Appendix 4.4. Additional Notes on K-Means Cluster Analysis Classification

A-4.4.1. Introduction to cluster analysis

The uncertainties and potential biases involved in grouping data by field observations and by geochemical pattern recognition require a non-subjective method of classification as an independent check, or to modify the previous groups. Cluster Analysis fits these requirements and can be used on its own as an independent partitioning method. Mathematical methods for partitioning multivariant data without *a priori* grouping (Cluster Analysis) have been commonly used since the mid 1960's. Many types of cluster analyses have been developed, particularly for use in medicine and biology. They can be divided broadly into hierarchical and non-hierarchical methods.

Statistical packages are available commercially which offer cluster analysis; Systat™ was chosen for this project.

Figure A-4.4.i shows the potential routes which could be taken in inputting and manipulating data and the various ways of clustering the data. The following section explains some of the theoretical considerations involved and outlines the choices made to analyze the geochemical data.

A-4.4.2. Data development and methodology

Data input

In addition to using the six (Na+ and K+ represent one) major ions to partition samples, it is also possible to use trace elements as was done with the "Trees" grouping method. The most important of the trace metals is iron, the presence of which (in the form of hydroxide precipitates) was used as a criterion in the field classification. Other trace metals, such as Ni, Co, Zn, Cd, Cu, W and Mo, could aid in the discrimination of clusters. However, as a number of these trace metals will be used to differentiate within groups (see Chapter 5), it is desirable to leave them out of the initial partition.

It was found by experimentation that Systat is capable of handling 120 cases with 7 variables. By experimenting with different combinations, it was determined that H^+ is the most revealing element (i.e., pH converted into mg/l H^+) and, therefore, was used as the seventh variable. High H^+ concentrations are often associated with the presence of Fe, which is probably the next most revealing element. Because of this association, the H^+ concentration can help to discern waters which have come into contact with iron sulphides.

All of the geochemical data used herein were reported by the laboratory in mass ratios (ppm, ppb, ppt). In this form, it is difficult to compare different ions because mass ratios are partially controlled by the atomic weight of the ion and do not take into consideration its charge. Equivalence units are more desirable because they allow comparison of all ions on the same scale. Another consideration is that the maximum magnitude of each of the major ions in the data set is similar when using equivalence units, although their statistical means differ considerably. Most clustering methods require that variables have similar scale and magnitude.

When dealing with trace elements, particularly metals, equivalence cannot be used because the dissolved species (usually a complex in which it is contained) is often not known and neither is its valence. The scale and magnitude of the mass ratio units (ppb, ppt), therefore, present a problem since they can vary by orders of magnitude both within and between variables. The H^+ concentration, with respect to that of major ions, is an example of this. This problem can be solved by using (the following) two techniques: standardization and log transformation.

Another way of inputting the major ion data is as cation and anion percentages. This will produce a different result because it completely eliminates the effect of the magnitude of the variables. This type of method is used when plotting Piper diagrams.

Data manipulation

There are three reasons for manipulating the raw geochemical results before cluster analysis is attempted. The first two, to standardize scale and magnitude across variables, have already been mentioned. The third reason is to weight individual variables so as to give them more or less significance with respect to the others.

Standardization of scale and magnitude must be considered for the variables when any sort of distance calculations are used in the clustering method. When calculating Euclidean (and other types of) distances between points, each element is a variable in its own dimension and contributes to the distance as a whole. If a particular variable is on a different scale than the others, such as ppb vs. ppm,

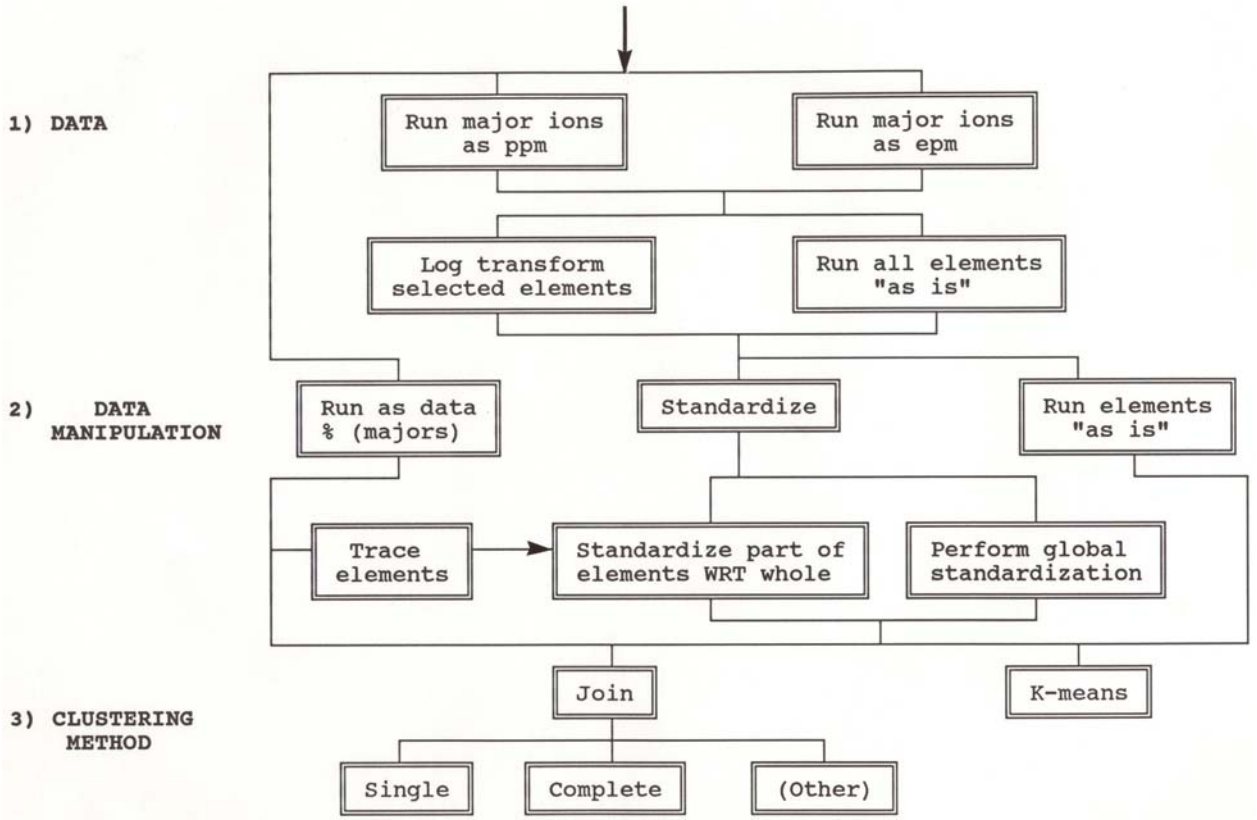

Figure. A-4.4(i) Flow chart to show potential routes for inputting, manipulating and clustering the data using cluster analysis. (from Hamilton, 1990, Fig. 4.5).

or has a drastically different magnitude, it will contribute a disproportionate amount to the distance calculation. This may result in groupings, which rely more on some variables or completely ignore others. Standardization is a form of proportional scaling that can be used to prevent this effect.

Standardization may mask true groupings by reducing the significance of important variables. Conversely, it may give too much significance to variables, which are not important in natural groupings. For example, if trace metals were included in the variables and all were standardized, then Mo, which is usually present in only minute amounts, would have the same significance as Ca^{2+}, which is almost always an important component. This problem is avoided if only selected variables are standardized.

Another form of data manipulation is the log transformation of variables, that finds the highest and lowest values in a sample population. It then re-scales the whole sample population to a fixed scale with the lowest value being 1 and the highest being 100. The log of each value is then taken. This reduces, but does not eliminate the effect of magnitude. It is an especially useful procedure when dealing with a variable or variables, which vary by orders of magnitude. It has the possible disadvantages of masking clusters due to magnitude and if all variables are log transformed it eliminates differences due to scale just like standardization. The results of log transformation can themselves be standardized. Indeed, if only several variables are transformed then one must standardize to a final scale similar to that of the other variables.

Expressing variables as a percentage of total charge eliminates the effect of their absolute magnitudes on clustering. A sample with only 50 ppm TDS content can thus be clustered in the same group as a sample of 3000 ppm TDS content if the relative proportions of the major ions in each sample are similar. This effect is not desirable because it will eliminate any groupings based on magnitude.

Another problem is that samples with very low concentrations (like recent snow melt or rainwater) can have their major ion ratios altered significantly by the addition of very small quantities of one or more of the major ions. This could take place in the soil zone, in a swamp or even by dissolution of aerosols and gases in the atmosphere. Because of this, major ion percentages in waters with very low TDS content may be indicative of processes totally unrelated to those taking place in waters with higher TDS content but similar major ion proportions.

Bearing in mind the potential problems with standardization and log transformation, it was decided to minimize data manipulation as much as possible. The six major ions all have similar maximum values, around 50 equivalence per million (epm), and, therefore, do not require standardization. The H^+ concentration varies by orders of magnitude therefore must be log transformed and subsequently standardized.

A-4.4.3. Clustering Method

Both hierarchical and non-hierarchical clustering methods may be suitable for analyzing geochemical data, and the statistics package used (Systat) offers one of each. The "Joining" method is hierarchical and the "K-means" is non-hierarchical.

Two requirements must be met before the joining algorithm can be employed. First, there must be a method for calculating "distance" between clusters in multidimensional space. Second, there must be a "linkage" method, which will amalgamate two clusters (which consist of one or more samples). The linkage method must include a definition of where the cluster is centered in order to calculate distance between it and other clusters. The term "joining", as used here, includes a number of different clustering methods. They differ largely in their distance calculation and linkage methods. They are similar in that they are all hierarchical, i.e., they partition points into a series of sets which are joined together to form nested sets which depend on how closely they resemble each other.

The earliest and best known linkage method is single linkage. In calculating the distance between two clusters it uses the smallest distance between any of the points in the two clusters, i.e., only a single link is needed between them. If that distance is sufficiently small, the clusters are joined.

Complete linkage is similar to single linkage, except that it uses the two most distant points between two clusters to establish a distance between them, i.e., all the points in each cluster can be considered linked. Again, if the established distance is sufficiently small the two clusters are joined. Complete linkage produces clusters that are more compact and farther apart than single linkage, which means that they represent more "ideal" groups. As a result, complete linkage was considered more discriminating and preferable to single linkage. Both methods were tried and complete linkage indeed produced the best results. Joining major ions and standardized (or weighted) H+ using complete linkage in Euclidean space produced good results. The method discriminated well between waters with differing chemistry.

The K-Means method is non-hierarchical and, therefore, produces a partition in which each point is assigned to only one cluster. Each point within a cluster must be closer to the cluster centre than to the centre of any other cluster. As the partitioning is wholly based on distance, it is essential that either a similar scale be used for each variable, or that data be standardized. The K-Means method of cluster analysis was "seeded" for this study by arbitrarily instructing the computer to start with 10 clusters. The outputs are listed in Table 4.4 (i). When a larger number was tried, the results were similar, except the more significant groups were further subdivided.

A-4.4.4. Discussion and results of cluster analysis

The following discussion is abbreviated in text Chapter 5.4.4. Based on the field evidence and on geochemical consistency within groups the "joining" method of clustering produced very satisfactory results. Text Figure 5.7 shows the hierarchical clustering of all points (except 026, 050, 117 and 118) using complete linkage in Euclidean space. The printout is organized hierarchically and gives visual representation of the process of clustering the data. Every sample starts off as a separate cluster and these clusters are joined in successive iterations until all samples are included in sets and all sets are joined into larger sets, the largest set being the sample population itself. The samples with the smallest distance between them are the most similar and are the first ones joined. The horizontal lines on the figure represent distances from a point or a cluster centre to the centre of the larger cluster in which it is contained. The vertical lines join samples or clusters together to form a new cluster and are positioned on the horizontal axis so as to give a representation of the distance that the new cluster

centre is from the origin.

The most significant groups of springs are the ones whose cluster centres are farthest from the centre of the cluster in which they are contained. For example, the most distinct group of springs is the one shown on the bottom, right hand side of the figure (group C1). This group has by far the largest distance between it and the higher cluster. Groups have been numbered on the figure in descending order of significance.

All of these groups have been identified before by at least one of the other two classification systems. The first six groups can be considered to represent "ideal" water types because they neatly pick out each of the distinctive groups of springs. They include 60% of all the samples taken. The remaining samples belong to the tentative groups C7t, C8t and C9t, which have been re-clustered. These show small total distances between samples and their clusters, which is mostly a result of low TDS contents. Definite hydrogeochemical environments are not apparent in any of the three groups because low TDS content causes poor differentiation of water types and because mixed and diluted waters from the above groups tend to be included here. Included in this manner are Fe-rich, neutral-pH waters which may be the result of mixing of C4 (or T5) type waters with high alkalinity waters from carbonate terrane (C5, or C7). The best chance of finding any environment-specific groups within C7t-C9t is to re-cluster the data outside the context of the larger spring groups.

Samples 11 and 34 were removed from the second run because these are known from field evidence to be examples of dilution of C6 type waters. The four samples removed in the first run for reasons of space (26, 50, 117 and 118) were added here. Three of these are thought to be examples of re-emergent surface water and, therefore, would likely have plotted in C7t-C9t. H^+ was re-standardized to a range equivalent to the range of major ion concentrations in clusters C7t-C9t. Log transformed and standardized Fe was added in the hope that it would help discern springs with water having iron above saturation limits.

Systematic weighting of variables can be tried in order to improve discrimination. Weighting is subjective and requires experimentation to see what will give the best results. The subjectivity is only in the way the samples are to be chosen, not in their actual grouping.

Two runs were attempted: one with variables as described above, and one with SO_4^{2-} and Fe increased by 50%. It was hoped that this weighting would help discern the samples with a character or component of C4 type waters. C4 type waters are particularly interesting because they indicated dissolution of sulphides. They are also problematic because most of the ones with neutral pH tend not to plot in C4.

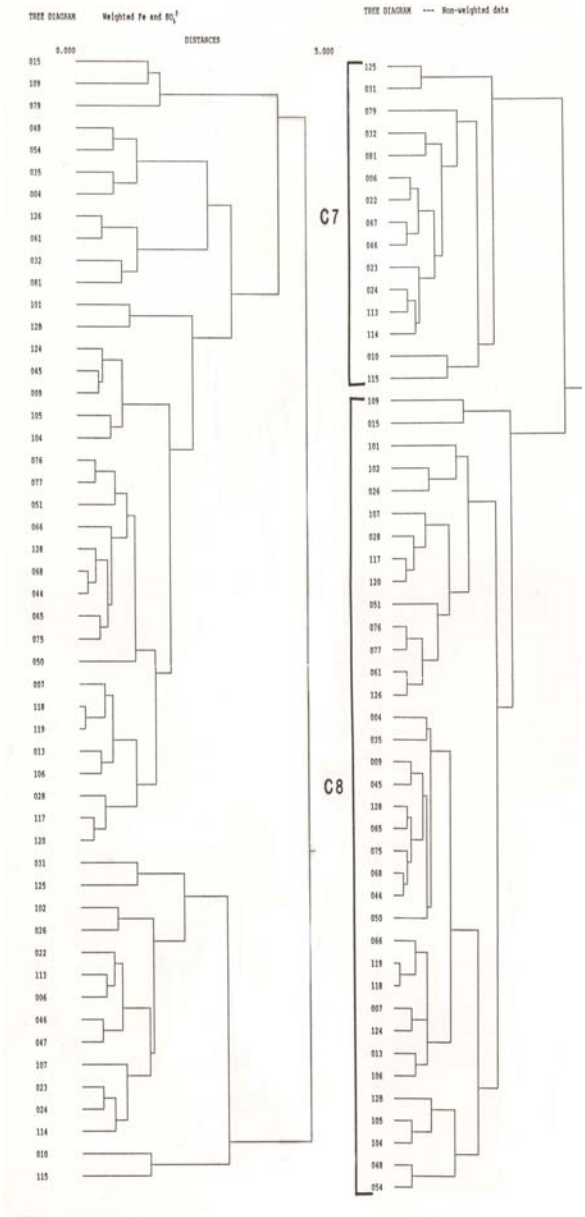

Figure A-4.4(ii). Output of cluster analysis on samples from groups C7t, C8t and C9t of Fig. 5.7 in text section 5.4.4. The run on the right is unweighted and the one on the left has had SO_4^{2-} and Fe increased by 50%. (from Hamilton, 1990, Fig. 4.7).

The results for each run are shown in Figure A-4.4.ii. The weighting improved the grouping slightly as can be seen by the increased distances between groups. However, it still did not clearly discern waters with a C4 component. Only a small number of the samples in the second run fit into reliable groups. The first group in the unweighted run (C7) contains many of the same samples as the first group in the weighted run. In both cases, there is a significantly larger distance between this group and the main group. The samples in the two groups are collectively referred to as C7 because they are considered to be significant.

Beyond this, groups have less significance and are collectively referred to as C8. The only common characteristic is that nearly all of the samples have low TDS contents. Low TDS content can result from of a number of subsurface conditions. The waters may not have had sufficient contact with the host rock to allow water-rock interaction due to short residence times. Oxidizing conditions may persist in the subsurface due to permeable media or a lack of reducing species for reaction. This will result in low concentrations of metals and other species, which have low solubility under oxidizing conditions. The system may be closed to the input of CO_2 or open to the loss of CO_2 (such as karstic systems) which will lower the solubility of carbonates. Surface water samples have similar chemistry to many of the low TDS samples and most are grouped among them in C8. Thus, the low TDS samples in C8 cannot be further grouped, with any degree of reliability by using cluster analysis. With a few exceptions, it seems better to examine these as one group than to force them into some of the other, relatively well defined groups, even though they show weak evidence of the same rock types.

The K-means method of cluster analysis produced problematic results. This method bases its calculations and output entirely on distance measurements (the error factor). Groups are made, based on total distance (magnitude) and this results in a classification based more on TDS content than on similar variability of elements. All the samples which have low to moderate TDS content were grouped together under K-means because several other groups (such as A, C and B1 in the final grouping) had very high TDS concentrations. This also happens with heirarchical cluster analysis but the heirarchical subdivision of the lower TDS groups successfully displays the further grouping of these samples. There is no way to subdivide groups under K-means except to instruct it to produce (or "seed") more groups. The result is that it further subdivides all the groups of high TDS samples because relatively small differences in the concentration of the elements in these groups will be greater than any differences possible within the low TDS groups. The K-Means method of clustering is therefore considered to be unsatisfactory for the Nahanni population of spring waters.

Table A-4.4.(i) Results of K-Means Cluster analysis (from Hamilton, 1990, Appendix 4). Major ions are reported in ppm. H^+ is given in standardized units which are explained in more detail in section 4.4.2.

SUMMARY STATISTICS FOR 10 CLUSTERS

VARIABLE	BETWEEN SS	DF	WITHIN SS	DF	F-RATIO	PROB
CA	6750.851	9	514.874	109	158.182	0.000
MG	1622.048	9	157.441	109	124.776	0.000
NA	10618.326	9	428.641	109	300.017	0.000
HCO3	5932.774	9	559.940	109	128.322	0.000
SO4	10286.479	9	740.614	109	168.213	0.000
CL	16332.056	9	101.320	109	1952.215	0.000
HST	9258.779	9	1725.493	109	64.987	0.000

CLUSTER NUMBER:1

MEMBERS		STATISTICS				
CASE	DISTANCE	VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
004	1.26	CA	0.06	1.83	5.70	1.46
006	1.59	MG	0.00	0.85	2.75	0.78
007	1.08	NA	0.00	0.78	9.99	1.64
009	1.12	HCO3	0.00	2.49	7.29	1.64
010	2.03	SO4	0.02	0.90	6.91	1.09
011	0.96	CL	0.00	0.06	0.71	0.11

013	0.58	HST	29.00	36.40	46.25	4.16
018	1.09					
019	1.47					
020	2.17					
022	1.32					
024	1.49					
028	1.17					
031	2.46					
033	3.26					
035	1.40					
038	1.13					
044	0.71					
045	1.43					
046	0.95					
047	1.64					
048	2.89					
051	1.44					
052	1.82					
053	3.05					
054	2.86					
060	3.32					
061	1.73					
065	0.58					
066	0.87					
068	0.46					
075	0.45					
076	1.21					
077	1.40					
079	1.24					
081	2.06					
12B	2.97					
18A	1.99					
18B	2.00					
33	3.74					
34	0.78					
53	3.40					
60	3.88					
101	2.71					
104	1.35					
105	1.28					
106	1.33					
107	1.50					
113	1.34					
114	1.70					
116	4.81					
119	1.05					
120	0.79					
124	0.89					
126	1.64					
128	0.34					

CLUSTER NUMBER: 2

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
063	0.83	Ca	21.50	24.24	25.42	1.59
064	1.14	Mg	7.57	7.67	7.74	0.06
64	1.45	Na	50.79	53.32	56.63	2.11
123	2.05	HCO_3	3.19	4.31	6.02	1.07
		SO_4	15.20	17.03	17.91	1.09
		Cl	62.31	65.21	66.70	1.71
		HST	26.50	27.42	30.02	7.50

CLUSTER NUMBER: 3

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
73A	1.56	CA	14.91	16.90	18.12	1.42
73B	3.75	MG	14.41	16.26	17.43	1.32
73C	2.28	NA	0.80	0.84	0.85	0.02
		HCO3	0.00	2.55	4.51	1.89
		SO4	31.09	32.90	34.40	1.37
		CL	0.03	0.03	0.03	0.00
		HST	16.44	25.45	31.05	6.43

CLUSTER NUMBER: 4

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
017	3034	Ca	22.76	28.01	34.14	4.02
056	2.56	Mg	4.19	5.82	9.16	1.94
057	3.45	Na	1.23	2.66	5.58	1.46
058	1.43	HCO_3	28.22	33.70	39.80	4.42
17	2.12	SO_4	0.66	2.53	6.35	2.28
58	2.59	Cl	0.08	0.26	0.63	0.24
		HST	22.76	24.13	27.97	1.78

CLUSTER NUMBER: 5

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
012	0.83	Ca	0.06	1.42	4.82	1.37
025	2.29	Mg	0.10	1.07	3.74	1.09
036	1.50	Na	0.02	0.19	1.67	0.43
042	0.41	HCO_3	0.00	0.45	5.99	1.54
049	3.43	SO_4	0.28	3.32	11.70	3.28
062	3.75	Cl	0.00	0.02	0.09	0.03
070	2.99	HST	4.40	13.41	23.12	5.16
080	3.48					
12A	1.05					
12C	2.02					
100	0.97					
103	3.67					
108	1.42					
121	3.01					

CLUSTER NUMBER: 6

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
072B	3.72	Ca	12.63	14.90	17.17	2.27
72B	3.72	Mg	12.46	14.97	17.48	2.51
		Na	0.70	0.73	0.76	0.03
		HCO_3	0.00	0.00	0.00	0.00
		SO_4	31.65	39.60	47.54	7.94
		Cl	0.06	0.06	0.07	0.01
		HST	4.77	5.25	5.73	0.48

CLUSTER NUMBER: 7

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
001	3.27	Ca	1.23	6.94	13.87	2.80
015	2.52	Mg	0.58	2.93	5.37	1.20
021	3.42	Na	0.00	1.62	8.99	2.55
023	2.00	HCO_3	1.05	6.06	12.60	2.90
027	1.99	SO_4	0.31	4.85	12.86	3.49
029	2.15	Cl	0.01	0.65	7.83	1.75
030	1.71	HST	25.55	30.34	36.70	2.69

032	1.98
037	2.03
040	2.46
041	3.53
043	1.37
059	1.99
067	2.02
074	2.62
078	2.67
082	1.30
083	4.33
27	1.93
82A	1.26
82B	1.46
102	2.78
109	3.08
110	4.39
111	3.39
115	1.62
122	3.80
125	1.90
127	1.33

CLUSTER NUMBER: 8

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
055	0.00	Ca	17.21	17.21	17.21	0.00
		Mg	3.46	3.46	3.46	0.00
		Na	0.53	0.53	0.53	0.00
		HCO_3	18.03	18.03	18.03	0.00
		SO_4	3.02	3.02	3.02	0.00
		Cl	0.15	0.15	0.15	0.00
		HST	21.14	21.14	21.14	0.00

CLUSTER NUMBER: 9

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
069	1.40	Ca	1.79	2.84	3.88	1.04
071	1.40	Mg	2.47	3.56	4.56	1.09
		Na	0.44	1.54	2.64	1.10
		HCO_3	0.00	0.00	0.00	0.00
		SO_4	11.97	12.34	12.72	0.38
		Cl	0.02	0.30	0.58	0.28
		HST	0.00	3.16	6.31	3.16

CLUSTER NUMBER: 10

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
072A	2.65	Ca	19.80	22.77	25.75	2.98
112	2.65	Mg	13.18	16.51	19.85	3.34
		Na	0.69	5.87	11.05	5.18
		HCO_3	0.52	1.20	1.88	0.68
		SO_4	47.76	48.89	50.03	1.13
		Cl	0.42	0.45	0.47	0.03
		HST	22.98	23.42	24.15	0.73