UVM Stormwater Project
Statistical Analysis of Watershed Variables

Julie Foley and Breck Bowden

October 28, 2005

INTRODUCTION

The Vermont Agency of Natural Resources (ANR) has contracted with the University of Vermont (UVM) to help develop a protocol that can be used to objectively identify targets for stormwater reduction and locations for priority permit actions relevant to the new stormwater rules enacted during 2004. This UVM project expands upon the Water Resources Board Docket of 2004 and the stormwater simulation project completed by TetraTech in 2005. The specific objectives completed by UVM shifted in response to changes in the nature and timing of deliverables from the TetraTech project. Clarifications and updates were transmitted on 10 December 2004, 2 March 2005, 28 April 2005, and 11 May 2005.

The purpose of this study was to determine if it is possible to identify which of the key input variables in the P-8 stormwater runoff model also seem to explain groupings between impaired and attainment watersheds in non-mountainous watersheds in Vermont. For the purposes of this study, impaired watersheds are those watersheds that have been identified by the state as having biotic characteristics that have been degraded by stormwater runoff. Attainment watersheds are watersheds, also identified by the state, that have been developed to some degree but currently attain the state's biocriteria standards. The P-8 stormwater model was selected by the ANR Stormwater Advisory Group Technical Sub-committee because it is process-based, requires input data that is relatively available, is widely used, generates reasonable results, and includes a snowmelt component. The P-8 model is not a sophisticated hydrologic model but is an appropriate model for the limited input and validation data that we currently have available in Vermont. The input variables to the P-8 model are watershed area, land use/land cover percentages, impervious cover, average slope and hydrological soil group. It is certainly possible to assemble a longer list of variables that might affect stormwater runoff in the field. However, these variables are not a part of the P-8 input and so have no way to affect the P-8 output. Logically, they should not be included in this analysis.

There are many statistical tests (and variations of these tests) that could be applied to this dataset. However, it is not appropriate to "try out" a number of different analyses to identify a test or set of tests that show a significant difference or trend. The tests selected for this study were chosen to explore particular characteristics of the data and were based on literature reviews and consultation with University of Vermont statisticians.

METHODS

The input variables to the P-8 model and the independent variables in the statistical analyses used in this study were:

Variable Name	Description
Area	Watershed area (acres)
Agri	% agricultural land use in watershed
Forest	% forest land use in watershed
Urban	% urban land use in watershed
Water	% water land use in watershed
Wetland	% wetland land use in watershed
Soil_A	% Hydrologic soil group A
Soil_B	% Hydrologic soil group B
Soil_C	% Hydrologic soil group C
Soil_D	% Hydrologic soil group D
IC	% Impervious cover
Slope	Average % slope of watershed

To devise flow targets for permitting we utilized the flow output from the P-8 model. This model output is described in separate reports from TetraTech, who parameterized and ran the model. Ten years of precipitation and temperature data where used to simulate 10 years of flow from each of the watersheds. The same input data were used for each watershed simulation; only the watershed characteristics changed. The 10-year flow record for each watershed was used to produce a flow duration curve (FDC) and from these flow duration curves we extracted one high-flow metric and one low-flow metric for each watershed. The high-flow metric was the $Q_{0.3}$ or the flow that is exceeded only 0.3% of the time. This is approximately the 1-day flow, which is a metric that is easily understood, relates to key channel-forming processes, and is a variable that stakeholders have agreed is important. The low-flow metric was the Q_{95} or the flow that is exceeded 95% of the time. This flow is indicative of baseflow conditions.

Systat® 11 software was used to conduct the statistical testing. This is a user-friendly program with a Windows® interface that can be run on any computer with a Microsoft Windows® based system.

Data Analysis and Transformation

The original expressions of all but one of the independent watershed variables were in percentages. These were re-expressed as decimal equivalents for processing. The original expression for Area was in acres and these values were not re-expressed. The Shapiro-Wilks test was used to test the normality of the independent watershed variables for the impaired and attainment watersheds, separately. All but two independent variables had non-normal distributions. The variables SoilD and Forest were normally distributed.

Because many of the independent variables were not normally distributed, we tested a number of transformations, including arc-sine square root, natural log, log 10 and reciprocal transformations, to normalize the data. The arcsine square root transformation provided the best result and could be used across most of the variables. Three variables (SoilA, SoilB and Slope) could not be transformed with any of the selected transformations and were left out of the remaining tests that employed transformed data. Descriptive statistics were computed separately for the un-transformed and transformed watershed variables.

For data sets with relatively small numbers of observations, such as the dataset in this study, normality is an important assumption but is difficult to assess. Furthermore, the clustering methods we used in this study are relatively robust to divergences from normality. If we used only the data that could be normalized, we could not use all of the variables available in the dataset. If we used all of the data, at least some of it could not be normalized. For this reason we carried out parallel analyses in which we used all of the untransformed data then repeated the same analyses with only the transformed data that could be normalized.

Correlations

Scatterplot Matrix (SPLOM) plots show correlations among all possible pairs of variables and between the impaired and attainment watersheds. Only the untransformed data was plotted, as trends in the the transformed data would be similar. Correlation matrices were produced for untransformed data for both the impaired and attainment watersheds.

Cluster Analysis

Cluster analysis is a method used to identify natural groupings in datasets. We used two common clustering approaches in this study. In situations where the number of clusters is known or suspected, *k-means* clustering can be used to sort the individual members of the entire population (cases) into *k* distinct groups, where *k* is specified by the analyst. In this study we assumed that there are two clusters (*k*=2): impaired and attainment watersheds. The objective of this analysis was to determine if the selected watershed population could be reliably sorted into two groups on the basis of the simple, measurable watershed characteristics. Successive removal of the lowest ranking variables was conducted to see if there was any influence on the resulting clusters.

Hierarchical clustering is used to sort cases into an unspecified number of clusters. This analysis searches for similarities among cases on the basis of shared characteristics among their independent variables. Thus, one cluster may contain cases that are similar on the basis of some variables, while a separate cluster will contain cases that are similar on the basis of other variables. Any number of clusters of ‘similar’ cases may be created and some clusters may ‘nest’ within other, higher-order clusters.

We conducted two hierarchical cluster analyses. The first hierarchical cluster analysis used the final set of variables obtained from the k-means cluster analysis. These represented the most influential variables and resulted in clustering based on the dominant watershed variables. The second hierarchical cluster analysis we did used the *lowest ranking* variables from the k-means clustering results to examine whether and how the watersheds clustered on the basis of variables that did not strongly distinguish impaired from attainment watersheds. Presumably these clusters represent ‘natural’ watershed groupings in the absence of variables that would be indicative of impairment. We reasoned that attainment watersheds in these groups should serve as appropriate targets for impaired watersheds in the same group. Area, Impervious Cover (IC), Urban, Forest and SoilD were the variable excluded from this analysis because they had the greatest influence on watershed discharge ($Q_{0.3}$ and Q_{95}).

There are a variety of algorithms that can be used to calculate similarities among cases in a cluster analysis. In all analyses in this study we used the *average linking* method, which clusters based on the average Euclidean distances between cases and clusters. We also assessed the *single linking* method but found that the clusters were less well defined and more difficult to interpret with confidence.

Principal Components Analysis

PCA is generally an exploratory tool used to transform correlated variables into a smaller number of uncorrelated variables or principal components. It was used to reduce the dimensionality of the raw data and explore whether there were a smaller set of key factors that explained a large portion of the variance in the data. This was performed on the whole data set (untransformed and transformed) excluding Area.

The correlation-based analysis was used in this study because it is the most appropriate when variables are of different scales or their variances are differ greatly. We set a minimum eigenvalue of 1 and used a two- or three-factor correlation matrix for extraction. Review of the data with more than three factors was conducted, but did not provide additional insight into the data. A varimax rotation was applied to all the PCA’s. Normality is not a requirement of this test, but it does improve the analysis.

Two-Sample t-Test

T-tests (normal comparisons) or Mann-Whitney U tests (non-normal comparisons) were used to test for differences in the watershed variables between impaired and attainment watersheds. Variables that passed the normality test were subjected to an unpaired two sample *t*-test. The null hypothesis of this test was that the means of the two groups were the same. The minimum criterion for rejecting the null hypothesis was P=0.05. However, more stringent probabilities (P=0.01 and P=0.001) were noted where appropriate.

If the data failed the test for normality they could not be analyzed appropriately by a *t*-test. In these cases, the nonparametric Mann-Whitney Rank Sum test was used. The null hypothesis of this test is that the *medians* of the two groups are the same. The data are ranked from low to high, regardless of what group (impaired or attainment) it came from. The ranks were summed for each group and these sums (rank sums) were compared. Identical criteria for significance were used for the Mann-Whitney and the *t*-test results.

Kruskal-Wallis MANOVA

To completely circumvent the need to make assumptions about the underlying distribution of the independent input data, a Kruskal-Wallace MANOVA was performed on the untransformed data only. This is a weaker approach but provides robust information about which variables in the data set are most different between the impaired and attainment watersheds.

RESULTS

The Raw Data

The watershed variables used in this analysis (Table 1) were provided by ANR and were used by TetraTech as input to the P-8 model. There are 12 impaired and 15 attainment streams. Note that Allen Brook has both an attainment and an impaired reach. TetraTech calibrated the P-8 model with gauged streams and estimated flow values for the study streams based on this calibration. We extracted the $Q_{0.3}$ (the ~1-day flow or the flow exceeded only 0.3% of the time) as our standard metric of storm flow. We extracted the Q_{95} (or flow exceeded 95% of the time) as our metric of low (base) flow.

Table 1. The raw, untransformed watershed data used in this analysis. Area values are in acres. All other values are in decimal percent.

	Watershed	Status	Area	Water	Urban	Agri	Forest	Wetland	Soil_A	SoilB	Soil_C	Soil_D	IC	Slope
1	Alder_A	A	6571	0.06	0.23	0.29	0.40	0.02	0.09	0.19	0.30	0.41	0.06	0.07
2	Allen	A	2475	0.02	0.15	0.33	0.49	0.01	0.00	0.00	0.46	0.54	0.04	0.07
3	BumpSchool	A	670	0.06	0.00	0.02	0.92	0.00	0.00	0.00	0.13	0.88	0.00	0.13
4	Hubbardton	A	10825	0.14	0.04	0.12	0.69	0.01	0.04	0.00	0.19	0.62	0.02	0.13
5	Laplatte	A	1651	0.03	0.10	0.26	0.59	0.01	0.19	0.06	0.06	0.69	0.03	0.12
6	LittleOtter	A	7368	0.04	0.06	0.38	0.45	0.03	0.27	0.16	0.19	0.31	0.02	0.09
7	Malletts	A	9318	0.04	0.13	0.21	0.58	0.02	0.15	0.04	0.23	0.57	0.04	0.10
8	MiltonPond	A	1515	0.06	0.06	0.13	0.74	0.00	0.07	0.00	0.07	0.87	0.03	0.16
9	Muddy Branch	A	8382	0.04	0.03	0.24	0.67	0.01	0.12	0.39	0.09	0.31	0.01	0.15
10	Rock	A	1225	0.01	0.02	0.06	0.86	0.04	0.00	0.00	0.15	0.85	0.01	0.16
11	SandHill	A	685	0.05	0.18	0.01	0.76	0.00	0.86	0.00	0.14	0.00	0.08	0.08
12	SheldonSpr	A	1886	0.03	0.07	0.13	0.76	0.01	0.09	0.09	0.27	0.55	0.03	0.11
13	Teney	A	2987	0.03	0.30	0.07	0.59	0.01	0.12	0.21	0.55	0.03	0.07	0.12
14	Willow	A	1478	0.05	0.02	0.10	0.83	0.00	0.00	0.00	0.19	0.81	0.01	0.14
15	Youngman	A	672	0.03	0.05	0.31	0.56	0.03	0.83	0.00	0.00	0.17	0.03	0.05
16	Allen_I	I	6635	0.03	0.26	0.33	0.35	0.01	0.07	0.07	0.34	0.51	0.07	0.07
17	Bartlett	I	736	0.06	0.62	0.21	0.10	0.01	0.25	0.25	0.25	0.25	0.17	0.06
18	Centennial	I	887	0.07	0.71	0.04	0.18	0.00	0.63	0.00	0.00	0.25	0.31	0.06
19	Englesby	I	605	0.04	0.96	0.00	0.00	0.00	0.17	0.17	0.17	0.33	0.27	0.05
20	Indian	I	4582	0.05	0.39	0.18	0.37	0.01	0.13	0.04	0.17	0.60	0.08	0.06
21	Moon	I	5546	0.04	0.49	0.01	0.44	0.01	0.14	0.11	0.53	0.16	0.13	0.13
22	Morehouse	I	263	0.07	0.88	0.01	0.04	0.00	0.25	0.50	0.00	0.00	0.32	0.06
23	Munroe	I	3492	0.06	0.29	0.39	0.25	0.01	0.00	0.14	0.16	0.68	0.04	0.06
24	Potash	I	4561	0.05	0.53	0.30	0.11	0.01	0.28	0.10	0.14	0.48	0.22	0.05
25	Rugg	I	1831	0.05	0.20	0.37	0.38	0.01	0.14	0.00	0.71	0.14	0.07	0.11
26	Stevens	I	2136	0.05	0.47	0.23	0.26	0.00	0.00	0.04	0.75	0.21	0.11	0.08
27	Sunderland	I	1320	0.08	0.76	0.04	0.11	0.01	0.86	0.14	0.00	0.00	0.11	0.06

Descriptive Statistics

Descriptive statistics for the raw, untransformed data are included in Tables 2 and 3. With the exception of Area (in acres), all of the values are presented in decimal percents. Table 3 contains descriptive statistics for the attainment and impaired watersheds separately.

Table 2. Descriptive statistics for the raw, untransformed data for all watersheds.

All Cases	AREA	WATER	URBAN	AGRI	FOREST	WETLAND	SOIL_A	SOIL_B	SOIL_C	SOIL_D	IC_FINAL	SLOPE	PER_CN
N of cases	27	27	27	27	27	27	27	27	27	27	27	27	27
Minimum	262.7	0.012	0.001	0.001	0.001	0	0	0	0	0	0	0.05	48
Maximum	10825	0.136	0.958	0.39	0.916	0.045	0.857	0.5	0.75	0.875	0.32	0.16	78
Range	10562.3	0.124	0.958	0.389	0.915	0.045	0.857	0.5	0.75	0.875	0.32	0.11	30
Sum	90301.5	1.359	7.986	4.786	12.478	0.29	5.735	2.697	6.244	11.194	2.38	2.53	1870
Median	1886	0.05	0.2	0.18	0.455	0.009	0.125	0.062	0.167	0.406	0.06	0.08	72
Mean	3344.5	0.05	0.296	0.177	0.462	0.011	0.212	0.1	0.231	0.415	0.088	0.094	69.259
95% CI Upper	4542.836	0.06	0.409	0.229	0.568	0.015	0.317	0.15	0.313	0.525	0.125	0.108	72.864
95% CI Lower	2146.164	0.041	0.183	0.125	0.356	0.007	0.108	0.05	0.15	0.304	0.052	0.079	65.654
Std. Error	582.982	0.005	0.055	0.025	0.052	0.002	0.051	0.024	0.04	0.054	0.018	0.007	1.754
Standard Dev	3029.261	0.023	0.285	0.131	0.268	0.01	0.264	0.126	0.206	0.28	0.092	0.036	9.113
Variance	9176423.3	0.001	0.081	0.017	0.072	0	0.07	0.016	0.042	0.078	0.008	0.001	83.046
C.V.	0.906	0.465	0.964	0.739	0.58	0.966	1.243	1.261	0.89	0.675	1.045	0.389	0.132
Skewness	1.083	1.762	0.977	0.123	-0.098	1.642	1.739	1.763	1.226	0.099	1.51	0.448	-1.27
SE Skewness	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448
Kurtosis	0.088	6.19	-0.12	-1.423	-1.067	3.312	1.987	3.333	0.969	-1.105	1.355	-1.214	0.484
SE Kurtosis	0.872	0.872	0.872	0.872	0.872	0.872	0.872	0.872	0.872	0.872	0.872	0.872	0.872
SW Statistic	0.849	0.863	0.87	0.92	0.962	0.846	0.723	0.787	0.868	0.946	0.792	0.895	0.807
SW P-Value	0.001	0.002	0.003	0.039	0.41	0.001	0	0	0.003	0.169	0	0.01	0

Table 3. Descriptive statistics for the raw, untransformed data grouped by attainment and impaired watersheds.

ATTAIN WS	AREA	WATER	URBAN	AGRI	FOREST	WETLAND	SOIL_A	SOIL_B	SOIL_C	SOIL_D	IC_FINAL	SLOPE
N of cases	15	15	15	15	15	15	15	15	15	15	15	15
Minimum	670	0.012	0.001	0.006	0.399	0	0	0	0	0	0	0.05
Maximum	10,825	0.136	0.299	0.378	0.916	0.045	0.857	0.393	0.545	0.875	0.08	0.16
Range	10,155	0.124	0.298	0.371	0.517	0.045	0.857	0.393	0.545	0.875	0.08	0.11
Median	1,886	0.043	0.061	0.135	0.674	0.008	0.091	0	0.187	0.545	0.03	0.12
Mean	3,847	0.047	0.096	0.178	0.659	0.014	0.189	0.076	0.202	0.506	0.032	0.112
95% CI Upper	5,824	0.063	0.144	0.244	0.743	0.021	0.342	0.14	0.284	0.668	0.045	0.131
95% CI Lower	1,870	0.031	0.048	0.112	0.575	0.007	0.035	0.012	0.12	0.343	0.019	0.093
Std. Error	922	0.007	0.022	0.031	0.039	0.003	0.072	0.03	0.038	0.076	0.006	0.009
Standard Dev	3,570	0.029	0.087	0.119	0.152	0.013	0.278	0.115	0.148	0.294	0.023	0.034
Variance	12,742,500	0.001	0.007	0.014	0.023	0	0.077	0.013	0.022	0.086	0.001	0.001
C.V.	0.928	0.613	0.903	0.668	0.231	0.922	1.473	1.512	0.731	0.581	0.72	0.308
Skewness(G1)	0.905	2.176	1.176	0.169	-0.019	1.164	2.05	1.759	1.123	-0.381	0.809	-0.243
SE Skewness	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
Kurtosis(G2)	-0.824	6.771	0.732	-1.293	-0.854	1.026	3.225	3.022	1.135	-0.947	0.048	-0.981
SE Kurtosis	1.121	1.121	1.121	1.121	1.121	1.121	1.121	1.121	1.121	1.121	1.121	1.121
SW Statistic	0.809	0.788	0.88	0.945	0.975	0.89	0.657	0.73	0.911	0.931	0.92	0.956
SW P-Value	0.005	0.003	0.047	0.455	0.928	0.068	0	0.001	0.142	0.281	0.192	0.621
IMPAIRED WS	AREA	WATER	URBAN	AGRI	FOREST	WETLAND	SOIL_A	SOIL_B	SOIL_C	SOIL_D	IC_FINAL	SLOPE
N of cases	12	12	12	12	12	12	12	12	12	12	12	12
Minimum	263	0.03	0.2	0.001	0.001	0	0	0	0	0	0.04	0.05
Maximum	6,635	0.08	0.958	0.39	0.442	0.015	0.857	0.5	0.75	0.676	0.32	0.13
Range	6,372	0.05	0.758	0.389	0.441	0.015	0.857	0.5	0.75	0.676	0.28	0.08
Median	1,983	0.052	0.506	0.195	0.215	0.01	0.155	0.103	0.167	0.25	0.12	0.06
Mean	2,716	0.055	0.546	0.176	0.216	0.007	0.242	0.129	0.268	0.301	0.158	0.071
95% CI Upper	4,093	0.064	0.701	0.272	0.309	0.01	0.404	0.217	0.436	0.443	0.221	0.087
95% CI Lower	1,339	0.045	0.39	0.081	0.123	0.004	0.081	0.042	0.1	0.159	0.096	0.055
Std. Error	626	0.004	0.071	0.043	0.042	0.001	0.073	0.04	0.076	0.064	0.028	0.007
Standard Dev	2,167	0.014	0.245	0.15	0.147	0.005	0.255	0.137	0.264	0.223	0.099	0.025
Variance	4,696,598	0	0.06	0.023	0.022	0	0.065	0.019	0.07	0.05	0.01	0.001
C.V.	0.798	0.263	0.449	0.853	0.679	0.712	1.051	1.06	0.986	0.742	0.622	0.348
Skewness(G1)	0.583	0.151	0.27	0.112	0.093	-0.471	1.66	1.971	0.922	0.322	0.636	1.761
SE Skewness	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637
Kurtosis(G2)	-1.096	-0.332	-0.965	-1.747	-1.402	-1.127	2.443	4.732	-0.329	-0.953	-1.107	2.385
SE Kurtosis	1.232	1.232	1.232	1.232	1.232	1.232	1.232	1.232	1.232	1.232	1.232	1.232
SW Statistic	0.902	0.977	0.962	0.877	0.939	0.822	0.804	0.805	0.858	0.943	0.891	0.733
SW P-Value	0.169	0.966	0.817	0.08	0.484	0.017	0.01	0.011	0.046	0.534	0.121	0.002

Normality of the Data

Many of the variables in the dataset exhibited significant skewness and/or kurtosis so that all but two variables (Forests and SoilD) failed the Shapiro-Wilk test for normality (see Table 2). Figure 1 shows histograms for the raw, untransformed watershed variables. All variables were transformed using several different common algorithms (arcsine square root for percentage data, natural log (ln) or log10 transformations for skewed data, and the reciprocal). In some cases more than one algorithm transformed the non-normal variables into normal variables. All but three variables (Soil A, Soil B and Slope) could be normalized by at least one of these standard transformations. These three variables were not used in any further analyses based on transformed data. The arcsine square root transform was found to normalize all the landuse variables, SoilC, SoilD and IC and so was used in all further testing with transformed variables. Area could not be transformed with the arcsine square root transform. However, as shown below, Area tended to obscure important differences in the data. Furthermore, the difference in magnitude between Area (100's to 1,000's) and all of the other variables (0 to 1) was large and might have unfairly influenced some of the statistical tests. Consequently, Area was removed as an independent variable in the final statistical tests reported here. Figure 2 shows histograms of the final transformed watershed variables.

Figure 1. Histograms of raw, untransformed watershed variables.

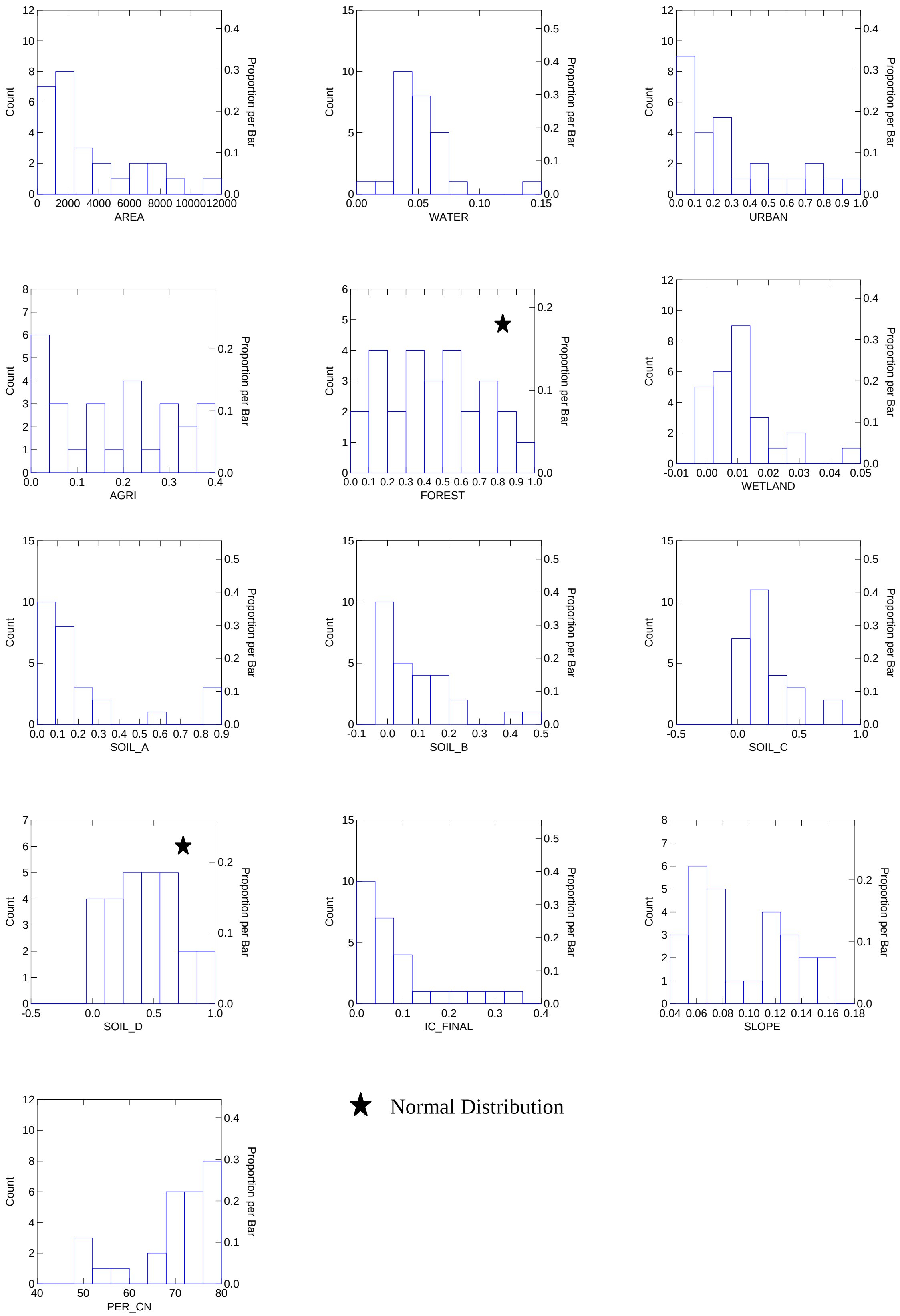
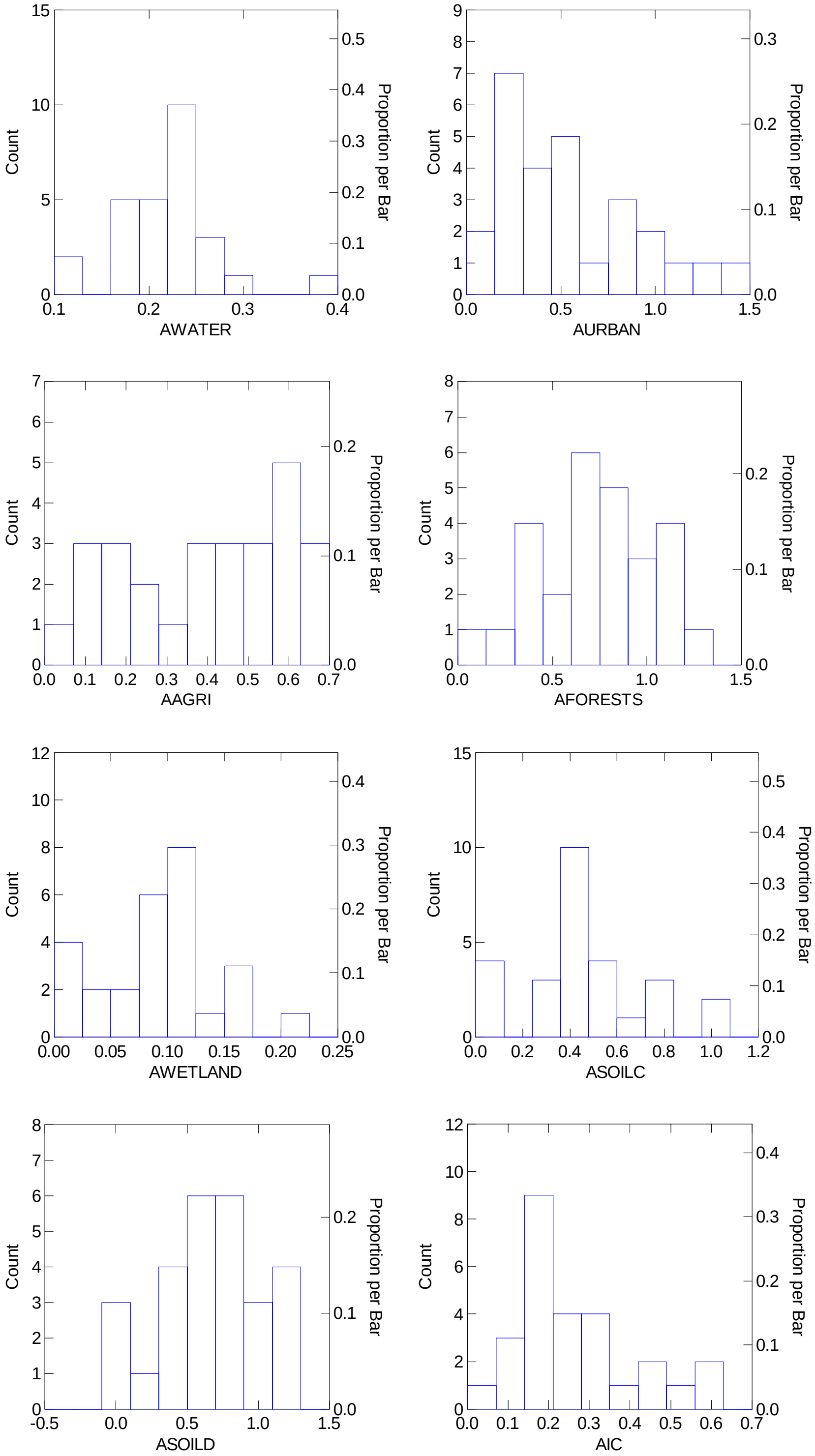


Figure 2. Histograms of the final arcsine square root transformed watershed variables. Three variables (Soil A, Soil B and Slope) could not be normalized with transformations. Area could not be transformed with the best fit arcsine square root transformation.



Correlations

Scatter plot matrices (SPLOM) were constructed to visualize the raw, untransformed data from the attainment and impaired watersheds (Figure 3). Tables 4 and 5 provide the Pearson correlation coefficients for these same data. There are clearly a number of strong internal correlations in the data. In the attainment watersheds, Urban positively correlated with SoilC and (as expected) with IC. Agri was negatively correlated with Forest. SoilA and SoilD were inversely correlated and IC tends to be lowest were SoilD is high. In the impaired watersheds, small watersheds tended to be urbanized and large watersheds tended to be forested. Oddly, Water was strongly associated with SoilA, the most permeable soil hydroclass. Urban areas were negatively correlated with Forest, Agri, and Wetland and positively correlated with IC, as should be expected. Similarly, IC was highly correlated with Urban, but negatively correlated with Forest, Agri, and Wetland. In both watersheds types, Forest was associated with higher Slope. In the attainment watersheds, Water, Urban, SoilA and IC were consistently low and appeared to be unaffected by the other watershed variables. Within the impaired watershed Water, Wetland, SoilA and Slope appeared to be unaffected by the other watershed variables. A number of the variables for both attained and impaired had outliers, though these can not be attributed to any particular watershed trend.

Figure 3. Scatterplot matrices (SPLOM) for Attainment and Impaired raw watershed variables

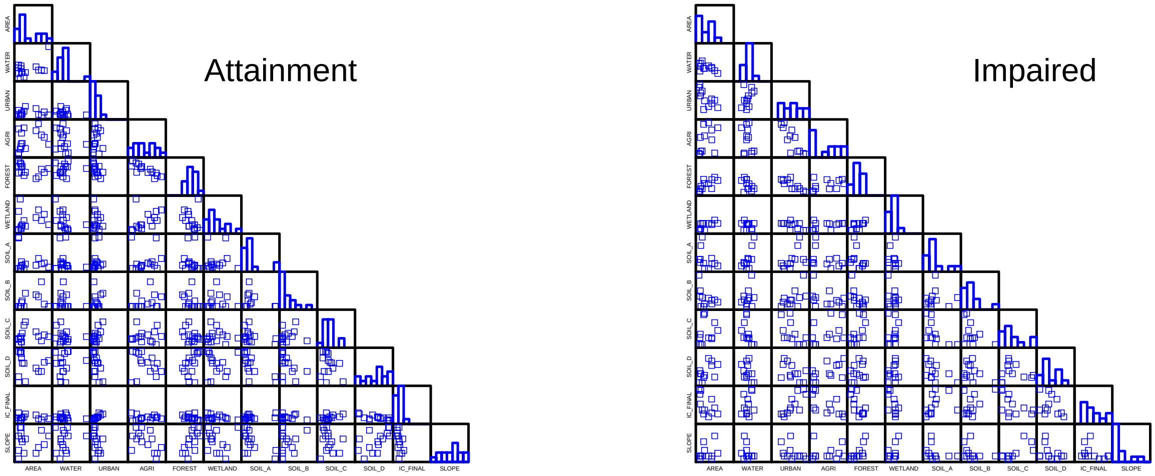


Table 4. Correlation matrix of attainment watershed variables. Correlations greater than ±0.6 are bold.

	Area	Water	Urban	Agri	Forest	Wetland	Soil_A	SoilB	Soil_C	Soil_D	IC	Slope
AREA	1.000											
WATER	0.511	1.000										
URBAN	0.031	-0.171	1.000									
AGRI	0.352	-0.238	0.057	1.000								
FOREST	-0.427	0.136	-0.561	**-0.839**	1.000							
WETLAND	0.196	-0.408	-0.155	0.409	-0.269	1.000						
SOIL_A	-0.242	-0.104	0.158	0.080	-0.150	0.056	1.000					
SOIL_B	0.446	-0.113	0.249	0.297	-0.362	0.052	-0.126	1.000				
SOIL_C	0.122	-0.167	**0.697**	-0.001	-0.352	-0.075	-0.364	0.192	1.000			
SOIL_D	-0.121	0.124	**-0.608**	-0.195	0.489	-0.028	**-0.686**	-0.438	-0.270	1.000		
IC_FINAL	-0.061	-0.100	**0.911**	-0.042	-0.442	-0.199	0.453	0.076	0.477	**-0.693**	1.000	
SLOPE	0.008	0.181	-0.449	-0.528	**0.662**	-0.144	-0.595	0.092	-0.204	0.598	-0.527	1.000

Table 5. Correlation matrix of impaired watershed variables. Correlations greater than ±0.6 are bold.

	Area	Water	Urban	Agri	Forest	Wetland	Soil_A	SoilB	Soil_C	Soil_D	IC	Slope
AREA	1.000											
WATER	**-0.615**	1.000										
URBAN	**-0.668**	0.427	1.000									
AGRI	0.418	-0.302	**-0.840**	1.000								
FOREST	**0.697**	-0.478	**-0.827**	0.394	1.000							
WETLAND	**0.635**	-0.190	**-0.604**	0.344	**0.628**	1.000						
SOIL_A	-0.413	**0.738**	0.515	-0.511	-0.410	-0.021	1.000					
SOIL_B	-0.414	0.320	0.574	-0.375	-0.597	-0.312	0.037	1.000				
SOIL_C	0.253	-0.565	-0.582	0.408	**0.624**	0.160	-0.580	-0.433	1.000			
SOIL_D	0.576	-0.427	-0.497	0.577	0.254	0.241	-0.499	-0.367	-0.090	1.000		
IC_FINAL	-0.558	0.327	**0.819**	**-0.658**	**-0.703**	**-0.665**	0.362	0.478	-0.511	-0.379	1.000	
SLOPE	0.313	-0.350	-0.447	0.038	**0.739**	0.465	-0.265	-0.284	0.740	-0.327	-0.373	1.000

k-Means Cluster Analysis

Untransformed Data

All untransformed watershed variables were included in the first cluster analysis (k=2). As illustrated in Figure 6, Area disproportionately influenced the clustering, reducing the influence of the remaining variables. The F-ratio for Area is 93.218, while the F-ratio for the next most influential variable is 3.447. An F test determines whether the ratio of two variance estimates is significantly greater than 1, which indicates the magnitude of the group variance from the total mean. In effect this analysis clustered the watersheds into ‘small’ and ‘large’ groups with a mix of attainment and impaired watersheds in each group. This was not a useful clustering of the watersheds.

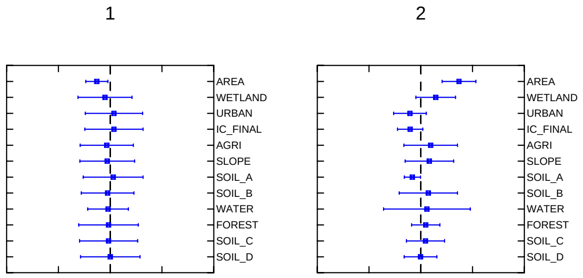


Figure 6. Cluster profile plot of untransformed watershed variables. Note the influence of the Area variable. The central vertical dotted line is the global mean in the data set. The ‘whiskers’ on each symbol indicate the variance in each variable, by cluster.

When Area was removed from the analysis, the influence of the remaining variables became apparent (Figure 7). In addition, clustering without Area produced a better separation of attainment and impaired watersheds. ICFinal was also removed as it is strongly auto-correlated with Urban.

Cluster 1 contained watersheds with slightly higher values of SoilD and Forest and lower values of Urban, IC_Final and SoilA. This cluster contained 15 cases, 12 of which were attainment watersheds. Cluster 2 contained watersheds with slight lower values of SoilD and Forest and higher values of Urban, IC_Final, and SoilA. This cluster contained 12 cases, 9 of which were impaired watersheds. SoilD, Urban and Forest were the most influential variables in this clustering.

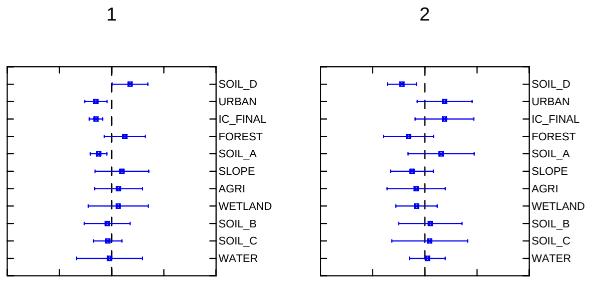


Figure 7. Cluster profile plot of untransformed watershed variables. Note the shift in variable influence when area is removed.

The k-means clustering was repeated 4 more times, each time removing the lowest ranking variable in the previous test (Water, SoilC, SoilB, Wetland). This resulted in a final clustering based on the most influential watershed variables (Figure 8). The final clustering was based on (in order of significance) SoilD, Urban, Forest, SoilA, Slope, Agri and Wetland. SoilD had an F-ratio of 45.054, twice that of Urban with a value of 22.731. Forest and SoilA have F-ratios of approximately 12. The remaining three variables had less disparate means and therefore, less influence on the clustering. The final watersheds included in each cluster are noted in Table 6.

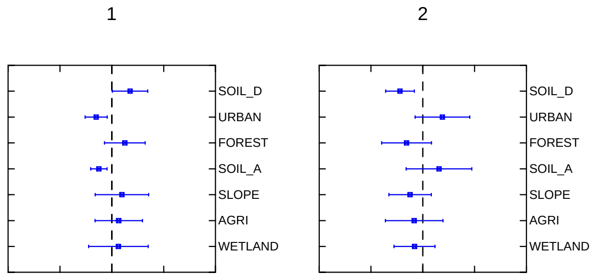


Figure 8. Final cluster profile plot of untransformed watershed variables.

Table 6. Table of k-means clustering results for the untransformed watershed variables.

	Cluster 1 - 15 Cases	
Case	Watershed	Status
1	Alder_A	A
2	Allen	A
3	BumpSchool	A
4	Hubbardton	A
5	Laplatte	A
6	LittleOtter	A
7	Malletts	A
8	MiltonPond	A
9	Muddy Branch	A
10	Rock	A
12	SheldonSpr	A
14	Willow	A
16	Allen_I	I
20	Indian	I
23	Munroe	I

	Cluster 2 - 12 Cases	
Case	Watershed	Status
11	SandHill	A
13	Teney	A
15	Youngman	A
17	Bartlett	I
18	Centennial	I
19	Englesby	I
21	Moon	I
22	Morehouse	I
24	Potash	I
25	Rugg	I
26	Stevens	I
27	Sunderland	I

Transformed Data

As shown in Figure 9, all transformed watershed variables were included in the first cluster analysis (k=2). For reasons described in previous sections, Area, SoilA and SoilB were not used in this analysis. AIC was also removed as it is autocorrelated with AUrban. Clustering based on the remaining seven variables resulted in Cluster 1 containing 16 watersheds (13 attainment and 3 impaired) and Cluster 2 containing 11 watersheds (9 impaired and 2 attainment). AUrban, ASoilD and AForest had the most influence on this clustering.

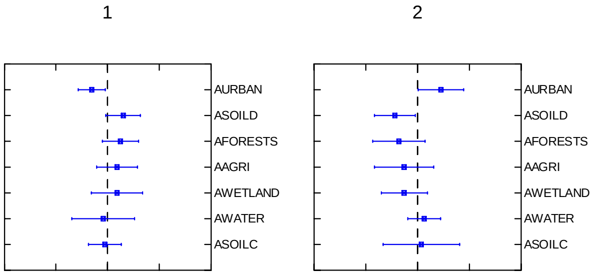


Figure 9. Cluster profile plot of transformed watershed variables.

The k-means clustering was repeated two more times, each time removing the lowest ranking variable in the previous test (ASoilC and AWater). This resulted in a final clustering based on the most influential watershed variables (Figure 10). The removal of ASoilC from the analysis resulted in a minor shift in the clustering such that Cluster 1 comprised 17 watersheds (4 impaired) and Cluster 2 comprised 10 watersheds (2 attainment).

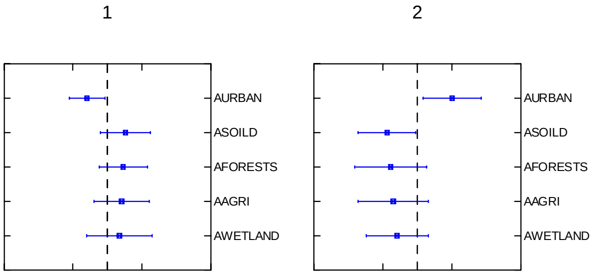


Figure 10. Final cluster profile plot of untransformed watershed variables.

The final clustering was based on (in order of significance) AUrban, ASoilD, AForest, AAgri and AWetland. AUrban had an F-ratio of 41.317, almost twice that of ASoilD which had a value of 22.607. The final watersheds included in each cluster are noted in Table 7.

Table 7. Table of k-means clustering results for the final transformed watershed variables.

	Cluster 1 - 17 Cases	
Case	Watershed	Status
1	Alder_A	A
2	Allen	A
3	BumpSchool	A
4	Hubbardton	A
5	Laplatte	A
6	LittleOtter	A
7	Malletts	A
8	MiltonPond	A
9	Muddy Branch	A
10	Rock	A
12	SheldonSpr	A
14	Willow	A
15	Youngman	A
16	Allen_I	I
20	Indian	I
23	Monroe	I
25	Rugg	I

	Cluster 2 - 10 Cases	
Case	Watershed	Status
11	SandHill	A
13	Teney	A
17	Bartlett	I
18	Centennial	I
19	Englesby	I
21	Moon	I
22	Morehouse	I
24	Potash	I
26	Stevens	I
27	Sunderland	I

Hierarchical Clustering

Untransformed Data

The final k-means clustering variables (SoilD, Urban, Forest, SoilA, Slope, Agri and Wetland) were used in the hierarchical cluster analysis. The resulting permuted data matrix is shown in Figure 10. This permuted data matrix is the hierarchical clustering of all the watershed variables (columns) crossed with all the watershed cases (rows). Distances between clusters are noted as different colors in the matrix. This provides an easy way to visualize how the different cases (watersheds) cluster and upon which variables (columns).

Figure 10 shows that SoilD and Forest form a distinct cluster. Slope and Wetland cluster together and Urban, SoilA and Agri cluster to these two at a greater distance. Nine attainment watersheds cluster out on the basis of SoilD and Forest as illustrated in the red cluster on the bottom right of Figure 10. Four impaired watersheds cluster loosely based on Urban and SoilA. Another less significant group of 10 mixed watersheds clusters out on the basis of Urban, Agri, SoilD and Wetland combined as illustrated in the light green cluster in Figure 10.

A second hierarchical cluster analysis was performed on the untransformed watershed dataset with Area and the next *most* influential variables (based on k-means clustering results) removed from the analysis (Figure 11). The resulting matrix identified five distinctive clusters that are primarily influenced by SoilA, Agri and SoilC.

Each of the five clusters identified in Figure 11 contained both impaired and attainment watersheds. The mean and standard deviation for the $Q_{0.3}$ flow and Q_{95} flow of the attainment watersheds in each cluster were calculated and are included in Table 8. The $Q_{0.3}$ attainment means were lower and the Q_{95} attainment means were higher than the calculated values for each impaired watershed in a given cluster. It should be noted that for Indian Brook the Q_{95} flow exceeds and the $Q_{0.3}$ flow is below the attainment average with the standard deviation taken into account.

Figure 10. Hierarchical cluster matrix using average linkage method. Input variables are the final cluster variables from the k-means clustering.

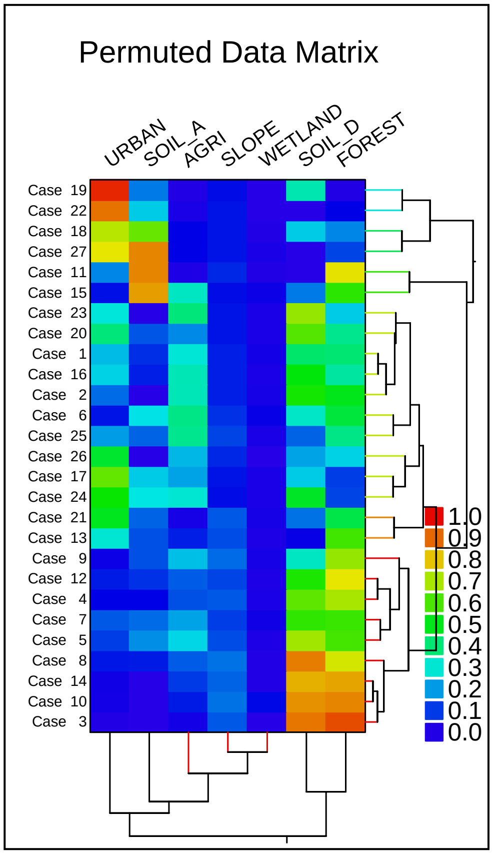


Figure 11. Hierarchical cluster matrix using average linkage method. Input variables are the lowest ranking variables from the k-means clustering.

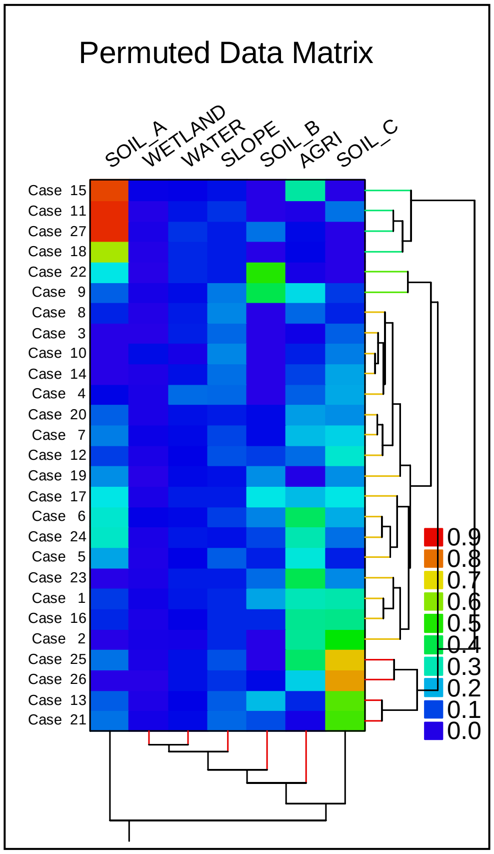


Table 8. Hierarchical clustering results for the raw, untransformed data. Mean attainment flow values for the Q 0.3% and Q 95% flow are identified for each cluster.

Cluster	Case #	Watershed	Status	Q 0.3%	Avg A Q 0.3%	Std Dev	Q0.3% + SD
1	18	Centennial	I	16.0399	7.9636	0.0849	8.0485
	27	Sunderland	I	8.2525			
	11	SandHill	A	8.0236			
	15	Youngman	A	7.9035			
2	22	Morehouse	I	16.8777	8.1448	--	-
	9	Muddy Branch	A	8.1448			
3	19	Englesby	I	15.4649	11.5276	1.1173	12.6449
	20	Indian	I	11.6373			
	3	BumpSchool	A	12.5317			
	4	Hubbardton	A	11.9623			
	7	Malletts	A	10.9241			
	8	MiltonPond	A	12.0885			
	10	Rock	A	11.9923			
	12	SheldonSpr	A	9.2432			
	14	Willow	A	11.9511			
4	17	Bartlett	I	11.3478	10.2719	1.7680	12.0399
	24	Potash	I	12.2374			
	5	Laplatte	A	11.5221			
	6	LittleOtter	A	9.0217			
5	16	Allen_I	I	11.7358	11.2695	0.0912	11.3607
	23	Munroe	I	12.0108			
	1	Alder	A	11.3340			
	2	Allen_A	A	11.2050			
6	21	Moon	I	9.9587	9.3369	--	-
	25	Rugg	I	11.3195			
	26	Stevens	I	11.9120			
	13	Teney	A	9.3369			

Cluster	Case #	Watershed	Status	Q 95%	Avg A Q 95%	Std Dev	Q95% - SD
1	18	Centennial	I	0.1875	0.2310	0.0035	0.2275
	27	Sunderland	I	0.2229			
	11	SandHill	A	0.2335			
	15	Youngman	A	0.2285			
2	22	Morehouse	I	0.1948	0.2176	--	-
	9	Muddy Branch	A	0.2176			
3	19	Englesby	I	0.1903	0.2116	0.0074	0.2042
	20	Indian	I	0.2108			
	3	BumpSchool	A	0.2100			
	4	Hubbardton	A	0.2116			
	7	Malletts	A	0.2177			
	8	MiltonPond	A	0.2027			
	10	Rock	A	0.2036			
	12	SheldonSpr	A	0.2239			
	14	Willow	A	0.2121			
4	17	Bartlett	I	0.2000	0.2190	0.0083	0.2107
	24	Potash	I	0.1964			
	5	Laplatte	A	0.2132			
	6	LittleOtter	A	0.2249			
5	16	Allen_I	I	0.2015	0.2206	0.0048	0.2158
	23	Munroe	I	0.2016			
	1	Alder	A	0.2240			
	2	Allen_A	A	0.2172			
6	21	Moon	I	0.2030	0.2399	--	-
	25	Rugg	I	0.2027			
	26	Stevens	I	0.1977			
	13	Teney	A	0.2399			

🟧	Q 95% flow exceeds or Q 0.3% flow is below the attainment average.
🟨	Q 95% flow exceeds or Q 0.3% flow is below the attainment average with standard deviation.

Transformed Data

The final k-means clustering variables (ASoilD, AUrban, AForest, AAgri and AWetland) were used in the hierarchical cluster analysis. The resulting permuted data matrix is shown in Figure 12. This permuted data matrix is the hierarchical clustering of all the watershed variables (columns) crossed with all the watershed cases (rows). Distances between clusters are noted different colors in the matrix.

Figure 12 Shows that ASoilD and AForest form a distinct cluster. AWetland, AAgri and AUrban cluster together more loosely, or at a greater distance. There are five resulting watershed clusters. Four attainment watersheds cluster out on the basis of ASoilD and AForest as illustrated in the red/orange/yellow cluster on the top right of Figure 12. The three middle clusters (mostly attainment) are driven primarily by AAgri and AUrban. The bottom ten cases (2 attainment and 8 impaired watersheds) cluster at the greatest distance based on varying influences from all the variables.

A second hierarchical cluster analysis was performed on the transformed watershed dataset with the most influential variables (based on k-means clustering results) *removed* from the analysis (Figure 13). AWater and AWetland cluster together and ASoilC and AAgri cluster together on the basis of watershed data. The low influence matrix identified six distinct clusters that are primarily influenced by AForests, ASoilC and AAgri. Two of these clusters (2 and 5) contain only attainment watersheds and the remainder comprise a mix of impaired and attainment watersheds (Figure 13). The mean and standard deviation for the $Q_{0.3}$ flow and Q_{95} flow of the attainment watersheds in each of the mixed clusters were calculated and are included in Table 9. The $Q_{0.3}$ attainment means were lower and the Q_{95} attainment means were higher than the calculated values for each impaired watershed in clusters three, four, five and six. The $Q_{0.3}$ attainment mean in cluster 1 is higher and the Q_{95} attainment mean is lower than the means for the impaired Sunderland watershed. It should be noted that for Indian Brook, the Q_{95} flow exceeds, and the $Q_{0.3}$ flow is below, the attainment average with the standard deviation taken into account. The $Q_{0.3}$ flow for Bartlett Brook behaves the same as for Indian Brook.

Figure 12. Hierarchical cluster matrix using average linkage method. Input variables are the final transformed cluster variables from the k-means clustering.

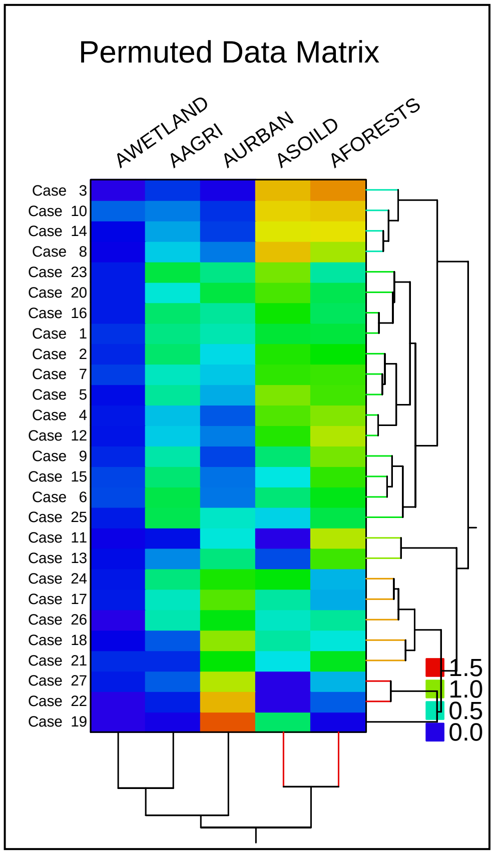


Figure 13. Hierarchical cluster matrix using average linkage method. Input variables are the lowest ranking transformed variables from the k-means clustering.

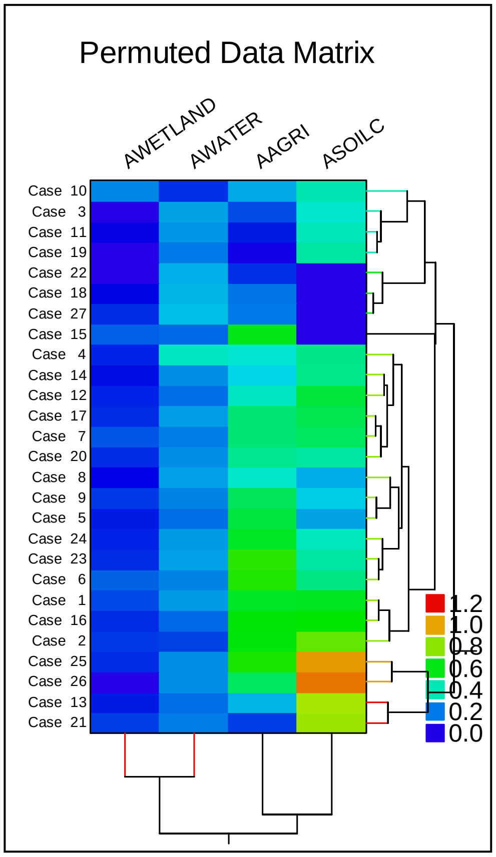


Table 9. Hierarchical clustering results for the transformed data. Mean attainment flow values for the Q 0.3% and Q 95% flow are identified for each cluster.

Cluster	Case #	Watershed	Status	Q 0.3%	Avg A Flow	Std Dev	Q0.3% + SD
1	18	Centennial	I	16.0399	9.3065	2.3268	11.6333
	19	Englesby	I	15.4649			
	22	Morehouse	I	16.8777			
	27	Sunderland	I	8.2525			
	3	BumpSchool	A	12.5317			
	10	Rock	A	11.9923			
	11	SandHill	A	8.0236			
2	15	Youngman	A	7.9035	NA	NA	-
3	17	Bartlett	I	11.3478	11.0202	1.2808	12.3009
	20	Indian	I	11.6373			
	4	Hubbardton	A	11.9623			
	7	Malletts	A	10.9241			
	12	SheldonSpr	A	9.2432			
	14	Willow	A	11.9511			
4	5	Laplatte	A	11.5221	NA	NA	-
	8	MiltonPond	A	12.0885			
	9	Muddy Branch	A	8.1448			
5	23	Munroe	I	12.0108	9.0217	--	-
	24	Potash	I	12.2374			
	6	LittleOtter	A	9.0217			
6	16	Allen_I	I	11.7358	11.2695	0.0912	11.3607
	1	Alder	A	11.3340			
	2	Allen_A	A	11.2050			
7	21	Moon	I	9.9587	9.3369	--	-
	25	Rugg	I	11.3195			
	26	Stevens	I	11.9120			
	13	Teney	A	9.3369			

Cluster	Case #	Watershed	Status	Q 95%	Avg A Q 95%	Std Dev	Q95% - SD
1	18	Centennial	I	0.1875	0.2157	0.0158	0.1999
	19	Englesby	I	0.1903			
	22	Morehouse	I	0.1948			
	27	Sunderland	I	0.2229			
	3	BumpSchool	A	0.2100			
	10	Rock	A	0.2036			
	11	SandHill	A	0.2335			
2	15	Youngman	A	0.2285	NA	NA	-
3	17	Bartlett	I	0.2000	0.2163	0.0057	0.2106
	20	Indian	I	0.2108			
	4	Hubbardton	A	0.2116			
	7	Malletts	A	0.2177			
	12	SheldonSpr	A	0.2239			
	14	Willow	A	0.2121			
4	5	Laplatte	A	0.2132	NA	NA	-
	8	MiltonPond	A	0.2027			
	9	Muddy Branch	A	0.2176			
5	23	Munroe	I	0.2016	0.2249	--	-
	24	Potash	I	0.1964			
	6	LittleOtter	A	0.2249			
6	16	Allen_I	I	0.2015	0.2206	0.0057	0.2148
	1	Alder	A	0.2240			
	2	Allen_A	A	0.2172			
7	21	Moon	I	0.2030	0.2399	--	-
	25	Rugg	I	0.2027			
	26	Stevens	I	0.1977			
	13	Teney	A	0.2399			

🟧 Q 95% flow exceeds or Q 0.3% flow is below the attainment average.
🟨 Q 95% flow exceeds or Q 0.3% flow is below the attainment average with standard deviation.

Principal Component Analysis

Principle components analysis (PCA) is used to explore if there are higher-order factors that might explain relationships between a complex array of individual variables. In other words, PCA can help reduce the 'dimensionality' of the data.

Untransformed Watershed Data

The first two factors of the PCA for the untransformed data explain 68% of the variance. Factor 1 separated out IC, Urban, SoilA and Agri from Slope, Forest, SoilD and Wetland. Factor 2 separated out IC, Urban, Soil and Slope from Agri, Wetland, SoilD and Forest. IC and Urban group together in one group. Slope, Forest and SoilD group together in a second group. Agri and Wetland group together in a third group (Figure 14). Including a third factor from the analysis help explain a further 21% of the variance in the data. The resulting factor plot (Figure 15) shows that adding this third factor separates SoilC from Agri and Wetlands, but otherwise retains the same three-group dimensionality of the two-factor analysis. The addition of the third factor does not appear to provide much important new information about the data and complicates the analysis.

It should be noted that Urban and IC are highly auto-correlated. Including both of these variables in the PCA therefore unfairly weights these variables relative to the others in the input data set. Removing one of these two variables (e.g., IC) slightly alters the rotation of the remaining PCA components. However, the composition, strength and orientation of the original three groups remained the same.

Figure 14. PCA factor loading plot for Factor 1 and Factor 2.

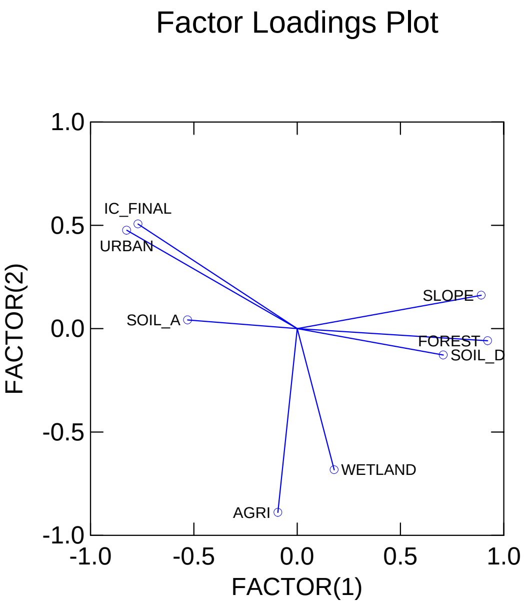
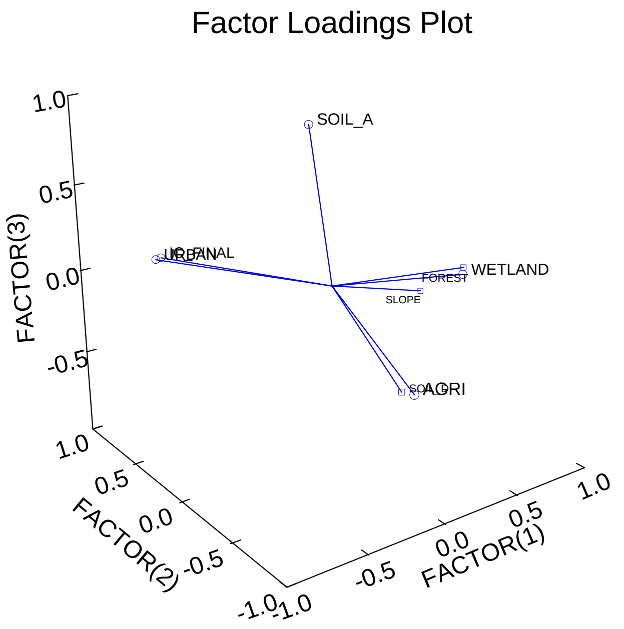


Figure 15. PCA factor loading plot for Factor 1, Factor 2 and Factor 3.



Transformed Data

All transformed watershed variables were included in the PCA analysis. The combined variance of the two factor analysis is 64%. Factor 1 separated out AIC, AUrban and AWater from AAgri, AWetland, ASoilD and AForests. Factor 2 separated out AIC, AUrban and AWater from ASoilD, AAgri, AWetland and ASoilD. ASoilC is plotted on the Factor 1 axis and AForests is plotted on the Factor 2 axis. AIC and AUrban have nearly identical loadings. AForest and ASoilD group together in a second group and AAgri and AWetland group together in a third group (Figure 16). Including a third factor in the analysis helps explain a further 12% of the variance in the data. The resulting factor plot (Figure 17) shows that adding this third factor separates ASoilC from AAgri and AWetlands, but otherwise retains the same three-group dimensionality of the two-factor analysis. The addition of the third factor does not appear to provide much important new information about the data and complicates the analysis.

It should be noted that AUrban and AIC are highly auto-correlated. Including both of these variables in the PCA therefore unfairly weights these variables relative to the others in the input data set. Removing one of these two variables (e.g., AIC) slightly alters the rotation of the remaining PCA components. However, the composition, strength and orientation of the original three groups remained the same.

Figure 16. PCA factor loading plot for Factor 1 and Factor 2 on transformed data.

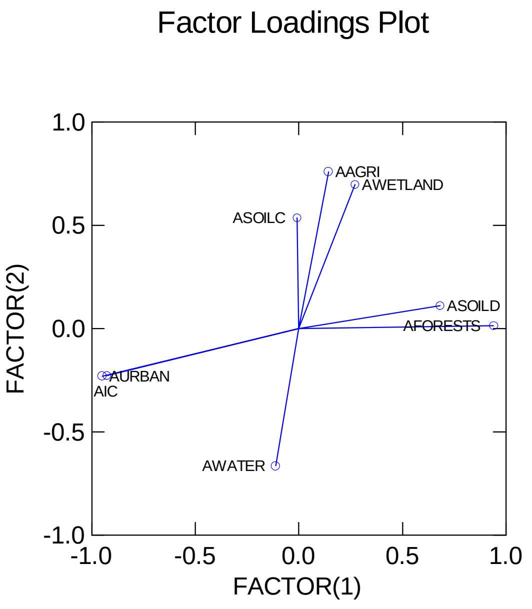
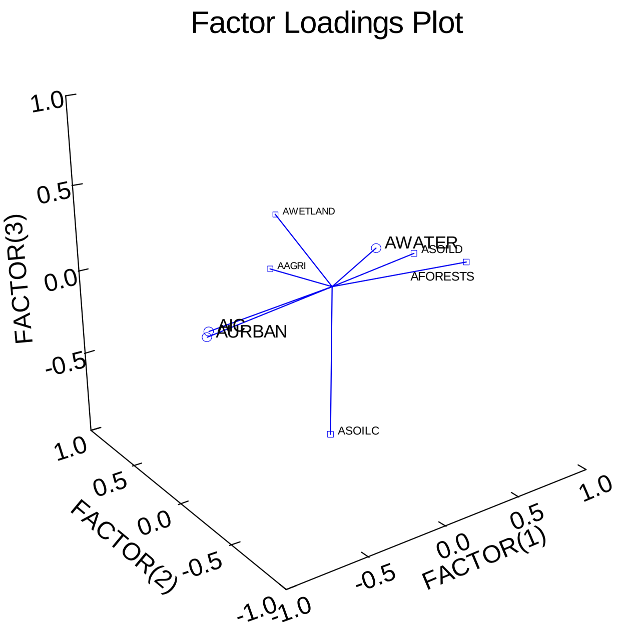


Figure 17. PCA factor loading plot for Factor 1, Factor 2 and Factor 3 on transformed data.



Kruskal-Wallis MANOVA

The KW test is non-parametric and so does not make any assumptions about the underlying distribution of the data. It is best suited to data such as the untransformed variables, which are known to be non-normal. When the characteristics of the data are unknown, non-parametric tests may be more appropriate to use. However, it should be recognized that non-parametric tests are inherently weaker than parametric tests if the data are in fact normally distributed. Furthermore, if the data are known to be non-normal (as here) non-parametric tests should not be viewed as a way to 'save' the data.

With these caveats in mind, the Kruskal-Wallis test identified statistically significant differences between impaired and attainment watersheds for the Urban, Forest, IC and Slope variables (Table 10). Several things should be noted about these variables. First, Urban and IC are clearly auto-correlated. In addition, the Urban and Forest variables are negatively correlated. Finally, Slope was one of the variables that could not be transformed.

Table 10. Results from the Kruskal-Wallis MANOVA and difference of means testing.

	Rank sums		M-W		Chi^2
Dependent Variable	**A**	**I**	**U**	**P(U)**	**1 df**
AREA	224.0	154.0	104.0	0.495	0.5
WATER	178.5	199.5	58.5	0.124	2.4
URBAN	124.0	254.0	4.0	0.000	17.6
AGRI	214.0	164.0	94.0	0.845	0.0
FOREST	299.0	79.0	179.0	0.000	18.9
WETLAND	230.5	147.5	110.5	0.315	1.0
SOIL_A	184.5	193.5	64.5	0.211	1.6
SOIL_B	179.0	199.0	59.0	0.121	2.4
SOIL_C	204.5	173.5	84.5	0.788	0.1
SOIL_D	250.0	128.0	130.0	0.051	3.8
IC	127.0	250.0	7.5	0.000	16.3
SLOPE	271.0	107.0	151.0	0.003	9.0

t-Test on Transformed Data

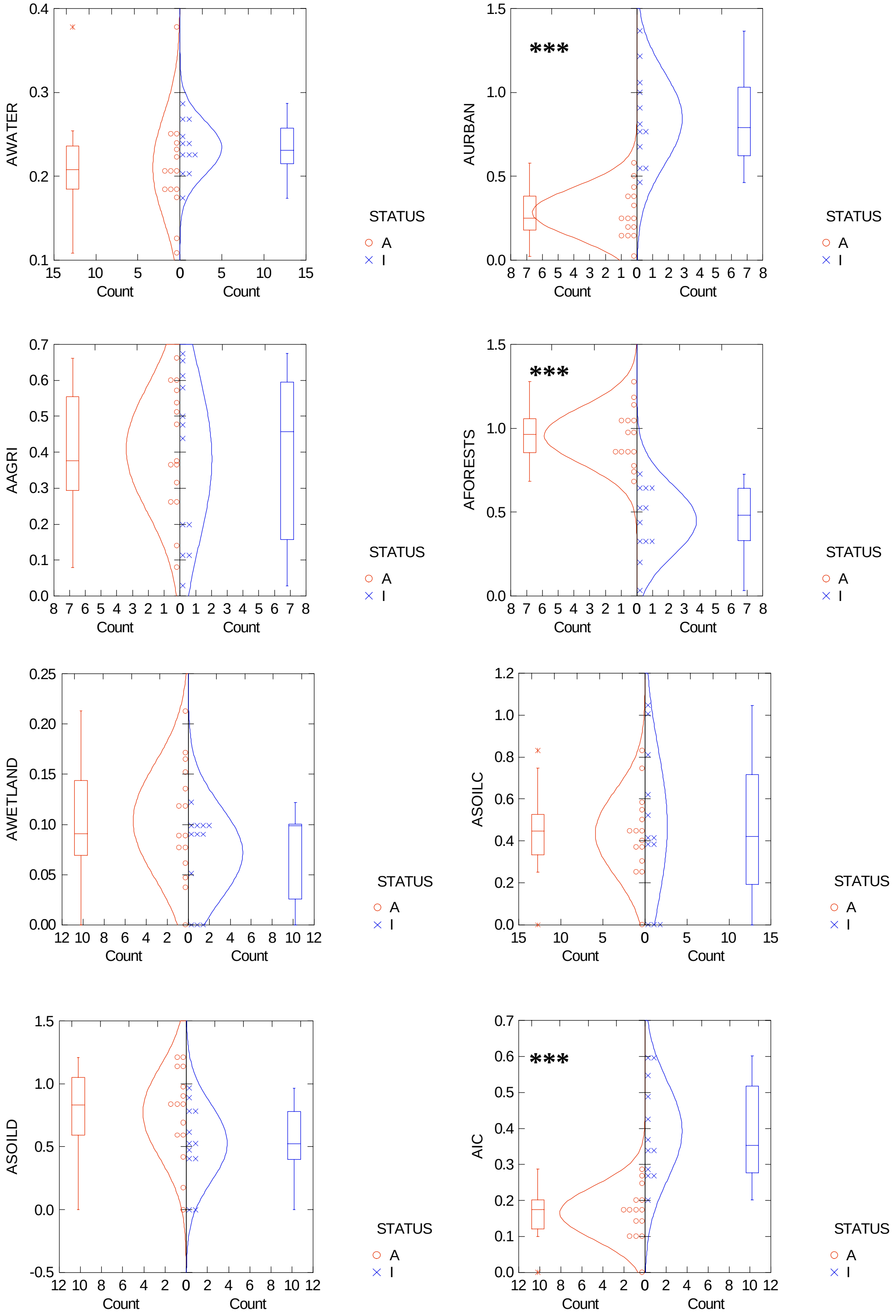
The AWater, AUrban, AAgri, AForests, AWetland, ASoilC, ASoilD and AIC watershed variables were analyzed with the *t*-test (Table 11). The difference in means for the AUrban, AForest and AIC variables was found to be extremely significant (P<0.001). The difference in means for the remaining variables was not significant. Figure 18 shows whisker plots and distribution curves of impaired and attainment groupings for all the transformed variables.

Table 11. Variable means by watershed status and t-Test results.

			t-Test		
Variable	**Attainment Mean**	**Impaired Mean**	**t**	**df**	**P**
AWATER	0.211	0.234	-1.178	25	NS
AURBAN	0.284	0.844	-6.706	25	***
AAGRI	0.409	0.382	0.336	25	NS
AFOREST	0.957	0.451	6.925	25	***
AWETLAND	0.104	0.072	1.559	25	NS
ASOIL_C	0.436	0.471	-0.3154	25	NS
ASOIL_D	0.771	0.529	1.827	25	NS
AIC_FINAL	0.166	0.393	-5.573	25	***

P Value		
Not Significant	>0.05	NS
Very Significant	<0.001	***

Figure 18. Plots of two-sample t-Test results for all transformed watershed variables. Plots labeled with "***" indicate a significant difference in means between the impaired and attainment watershed groupings. All others did not have significant P values.



DISCUSSION

k-Means Cluster Analysis

The final watershed variables included in the k-means two cluster analysis for both the raw, untransformed data and the transformed data resulted in very good separations of impaired and attainment watersheds. The most influential variables for both analyses were Urban, SoilD and Forest. There were consistently different means for these variables between the attainment and impaired watersheds, making these variables good indicators of watershed status. Sand Hill Brook, Teney Brook, and Youngman Brook clustered in the predominantly impaired Cluster 2 for both analyses. These may be good watersheds to evaluate as attainment targets, as they comprise similar influential watershed characteristics as the majority of the impaired watersheds.

Hierarchical Cluster Analysis

There is consistency in the clustering of the watershed variables between the untransformed and transformed data. In both, SoilD and Forest cluster together with higher values generally corresponding with attainment watersheds. In both datasets Agriculture and Wetlands cluster together along with Slope in the untransformed variables. This would be expected as Wetlands and Agriculture would decrease with the increase of Slope. Urban is clustered with Wetlands and Agriculture at a greater distance indicating that the relationship is less significant, but likely still associated with this Slope factor.

Both the untransformed and the transformed matrices clustered the watersheds in generally the same small groupings. Both result in two impaired clusters, three attainment clusters and two mixed clusters. This indicates that the chosen variables are resulting in meaningful clusters, though this does not address the goal of matching attainment watersheds with impaired watersheds. Both matrices also indicate three larger clusters, though these too are generally skewed toward impaired or attainment.

The clusters in the matrix of lowest ranking variables resulted in better within-cluster mixing of attainment and impaired watersheds than the clusters based on the most influential variables (Area, IC, Urban, SoilD and Forest). We think this is the most meaningful way to group impaired watersheds with appropriate attainment watersheds. One should use caution though, as the transformed matrix does not consider a number of watershed variables. It would be worthwhile to look into additional transformations that would normalize the Slope, SoilA and SoilB data so this method of grouping might be used on normalized data with more confidence.

Principal Components Analysis

The 2-factor loading plots for the untransformed and the transformed data look similar in the clusters that are illustrated, though the Factor 2 axes are reversed. In addition, the untransformed plot includes SoilA and Slope, while the transformed plot includes Water and SoilC. The resulting interpretation of these graphs is the same for both the untransformed and the transformed data. Factor 1, which separates IC and Urban from Forest, SoilD and Wetland, we interpret to represent a land disturbance factor. Factor 2 separates IC and Urban from

Agriculture and Wetlands, which we interpret to represent characteristics related to built infrastructure. Adding a third factor to the analysis does not change this general interpretation of Factor 1 and Factor 2. In fact, it further removes the variables unique to one plot.

Kruskal-Wallis MANOVA and Two-Sample t-Test

The Kruskal-Wallis test for untransformed data and the t-test for the normal, transformed data identified significant differences between the impaired and attainment watersheds for the Urban, Forest and IC variables. This is consistent with and supports the k-means cluster analysis, which is a more efficient and inclusive way to assess the data. The Kruskal-Wallis test also identified that the untransformed Slope data was significantly differences between the impaired and attainment watersheds. Slope could not be normalized with the selected transformations so the t-test was not appropriate. These watershed variable data support the hypothesis that urban land use and % impervious cover are the key factors driving the differences between impaired and attainment watersheds. Forest is generally inversely related to Urban and is significantly different between impaired and attainment watersheds, so this too could be a key indicator of watershed status.

CONCLUSIONS

We found that k-means cluster analysis combined with hierarchical cluster analysis could produce a statistically defensible way to group impaired and attainment watersheds to set watershed-level flow targets for stormwater permits. The k-means cluster analysis identified the variables that were most influential in separating impaired and attainment watersheds. These tended to be variables directly related to development (e.g., urban and impervious cover) and variables strongly autocorrelated with development (e.g., Soil D). When these variables are removed from the data set, the remaining variables describe the ‘inherent’ characteristics of the watersheds. Hierarchical cluster analysis of these less influential watershed variables produced natural groupings of watersheds that included both impaired and attainment streams. With the exception of Sunderland Brook watershed in the transformed dataset, the $Q_{0.3}$ attainment means were lower and the Q_{95} attainment means were higher than the corresponding flow values for each of the impaired watersheds in a given cluster. Thus, these attainment means could be used as watershed-level flow targets for the corresponding impairment watersheds.

We found that there was little difference in any of the statistical test we did when we used transformed versus untransformed data as the input. This suggests that these tests were relatively robust to the conditions of non-normality in the data. Care should be taken, however, when applying this methodology to the transformed data, as it accounts for fewer watershed variables than the clustering based on the untransformed data.

REFERENCES

Bonta, J.V. and B. Cleland, 2004. Incorporating Natural Variability, Uncertainty, and Risk into Water Quality Evaluations Using Duration Curves. Journal of the American Water Resources Association 39(6):1481-1496.

Caratti, J.F., Nesser, J.A., Maynard, C.L., 2004. Watershed Classification Using Canonical Correspondence Analysis and Clustering Techniques: A cautionary note. Journal of the American Water Resources Association 40(5):1257-1268.

Clausen, B., and B.J.F. Biggs, 2000. Flow Variables for Ecological Studies in Temperate Streams: Groupings Based on Covariance. Journal of Hydrology 237:184-197.

Haan, C.T., 1977. *Statistical methods in hydrology*. Ames: The Iowa State University Press.

Kent, M. and Coker, P., 1992. *Vegetation description and analysis: A practical approach.* England: John Wiley & Sons.

Olden, J. D. and N. L. Poff, 2003. Redundancy and the Choice of Hydrologic Indices for Characterizing Streamflow Regimes. River Research and Applications 19(2):101-121.

Santos-Roman, D., G.S. Warner, and F. Scatena, 2003. Multivariate Analysis of Water Quality and Physical Characteristics of Selected Watersheds in Puerto Rico. Journal of the American Water Resources Association 39(4):829-839.

StatSoft, Inc. (2004). *Electronic Statistics Textbook.* Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html.

SYSTAT Software, Inc., 2004. *Getting Started, Statistics I, Statistics II and Statistics III.*