets and migratory birds was insufficient to be incorporated into the multivariate analysis.

The use of GIS assisted in disaggregating the data available at county level or state level into a watershed level for analysis. However, an extensive data collection at the watershed level would greatly improve the results. The incorporation of location specific knowledge regarding application of manure on the land surface would increase the accuracy of the results. Similarly, quantifying the failing septic systems on a watershed basis would improve the results.

The study showed that the watersheds within a given cluster yielded similar model results using the same model input parameters. Therefore, the level of effort required to calibrate the model for a different watershed in the same cluster may be reduced.

ACKNOWLEDGMENTS

This project was financed through grants from the U.S. Environmental Protection Agency through the Texas Commission on Environmental Quality. The authors would like to acknowledge the many TCEQ personnel for their timely help in obtaining the datasets.

LITERATURE CITED

Alberto, W.D., M.L. 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. A Case Study: Suquía River Basin (Córdoba-Argentina). Water Res. 35(12):2881-2894.

Box, G.E.P. and D.R. Cox, 1964. An Analysis of Transformations. J. Royal Stat. Soc. 26:211-253.

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.

Cattell, R.B., 1966. The Scree Test for the Number of Factors. Multivariate Behavioral Res.1:245-276.

Choi, S., W. Chu, J. Brown, S.J. Becker, V.J. Harwood, and S.C. Jiang, 2003. Application of *Enterococci* Antibiotic Resistance Patterns for Contamination Source Identification at Huntington Beach, California. Marine Pollution Bulletin.

Clucas, S.R., 1997. Construction as a Curriculum Organizer for Technology Education. PhD. Dissertation, Virginia Polytechnic Institute and State University, Department of Teaching and Learning, Blacksburg, Virginia.

ESRI (Environmental Systems Research Institute), 1999. Getting to Know ArcView GIS: The Geographic Information System (GIS) for Everyone (Third Edition). Environmental Systems Research Institute, Redlands, California.

Haan, C.T., 2002. Statistical Methods in Hydrology (Second Edition). Iowa State University Press, Ames, Iowa.

Helena, B., R. Pardo, M. Vega, E. Barrado, J.M. Fernandez, and L. Fernandez, 2000. Temporal Evolution of Groundwater Composition in an Alluvial Aquifer (Pisuerga River, Spain) by Principal Component Analysis. Water Res. 34(3):807-816.

Jenerette, G.D., J. Lee, D.W. Waller, and R.E. Carlson, 2002. Multivariate Analysis of the Ecoregion Delineation for Aquatic Systems. Env. Mgmt. 29(1):67-75.

Johnson, R.A. and D.W. Wichern, 2002. Applied Multivariate Statistical Analysis (Fifth Edition). Prentice Hall International, Upper Saddle River, New Jersey.

Kaiser, H.F., 1960. The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement 20:141-151.

Kalkstein, L.S., G. Tan, and J.A. Skindlov, 1987. An Evaluation of Three Clustering Procedures for Use in Synoptic Climatological Classification. J. Climate and Applied Meteorology 26:717-730.

Karson, M.J., 1982. Multivariate Statistical Methods. Iowa State University Press, Ames, Iowa.

Milton, J.S. and J.C. Arnold, 1995. Introduction of Probability and Statistics. Principles and Application for Engineering and the Computing Sciences (Third Edition). McGraw-Hill, Inc., New York, New York.

Nathan, R.J. and T.A. McMahon, 1990. Identification of Homogeneous Regions for the Purposes of Regionalization. J. Hydrology 121(1990):217-238.

NCDC (National Climatic Data Center), 2002. NCDC: Locate Weather Observation Station Record. National Climatic Data Center, Asheville, North Carolina. *Available at* http://lwf.ncdc.noaa.gov/oa/climate/ stationlocator.html. *Accessed on* January 9, 2002.

Parveen, S., K.M. Portier, K. Robinson, L. Edmiston, and M.L. Tamplin, 1999. Discriminant Analysis of Ribotype Profiles of *Escherichia coli* for Differentiating Human and Nonhuman Sources of Fecal Pollution. Appl. Environ. Microbiol. 65:3142-3147.

Paul, S. 2003. Bacterial Total Maximum Daily Load (TMDL): Development and Evaluation of a New Classification Scheme for Impaired Water Bodies of Texas. PhD. Dissertation, Texas A&M University, Department of Biological and Agricultural Engineering, College Station, Texas.

Paul, S., M. Matlock, P.K. Haan, S. Mukhtar, and S. Pillai. 2002. Uncertainty Analysis as a First Step of Developing a Risk-Based Approach to Nonprofit Source Modeling of Fecal Coliform Pollution for Total Maximum Daily Load Estimates. Technical Report No. 192, Texas Water Resources Institute, Texas A&M University, College Station, Texas.

SAS (SAS Institute Inc.), 1999. SAS/STAT User's Guide. SAS Institute, Inc., Cary, North Carolina, Version 8, Volumes 1-3.

Srivastava, M.S. and E.M. Carter, 1983. An Introduction to Applied Multivariate Statistics. Elsevier Science Publishing Co., New York, New York.

TNRCC (Texas Natural Resources Conservation Commission), 2000. Texas 2000 Clean Water Act Section 303(d) List. Texas Commission on Environmental Quality, Austin, Texas. *Available at* http://www.tnrcc.state.tx.us/water/quality/00-303 dlist.pdf. *Accessed on* September 18, 2001.

TNRIS (Texas Natural Resources Information System), 2002. Digital Elevation Models (Dems) in 30m Resolution. Texas Natural Resources Information System, Austin, Texas. *Available at* http://www.tnris.state.tx.us/DigitalData/DEMs/dems.htm. *Accessed on* April 10, 2002.

U.S. Census Bureau, 1990. 1990 Census. U.S. Census Bureau, Washington, D.C. *Available at* http://www.census.gov/main/ www/cen1990.html. *Accessed on* October 30, 2000.

U.S. Census Bureau, 2000. Census 2000 TIGER/Line® Files. U.S. Census Bureau, Washington, D.C. *Available at* http://www.census.gov/geo/www/tiger/tiger2k/tgr2000.html. *Accessed on* October 30, 2000.

USDA (U.S. Department of Agriculture), 2002. USDA-NASS Published Estimates Data Base: 2000-2001. National Agricultural Statistics Database. USDA National Agricultural Statistics Service, Washington, D.C. *Available at* http://www.nass.usda.gov:81/ipedb/. *Accessed on* February 25, 2002.

USEPA (U.S. Environmental Protection Agency), 1996. TMDL Development Cost Estimated: Case Studies of 14 TMDLs. USEPA 841/R/96/001, Office of Water, U.S. Environmental Protection Agency, Washington, D.C.

USEPA (U.S. Environmental Protection Agency), 1998. Clean Water Action Plan. Washington D.C.: U.S. Environmental Protection Agency, Washington, D.C.. *Available at* http://www.epa.gov/history/topics/ cwa/03.htm. *Accessed on* December 12, 1999.

USEPA (U.S. Environmental Protection Agency), 2001a. Better Assessment Science Integrating Point and Nonpoint Sources, User's Manual for Release 3.0. EPA/823/B-01/001, Office of Water, U.S. Environmental Protection Agency, Washington, D.C.

USEPA (U.S. Environmental Protection Agency), 2001b. Protocol for Developing Pathogen TMDLs (First Edition). EPA/841/R-00/002, Office of Water, U.S. Environmental Protection Agency, Washington, D.C.

USEPA (U.S. Environmental Protection Agency), 2002. Index of /waterscience/ftp/basins/. U.S. Environmental Protection Agency, Washington, D.C. *Available at* http://www.epa.gov/waterscience/ftp/basins/. *Accessed on* April 10, 2002.

USGS (U.S. Geological Survey), 2002a. National Hydrography Dataset. U.S. Geological Survey, Reston, Virginia. *Available at* http://nhd.usgs.gov/index.html. *Accessed on* March 12, 2002.

USGS (U.S. Geological Survey), 2002b. National Land Cover Characterization Project. U.S. Geological Survey, Reston, Virginia. *Available at* http://landcover.usgs.gov/ nationallandcover.asp. *Accessed on* April 10, 2002.

USGS (U.S. Geological Survey), 2002c. Daily Streamflow for the Nation. U.S. Geological Survey, Reston, Virginia. *Available at* http://nwis.waterdata.usgs.gov/usa/nwis/discharge/?state_cd=48& format=station_list&sort_key=site_no&group_key=NONE&sitefile_output_format=html_table&column_name=agency_cd&column_name=site_no&column_name=station_nm&column_name=lat_va&column_name=long_va&column_name=state_cd. *Accessed on* January 9, 2002.

Vega, M., R. Pardo, E. Barrado, and L. Debn, 1998. Assessment of Seasonal and Polluting Effects on the Quality of River Water by Exploratory Data Analysis. Water Res. 32(12):3581-3592.

Whitlock, J.E., D.T. Jones, and V.J. Harwood, 2002. Identification of the Sources of Fecal Coliforms in an Urban Watershed Using Antibiotic Resistance Analysis. Water Res. 36:4265-4274.

Wiggins, B.A., R.W. Andrews, R.A. Conway, C.L. Corr, E.J. Dobratz, D.P. Dougherty, J.R. Eppard, S.R. Knupp, M.C. Limjoco, J.M. Mettenburg, J.M. Rinehardt, J. Sonsino, R.L. Torrijos, and M.E. Zimmerman, 1999. Use of Antibiotic Resistance Analysis to Identify Nonpoint Sources of Fecal Pollution. Appl. Environ. Microbiol. 65:3483-3486.

Yung, Y.-K., C.K. Wong, K. Yau, and P.Y. Qiah, 2001. Long-Term Changes in Water Quality and Phytoplankton Characteristics in Port Shelter, Hong Kong, From 1988-1998. Marine Pollution Bulletin 42(10):981-992.