EXAMINING THE VALIDITY OF SAMPLE CLUSTERS USING THE BOOTSTRAP METHOD

by

DAVID MICHAEL SHERA

B.A. Rice University
(1981)

Submitted to the Department of
Mathematics
in Partial Fulfilment of the
Requirements of the Degree of

MASTER OF SCIENCE IN APPLIED MATHEMATICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 1983

Signature of Author _____

Department of Mathematics, August 5, 1983

Certified by _____

Professor M. A. Wong, Thesis Supervisor

Accepted by

Professor D. J. Benney, Appl... Math Chairman

EXAMINING THE VALIDITY OF SAMPLE CLUSTERS USING THE BOOTSTRAP METHOD

by

DAVID MICHAEL SHERA

Submitted to the Department of Mathematics
in Partial Fulfillment of the Requirements of the Degree of

MASTER OF SCIENCE IN APPLIED MATHEMATICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 1983

ABSTRACT

An important problem in clustering research is the stability of sample clusters. Cluster diagnostics, based on subsampling procedures like the bootstrap and cross-validation methods, will be developed in this thesis to aid the users of cluster analysis in assessing the stability and validity of sample clusters.

Thesis Supervisor : Dr. M. A. Wong

Title : Associate Professor of Management Science

1. INTRODUCTION

An important problem in clustering research is the stability and validity of the sample clusters. For Euclidean data, Hartigan (1981), Wong (1982), and Wong and Lane (1983) have developed procedures to evaluate sample clusters using the density-contour clustering model. However, it is often true that in clustering the nations of the world, or the political states of the United States, or the companies of a major industry, the objects under study cannot be reasonably viewed as a sample from some such underlying population. Under these circumstances, it is not reasonable to talk about sampling errors in the computed clusters. But, the stability of the sample clusters can be evaluated in other ways.

In the approach taken by Baker (1974), Hubert (1974), and Baker and Hubert (1975), the question of how the clustering solutions provided by the single and complete linkage methods are affected by changes in the distance or similarity matrix is addressed. This is a reasonable question as different data collected on the objects might change the distances or similarities. Ling (1973) took a different approach and developed an exact probability theory for testing the compactness and isolation of single linkage clusters, under the (admittedly unrealistic) assumption that the rank order of the entries in the distance matrix is completely random.

Another approach is appropriate when the sample similarity matrix provides an approximation of the population similarity matrix for a fixed set of objects. For example, in marketing research, if the brand-switching behavior of the total population of coffee consumers were known, a population similarity matrix (in terms of relative frequency of switching between brands) between various coffee brands could be obtained. For such a finite population, a true hierarchical clustering can be defined on the objects (here, coffee brands) using the population similarity matrix (e.g., the block-distance clustering model can be used). However, only the brand-switching behavior of a sample of coffee consumers is available in practice, and the sample similarity matrix obtained only provides an approximation of the true aggregate similarities between brands. Consequently, the sample hierarchical clustering obtained from this similarity matrix is merely a sample estimate of the true hierarchical clustering.

In this type of study, standard sampling procedures like the bootstrap or cross-validation methods described in Efron (1979a, 1979b), or the error analysis scheme given in Hartigan (1969, 1971) can be usefully applied to the sample subjects (e.g., coffee consumers) to perform two main tasks:

1. to assess the similarity between the sample and population hierarchical clusterings, and

2. to assess the stablility of the sample clusters

both using the Bk measure developed by Fowlkes and Mallows at Bell Laboratories (1983).

In Section 2, we describe the theoretical background for this study. There are discussions on the type of data we are concerned with, the clustering methods used, the statistics to compare clustering trees, and how the bootstrap method is used. The sampling experiments are described in detail in Section 3. The simulation results and their implications are also discussed there. Section 4 draws conclusions and describes how the techniques can be used most effectively.

2. RESEARCH BACKGROUND

In this study, our main concern is to develop diagnostic tools for assessing the stability and validity of sample clusters in the case where the sample similarity matrix is merely an approximation of the population similarity matrix. For a finite population, a true hierarchical clustering can be defined on the objects by the ultrametrics model (Johnson 1967). In order to evaluate the clusters obtained by various clustering techniques from the sample similarity matrix, we need to examine the degree of agreement between the population and sample clusters and its distribution in a series of simulated sampling experiments. We will adopt Fowlkes and Mallows' Bk statistic as a measure of the degree of agreement. The clustering techniques used in this study are described in section 2.1, and the Bk statistic will be reviewed in section 2.2. In section 2.3, the bootstrap procedure will be outlined.

2.1. CLUSTERING METHODS

Clustering is a splitting of the set of objects into partitions, or sets, or groups of one or more objects. A heirarchical clustering is a set of clusterings, { Ci }, of the objects indexed from 1 to N, the number of objects. The

index i is the number of clusters in partition Ci. What makes it hierarchical is that if objects X and Y are in the same cluster in partition Ci then they are in the same cluster for all Cj where j < i. A tree will mean a dendrogram, a taxonomy, and a number of other things which all can mean hierarchical clustering. The term tree reflects the property that a hierarchical clustering can be represented on paper by something which resembles an upside-down botanical tree.

Here is a more dynamic definition of a tree. Given a distance matrix, each clustering method proceeds as follows: Start with each object in it's own cluster. At each step combine two clusters into a single cluster. Continue linking until all the objects are in a single cluster. Thus for N objects, we will have N-1 linkings and a tree of N clusterings, the first and last being trivial.

The differences in the linking methods come from different criteria used to decide which two clusters to combine at each step. One of the most common criteria is the distance between the closest two objects of the two clusters. This is known as single linkage. Complete linkage defines the distance between two clusters to be the maximum of the distances between any object in the first cluster and any object of the second. At each step, this method links the two clusters which are closest together by that distance

measurement. Average linkage is the same except it uses the average of the distances between objects in each cluster.

Wong and Lane (1983) have proposed a method which involves looking at the Kth nearest neighbor to each object where K is a parameter chosen by the data analyst. The rationale comes from density estimation so that it is like estimating the density of the true distribution at each object. Clusters are linked together by highest density first and only on the additional condition that at least some object in one cluster be within the Kth nearest neigborhood of some object in the other cluster. As a result, this method also produces an estimate for the number of clusters. Wong and Lane have suggested a method to decide which K to use by looking at the results for all values of K and see what the most common value of the number of clusters is.

There are many other existing clustering methods and criteria , but they will not be considered here.

2.2. THE Bk STATISTIC

The definition of Bk is as follows: The k in Bk refers to the number of clusters and we will compare the clusterings with k clusters in the two trees. Consider a table Mij where i and j run from 1 to k and where Mij equals the number of

objects in cluster i of clustering 1, the clustering from the first tree, and cluster j of clustering 2, the clustering from the second. Let Mio be the number of objects in cluster i of the first clustering, and Moj be the number of objects in cluster j of the second. Mio and Moj are the marginal totals of the rows and columns of table Mij.

Given k, let

$$Pk = \sum_i Mio^2 - N$$

$$Qk = \sum_j Moj^2 - N$$

$$Tk = \sum_{ij} Mij^2 - N$$

and then

$$Bk = \frac{Tk}{\sqrt{Pk\ Qk}}$$

Fowlkes and Mallows (1983) calculated a null distribution for Bk, that is, the distribution for two completely independent trees. However, it is difficult to imagine a situation in which the null distribution should be considered since there is almost always going to be clustering of some kind. Most situations will have significant deviation from the null case.

Rand (1971) also proposed a statistic to measure the similarity between clusterings. With the same P, Q, and T as defined above for Bk, Rand's statistic was

$$Rk = \frac{Tk}{C} - \frac{Pk+Qk}{2C} + 1$$

where C equals Comb (n, 2), the number of combinations of n objects taken 2 at a time.

In the paper introducing Bk, Fowlkes and Mallows show clearly that Rand's statistic was insensitive since it did not indicate important differences in situations where Bk rightfully did. We will use the Bk statistic in this study.

2.3. THE BOOTSTRAP METHOD

Bootstrapping is a numerical technique proposed by Efron (1979a,b,1983) to estimate the distribution of a statistic. Given a true probability distribution F, a set of N random variables X, = {Xi}, independent and identically distributed (iid) with distribution F, and a statistic R (X, F), the objective is to estimate the distribution of R. To do the bootstrap, consider the empirical distribution of X, i.e. each Xi has probability 1 / N and call this distribution G. Note that not each value of Xi has equal probability because Xi might have the same value for two or more values of i.

Now draw a number of new samples Yj, = {Yi}j, from G where the size of each sample Yj is the same as that of X, i.e. i runs from 1 to N. Then calculate R (Yj, G) for each bootstrap sample Yj.

The main assumption is that G is a good approximation of F. With this assumption, we say that the empirical distribution of the R(Yj,G)'s is a good approximatien of the true distribution of R (X,F). This is not an unreasonable assumption: the same assumption applies when a statistician rejects a model because it lies outside some confidence region. He or she is assuming that the data is not unusual or exceptional. Standard probablility theory tells us that for a large N, G approaches F almost surely with suitable, but hardly restrictive conditions. The data analyst may decide to smooth the data by adding random noise or fitting a smooth distribution to it if he or she feels that is appropriate.

A bootstrap sample is a sample based on the original sample by being drawn from the empirical distribution, G, defined by the original sample, or perhaps a smoothed version. The term sample by itself will mean a sample from the true distribution, not a bootstrap. In this study, we have an unknown "population" similarity matrix and a sample similarity matrix. The bootstrapping procedure will be

applied to the elements of the sample similarity matrix independently to obtain subsample matricies.

2.4 SIMULATION DATA

It is difficult to make general statements about data because it comes in so many different forms. In fact, translating data to computer usable form usually requires writing a new program for each problem. Here we describe our data and how it is treated in this paper.

Objects are the things of interest which we want to cluster. These objects might be variables or countries or brands of chewing gum. N will denote the number of objects here. Data is made up of responses or single values, e.g. a single draw Xi from a distribution F is a response. A sample is a collection of independent responses.

In this study, the set of objects of interest will not change within the scope of a single problem. On the other hand, getting many samples is an important part of the bootstrap technique and so there will be many different samples and resamples. The bootstrap procedure we used has the resample size fixed equal to the original sample size. (Some studies have been done where the sample size does change for different samples (Hartigan 1981), but we will not be concerned with that here.)

Since all the clustering methods we will be using involve a distance matrix, eventually, through some operations, we want to change the raw data into a distance matrix between the objects. The resulting distance need not be a true metric in the mathematical sense, i.e. it need not satisfy the triangle inequality. Since this is not the focus of the paper, we will assume that the transformation from X to its distance matrix is clear and done implicitly. Thus, the techniques presented here will apply to any situation where one can create a distance or similarity matrix from the data.

Here are some examples of responses and their relation to distance matrices:

a) A sample of distance matricies, one for each individual giving his or her associations between objects combined in some fashion. Some average of these responses would then become the distance matrix corresponding to the sample.

b) A vector of values for each individual where we are trying to cluster the variables. We might measure the association between variables by linear or monotone correlations, or something more sophisticated.

c) A simple distance matrix between the objects. This is the case addressed by Hubert (1974) and will not be approached here.

In our experiments, the ultrametrics tree model will be used to generate the population distance matrix. Below is the distance matrix for data set A1; it corresponds to a hierarchical clustering of 10 objects with 3 well defined clusters. Following that is the tree corresponding to set A1. Other data sets will be introduced in Section 3 as they are used.

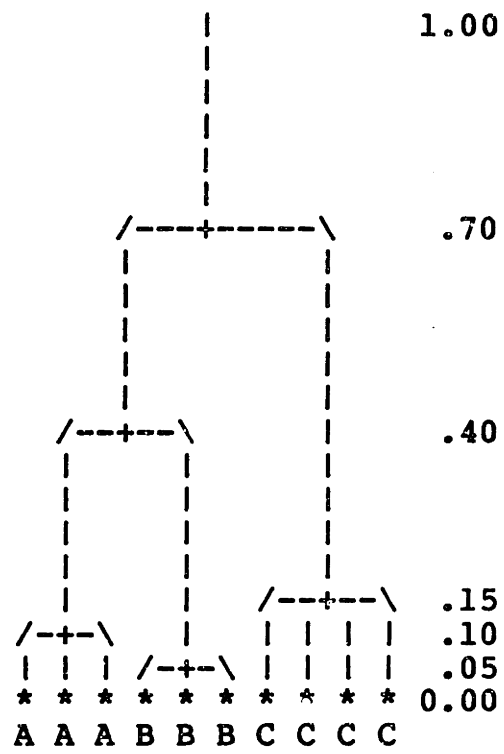
Table 2.4.1: True Distances of Set A1

	A1	A2	A3	B1	B2	B3	C1	C2	C3	C4
A1 -		.10	.10	.40	.40	.40	.70	.70	.70	.70
A2 -			.10	.40	.40	.40	.70	.70	.70	.70
A3 -				.40	.40	.40	.70	.70	.70	.70
B1 -					.05	.05	.70	.70	.70	.70
B2 -						.05	.70	.70	.70	.70
B3 -							.70	.70	.70	.70
C1 -								.15	.15	.15
C2 -									.15	.15
C3 -										.15

The application we have in mind can be described as follows: Suppose the objects were brands of detergents and the respondents were consumers. The real distance between detergent 1 and detergent 2 is the proportion of consumers who would not use each as a substitute for the other. In

marketing research studies, a finite sample of , say, 100 people would be tested if they would accept one product as a substitute for the other. So if 80 percent of the people switched, then the distance is .20, and it is a reasonable estimate of the true distance in the population.

Figure 2.4.2: Tree for Data Set A1



To simulate the survey described above, the natural thing to do is to create binomial random variables with probability parameters equal to the distance and with n = 100 trials. Sample distances are distributed Binomial (Pij,100)/100 where Pij is the true distance between objects i and j. Likewise, if Sij is the sample distance, the bootstrap distances are distributed Binomial (Sij,100)/100. The next two distance matrices are examples of a sample from set A1 and then one of the bootstrap samples based on that sample.

Table 2.4.3: Example of Sample Distances

	A1	A2	A3	B1	B2	B3	C1	C2	C3	C4
A1 -		.14	.12	.38	.46	.38	.69	.74	.67	.67
A2 -			.11	.45	.41	.46	.72	.69	.67	.75
A3 -				.43	.46	.37	.70	.66	.76	.75
B1 -					.05	.02	.74	.81	.64	.67
B2 -						.07	.69	.71	.73	.64
B3 -							.73	.62	.71	.74
C1 -								.15	.13	.14
C2 -									.15	.14
C3 -										.17

Table 2.4.4: Example of Bootstrap Sample

	A1	A2	A3	B1	B2	B3	C1	C2	C3	C4
A1 -		.18	.14	.32	.53	.30	.69	.74	.66	.63
A2 -			.10	.45	.37	.39	.75	.74	.58	.74
A3 -				.46	.41	.39	.73	.66	.73	.71
B1 -					.07	.01	.78	.82	.61	.72
B2 -						.08	.63	.69	.74	.63
B3 -							.72	.63	.69	.71
C1 -								.05	.07	.18
C2 -									.17	.06
C3 -										.17

3. RESULTS

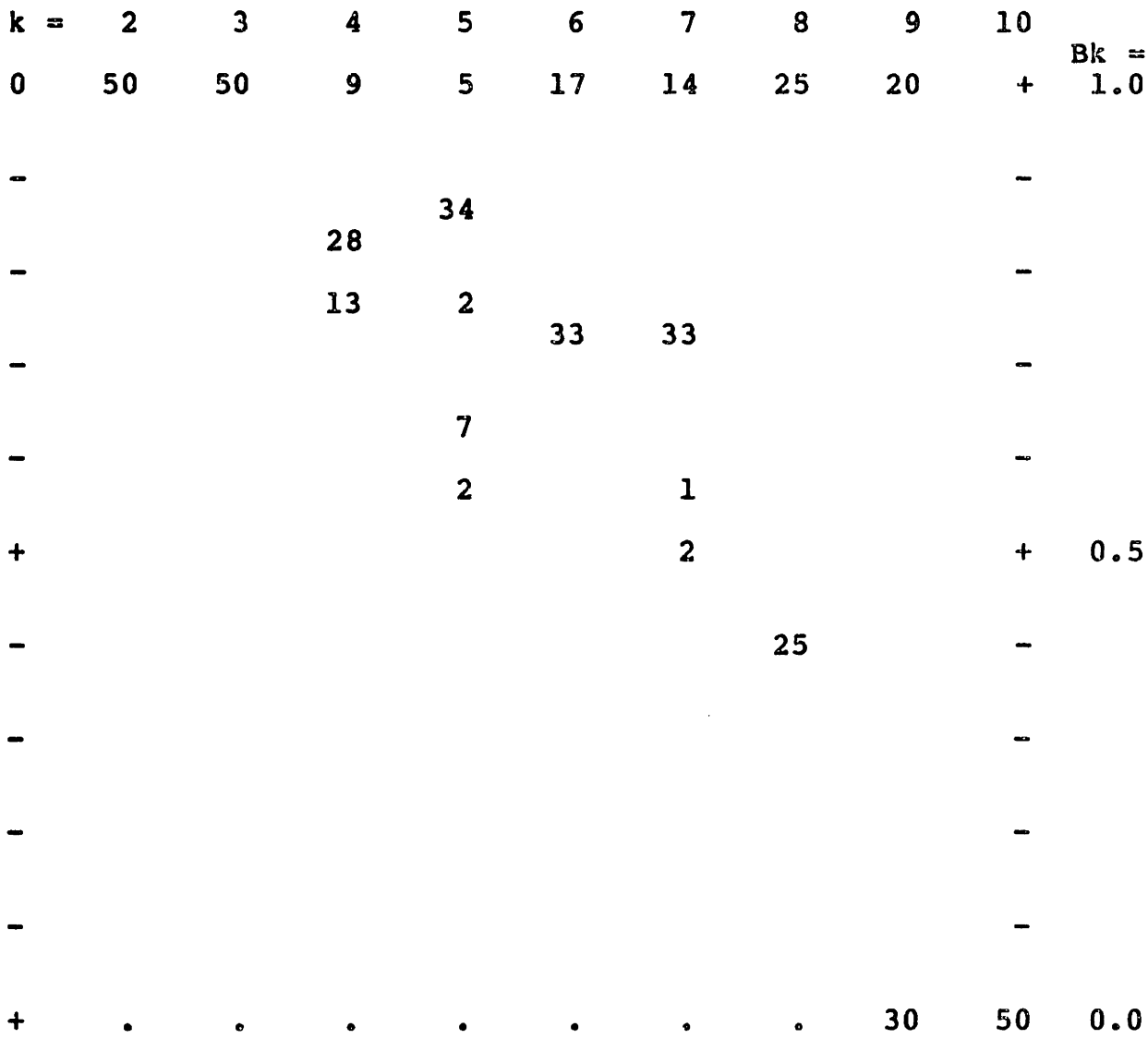
3.1 BK PLOTS

3.1.1 Data Set A1, Complete Linkage

In the first experiment based on data set A1, 50 samples from the true distribution were generated and complete linkage was used to make 50 trees. Bk was calculated by comparing each of the 50 trees with the true tree.

The distribution of Bk for each k is shown in the figure 3.1.1.1. The horizontal scale runs from 1 to N, (N = 10); 1 and N lie at the left and right borders, respectively. The vertical scale runs from 0 at the bottom to 1 at the top and hatch marks are at increments of .10. For each k, the numbers count how many Bk's took on that value.

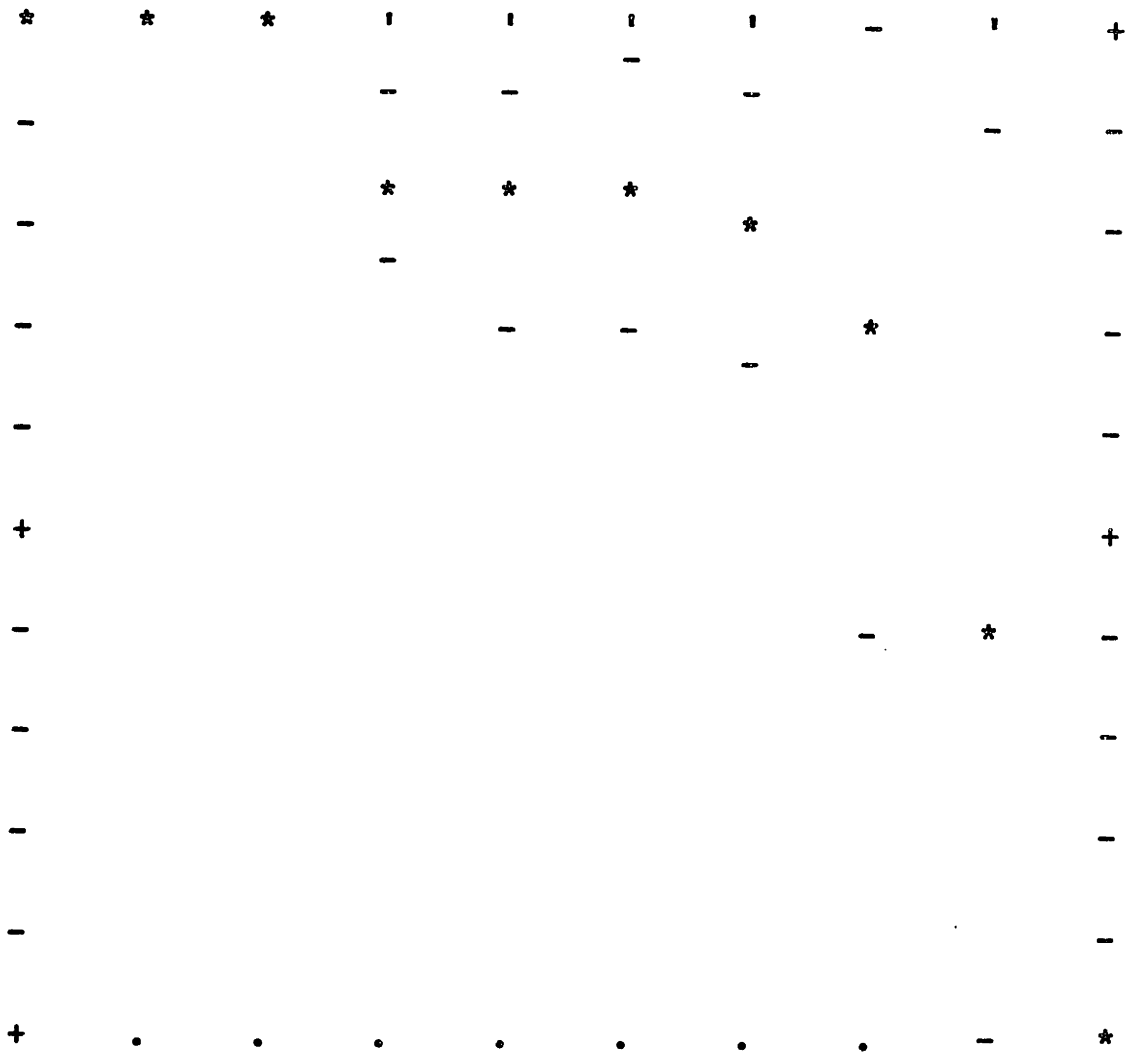
Figure 3.1.1.1: Plot of Bk versus k
Set A1, Complete Linkage



At k = N, Bk is always 0 and as k goes back towards 1, clusters link up and raise, or sometimes lower, the value of Bk. At k = 1, Bk is always equal to 1. It is clear from these plots that the distribution of Bk for a given k is often skewed and highly discreet. In fact, for k = 9, (k = N - 1 in general), the only possible values for Bk are 1 and 0.

Figure 3.1.1.2 shows the means of Bk, indicated by stars, "*", and plus and minus one sample standard deviation on each side, indicated by dashes, "-". They are truncated at the top and bottom boundaries because Bk will only lie in the unit interval.

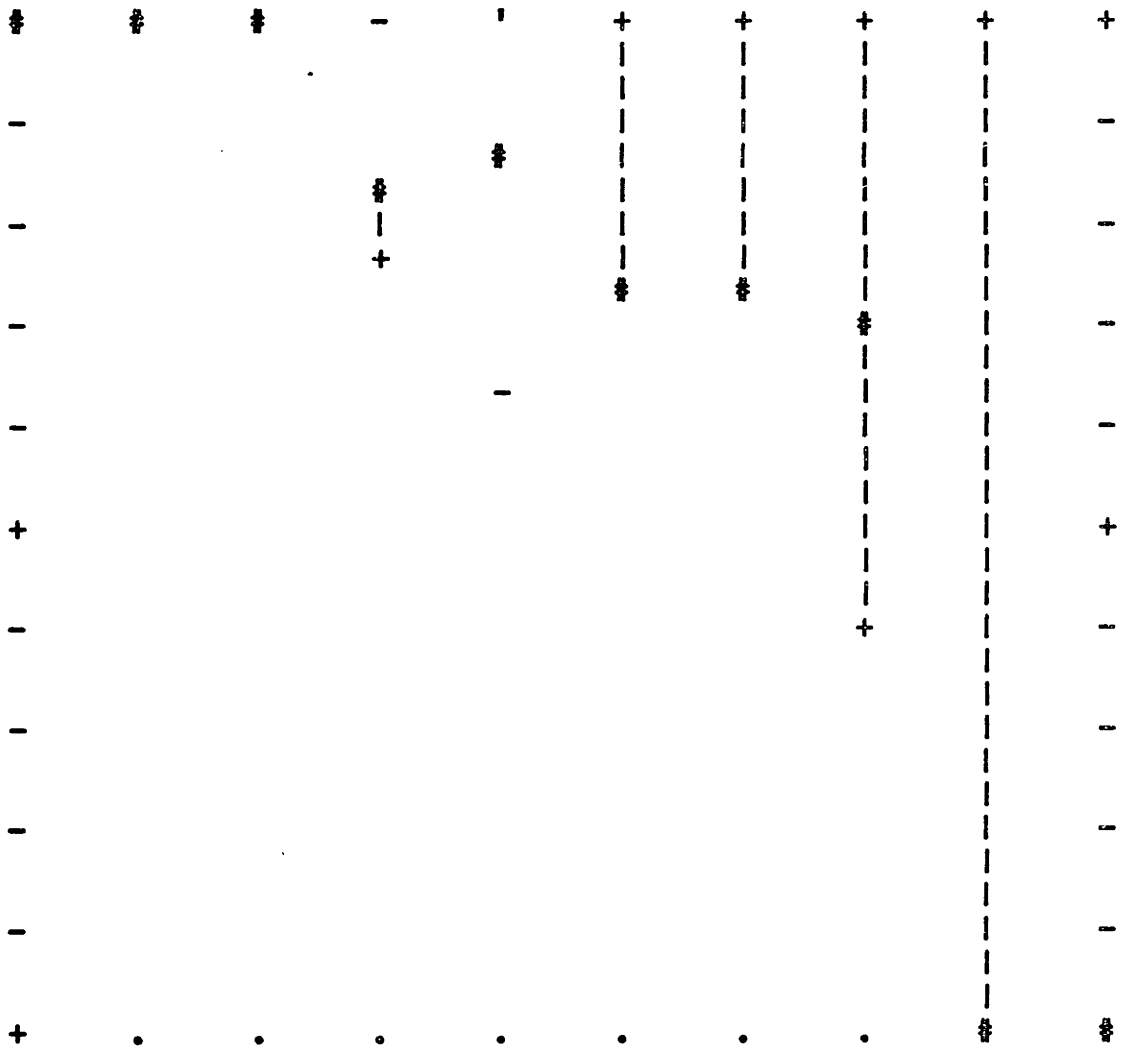
Figure 3.1.1.2: Plot of Bk versus k
Set A1, Complete Linkage



The following are box plots of the same results. The median is represented by a number sign, "#", the quartiles by a plus sign, "+", and the first and seventh eighths by minus signs or dashes, "-". The regions in between the quartiles was filled in with vertical bars, "|", to make the boxes in the box-plots more visible. Often, because the eighths are equal to the quartiles and/or the quartiles equal to the

median, only one of the symbols is shown. The discreteness of the distribution causes this problem. It also often causes rather large boxes for higher values of k.

Figure 3.1.1.3: Plot of Bk versus k
Set A1, Complete Linkage



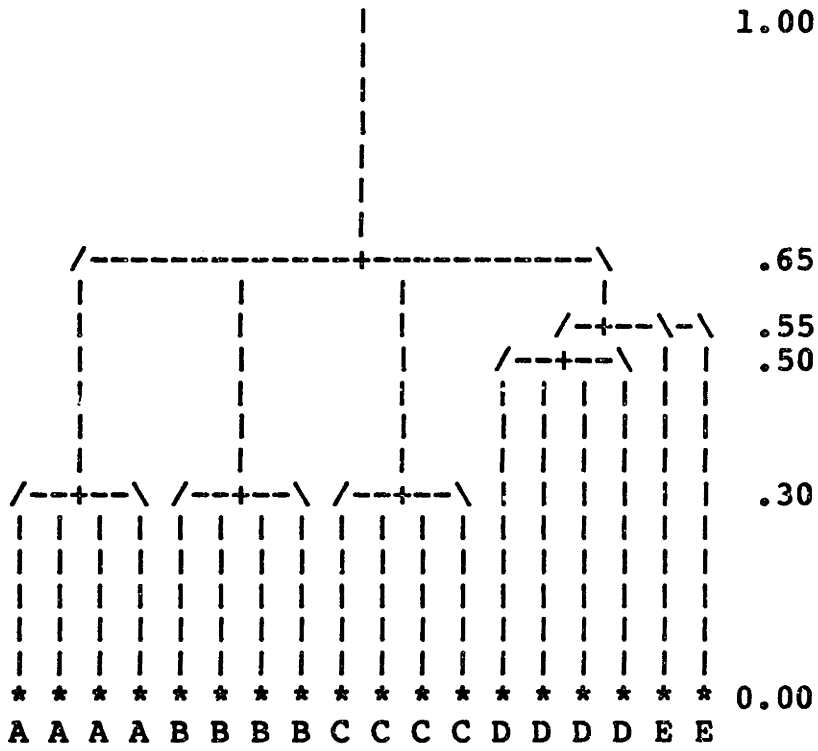
Because of the discreteness, simply finding the average may not be an appropriate summary of the information. The box plots seem to convey the variability of Bk most efficiently. The frequency plot contains the most information, but can be difficult to read, especially when there are many objects.

Before we continue with more experiments, let us explain the important feature of the Bk plot. It is clear that the plot of Bk is not a smooth one but jagged and contains many peaks, valleys, and steep cliffs. The reason is natural, and actually desirable, because the peaks indicate the more stable and relevant clusterings.

To see the reason, consider the structure of Set A1 and the process of linking together clusters. Even with the perturbation to for the samples, the objects labelled with B's are most likely to link before those with A's or C's. After 2 linkings, when k = 8, the B cluster is usually linked and Bk is higher. At k + 1, (9), none of the clusters is completely linked, and only random pieces have been linked. Then, Bk is smaller. Similarly, for k - 1, (7), Bk is smaller because the incomplete cluster of A's is only beginning to link together. The peak is not always exactly at 8 though because sometimes some A's may get a small distance in the sample, or the B's a large distance, and then A's link before the B cluster is complete. So by looking at the clustering at level k when k is one of these peaks, one will likely find stable clusters.

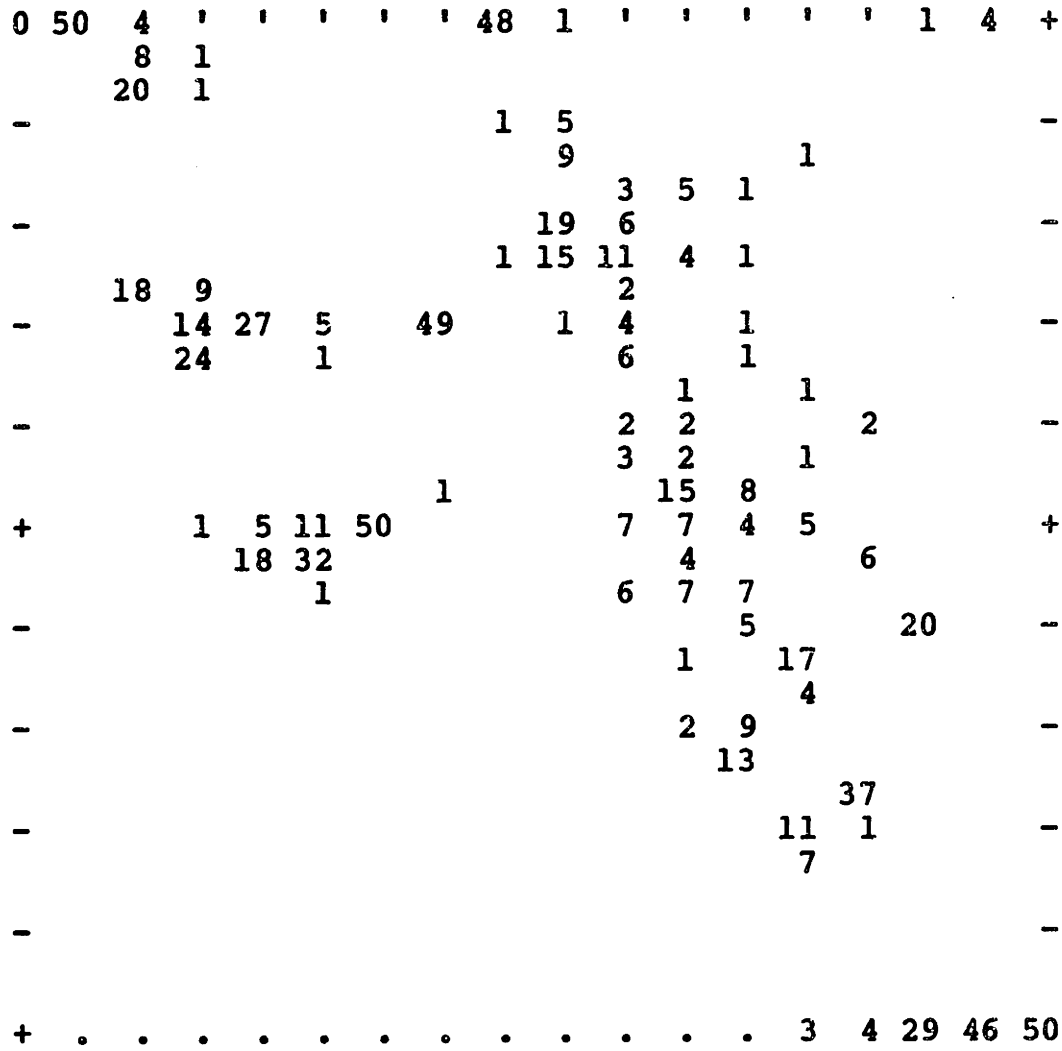
3.1.2. Data Set B1, Complete Linkage

Figure 3.1.2.1: Tree for Data Set B1



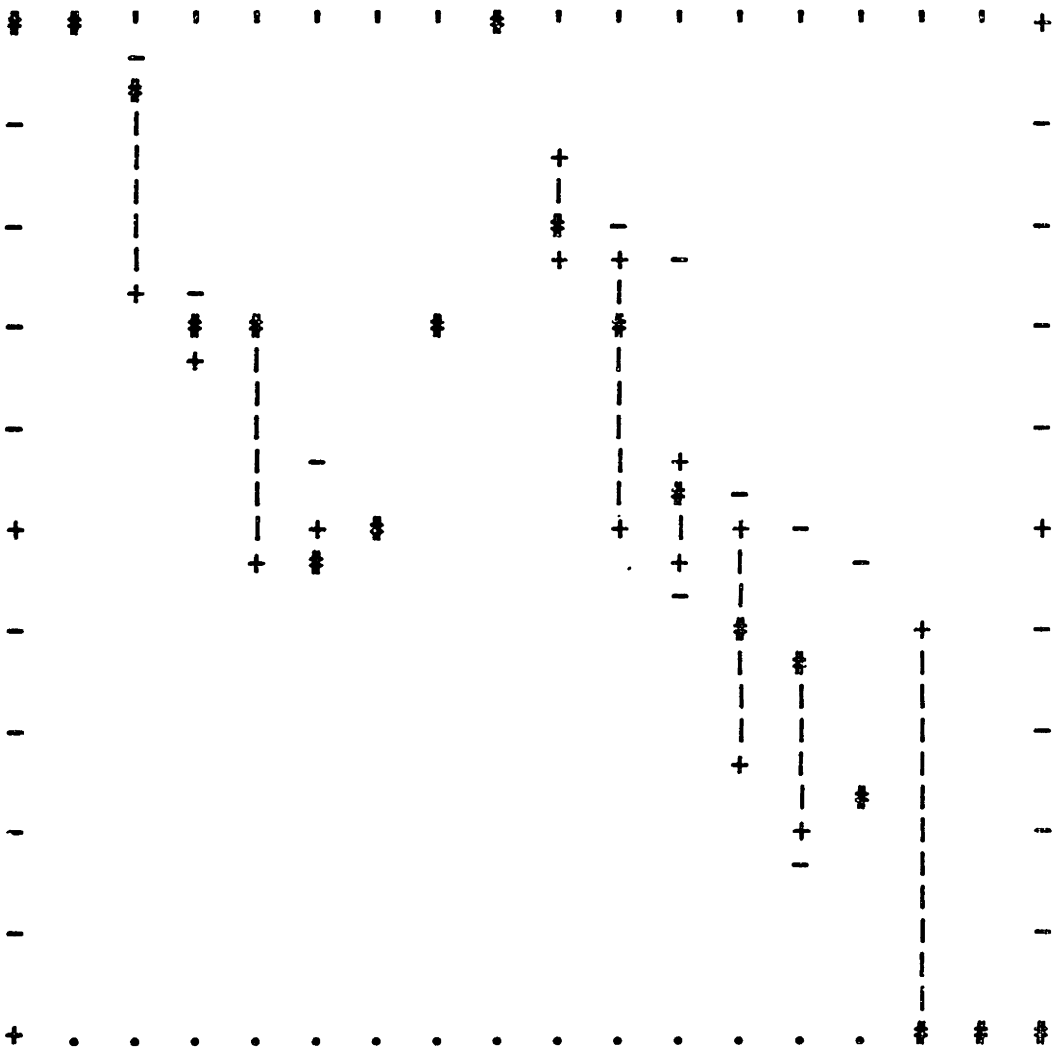
The three figures here are for Data Set B1, consisting of 18 objects, which represents a different population. There are 3 distinct clusters of 4 objects each labelled with A's, B's, and C's, and 6 other objects labelled with D's and E's which are in a less distinct cluster.

Figure 3.1.2.2: Plot of Bk versus k
Set B1, Complete Linkage, 50 Samples



More objects made the distribution smoother, except for the very high values of k; but it is still not smooth enough for us to use only the sample mean and standard deviation as descriptors for the distribution.

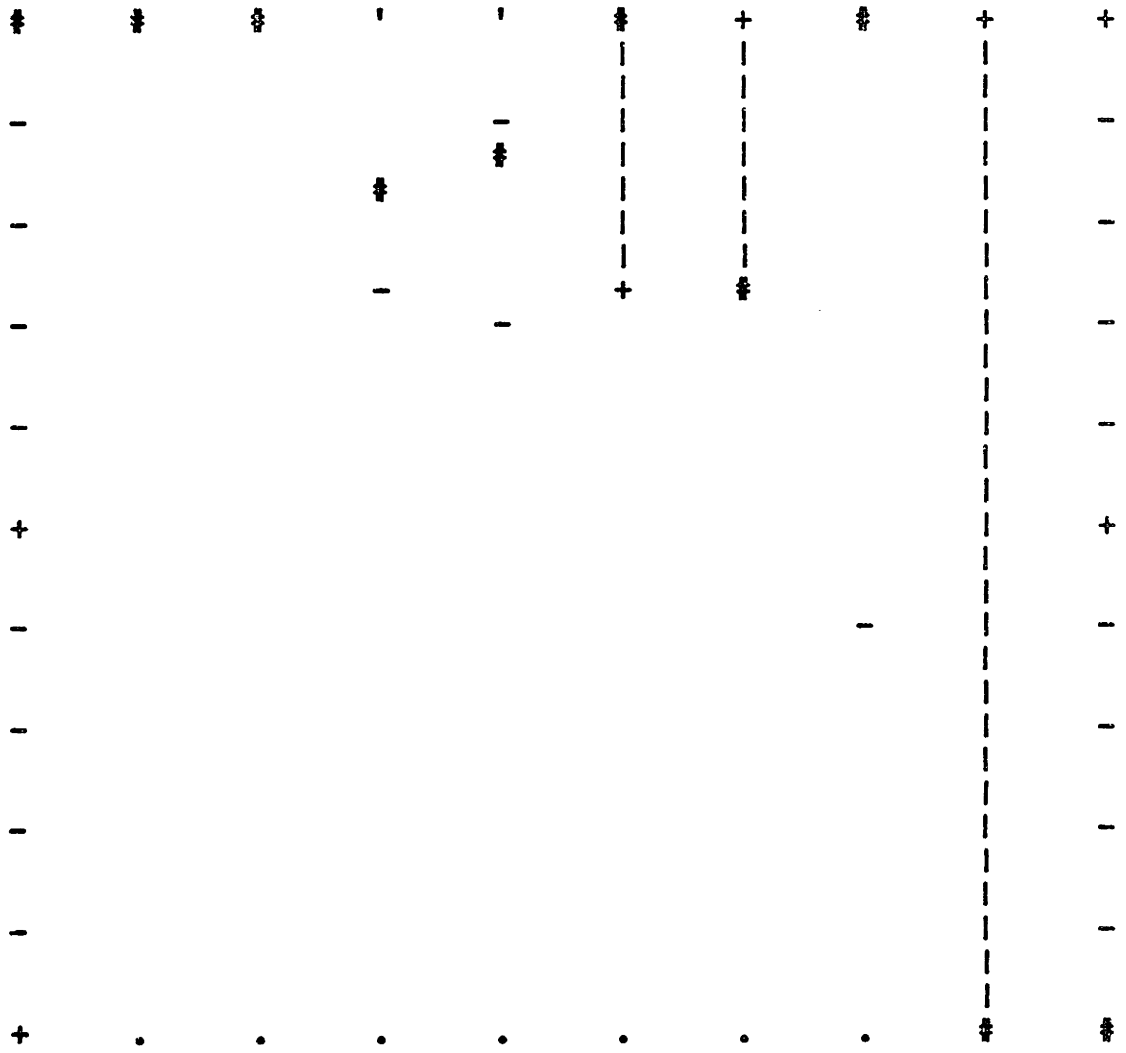
Figure 3.1.2.3: Plot of Bk versus k
Set B1, Complete Linkage, 50 Samples



3.1.3. Data Set A1, Other Methods

Section 3.1.1 contained the plots of Bk for 50 samples when complete linkage was the method used to form the trees. Next are the results when single linkage, average linkage, and the kth nearest neighbor algorithm are used.

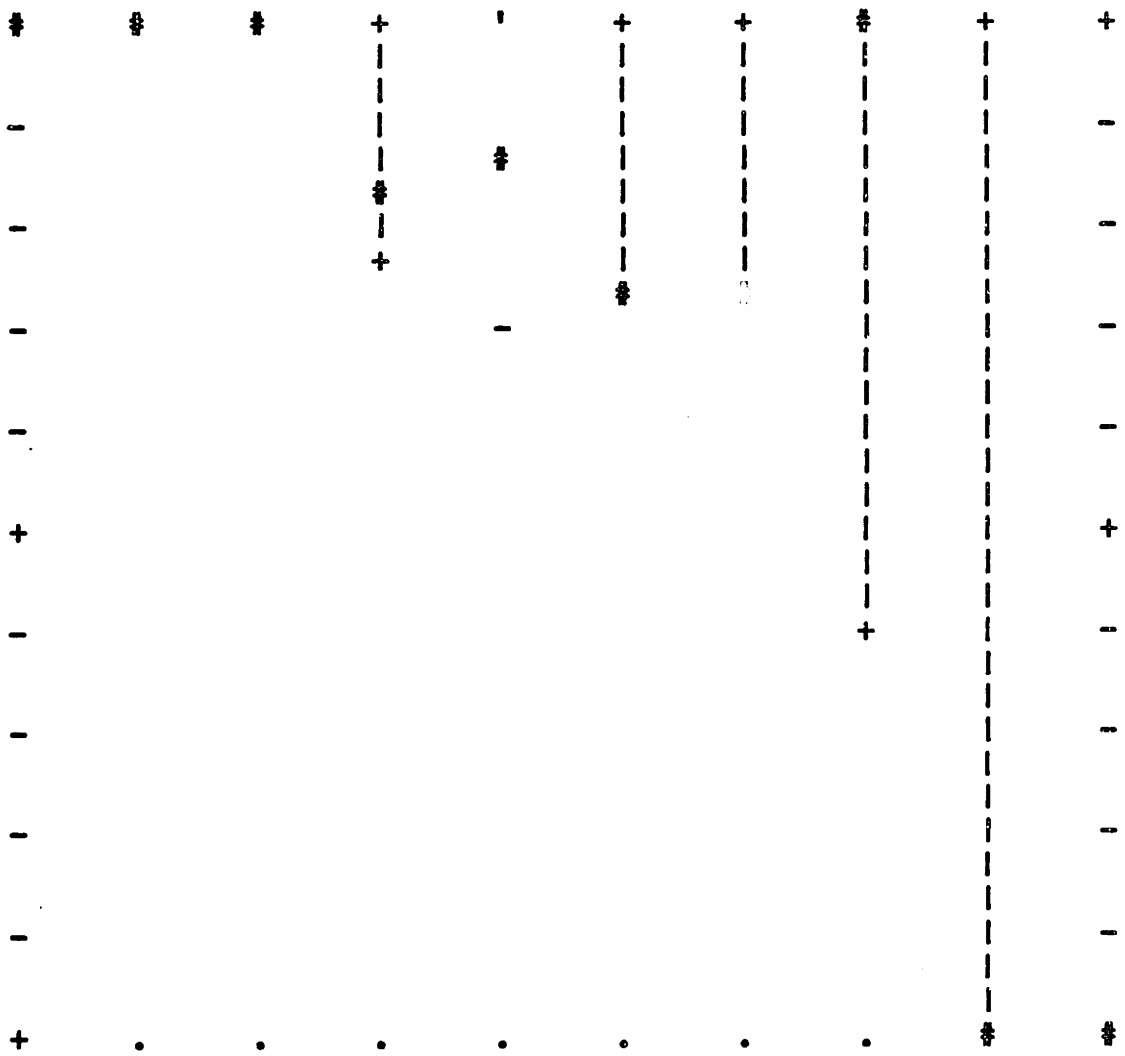
Figure 3.1.3.1: Plot of Bk versus k
Set A1, Single Linkage, 50 Samples



The Bk's for single linkage are typically higher, especially for high values of k near N, the number of objects. Complete linkage finds the stable clusters at lower values of k, near 1. In their experiments, Fowlkes and Mallows also see that the complete linkage decayed faster than single linkage. This is easily explained by the fact that single linkage is "continuous" while complete is known

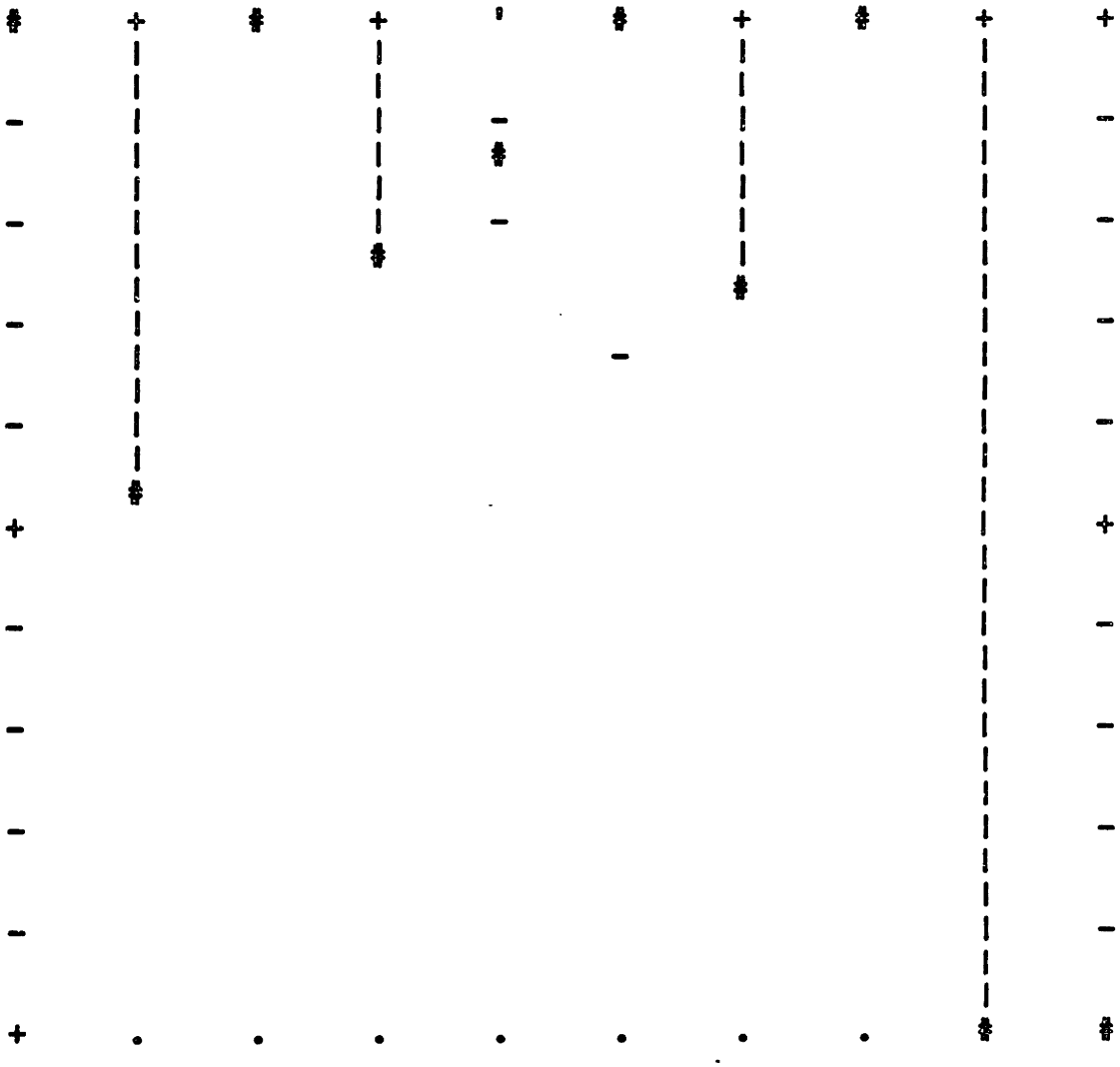
to be "discontinuous". Discontinuous means that drastic changes in the result can come from small change in the data. However, single linkage has its problems, too. It is sensitive to the small distances which can lead to "chaining", or linking clusters earlier than they should normally be linked.

Figure 3.1.3.2: Plot of Bk versus k
Set A1, Average Linkage, 50 Samples



By using the average distance between objects in clusters instead of the minimum or maximum, chaining and discontinuity might be avoided. As one would expect, Bk with average linkage (figure 3.1.3.2) usually ends up between the single and complete linkage results.

Figure 3.1.3.3: Plot of Bk versus k
Set A, Nearest Neighbor (2) Linkage, 50 Samples



In the plot above, it is shown that the kth nearest neighbor method failed to identify the two population clusters, indicated by low Bk's at k = 2, although it did find stable clusters for k = 3. The reason for this confusion is that the kth nearest heighbor method is oriented towards picking out the number of clusters and the clusters themselves and not towards the reproducing whole tree. Here,

3 is its choice of the number of clusters and that is a perfectly good answer. Thus, it is not appropriate to evaluate the kth nearest neighbor method by how well it reproduces the tree nor use it together with Bk. One additional point, this method is not designed for a small set of objects but more for Euclidean data with more than 100 objects. When other parameters besides 2 were used, reproduction of the tree was worse than for 2. For these reasons, we will not consider this method for the remainder of the study.

3.1.4. Data Set B1, Other Methods

Recall figure 3.1.2.3. The graph showed stable clustering at k = 9, when only clusters A, B, and C are linked. The graph of Bk would tell us that the D's and E's are not really clusters. However, by simply looking at the tree, figure 3.1.2.1, one might be mislead into declaring that there were 4 clusters. Complete linkage says 4 clusters in unstable while single linkage, above, shows stable clusters for 2, 3, 4, 5, 6, 7 and 9 clusters! Average linkage has stable clusters at 2, 3, 4, and 5, and of course 9. It is difficult to say which method is more correct.

Figure 3.1.4.1: Plot of Bk versus k
Set B1, Single Linkage

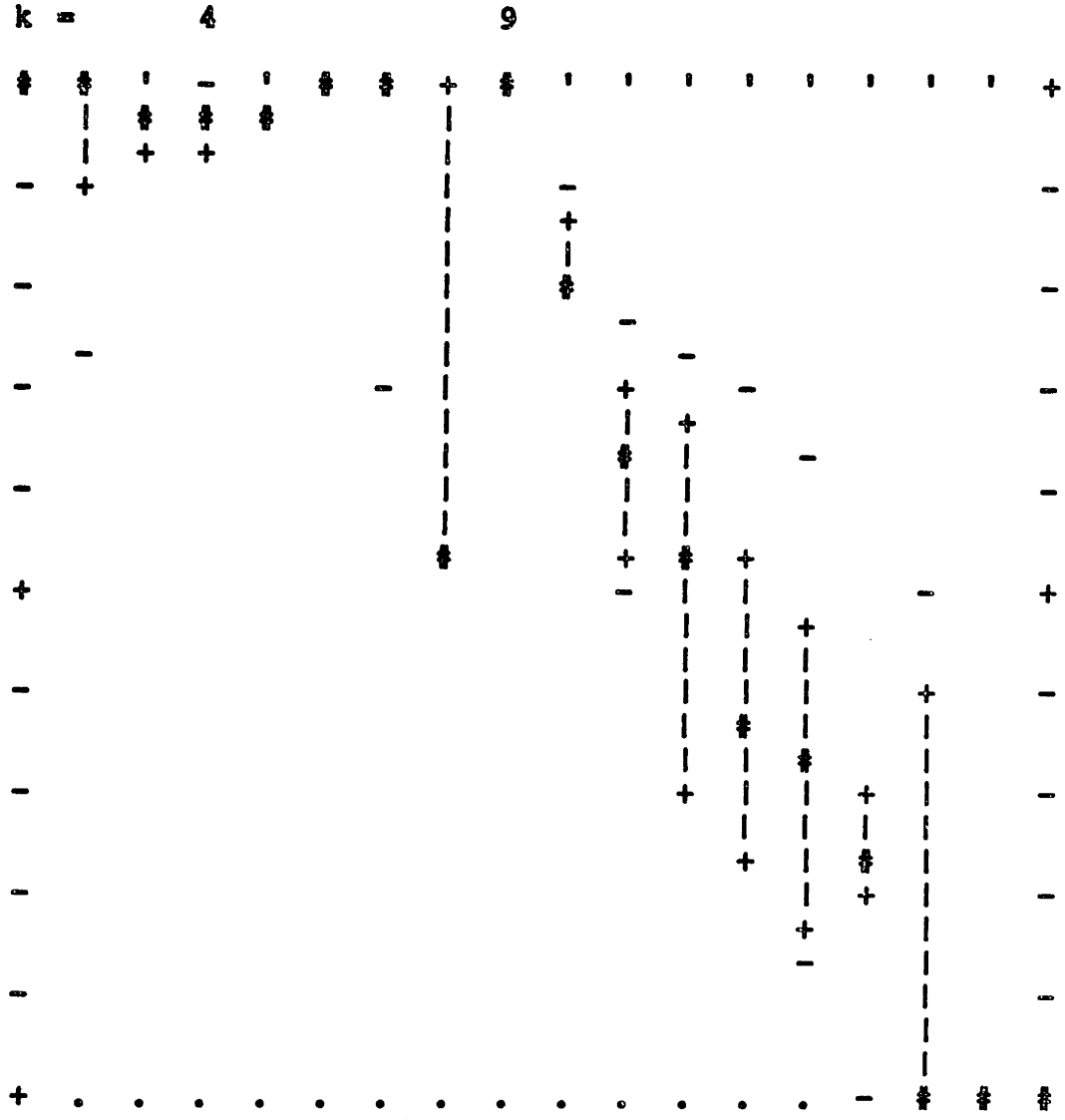
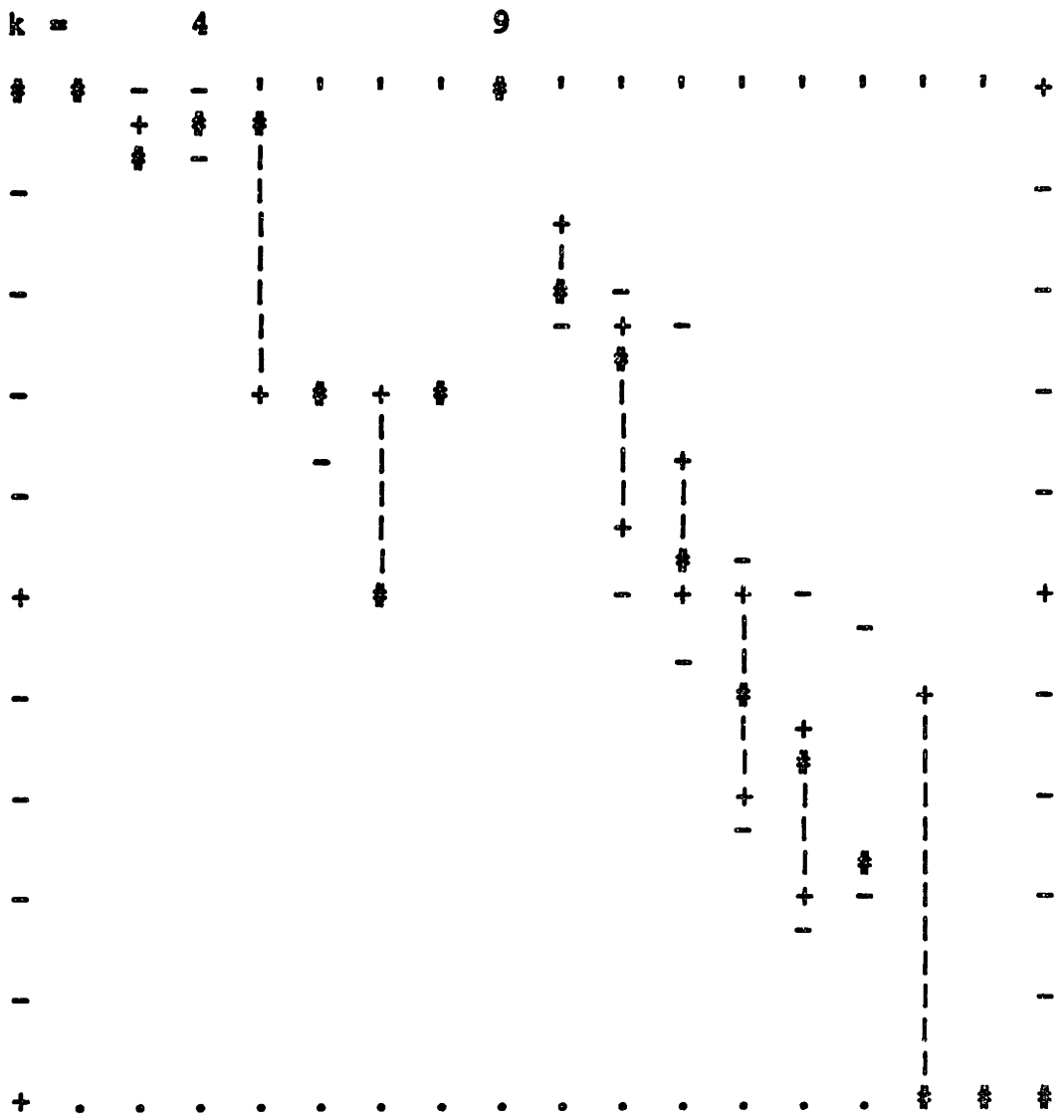


Figure 3.1.4.2: Plot of Bk versus k
Set B1, Average Linkage, 50 Samples



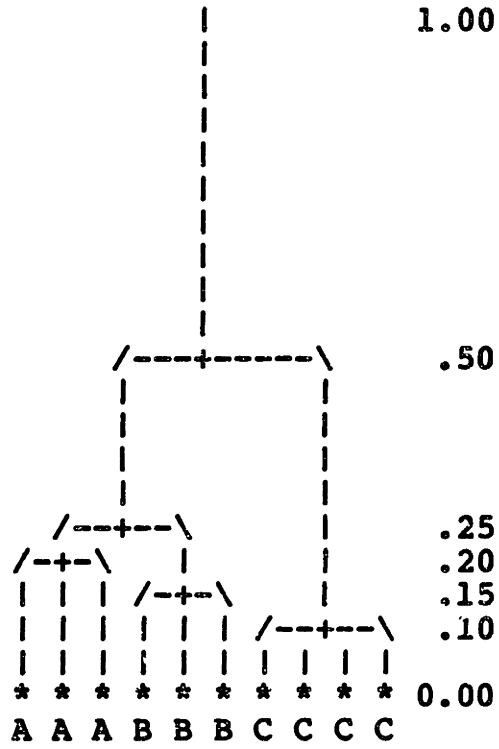
3.1.5. Data Sets A2 and A3, Complete Linkage

Set A1 is an unusually nice set with well separated clusters. By changing the distances but conserving the topology of the tree, set A2 is formed with less distinct clusters.

Table 3.1.5.1: Distances for Data Set A2

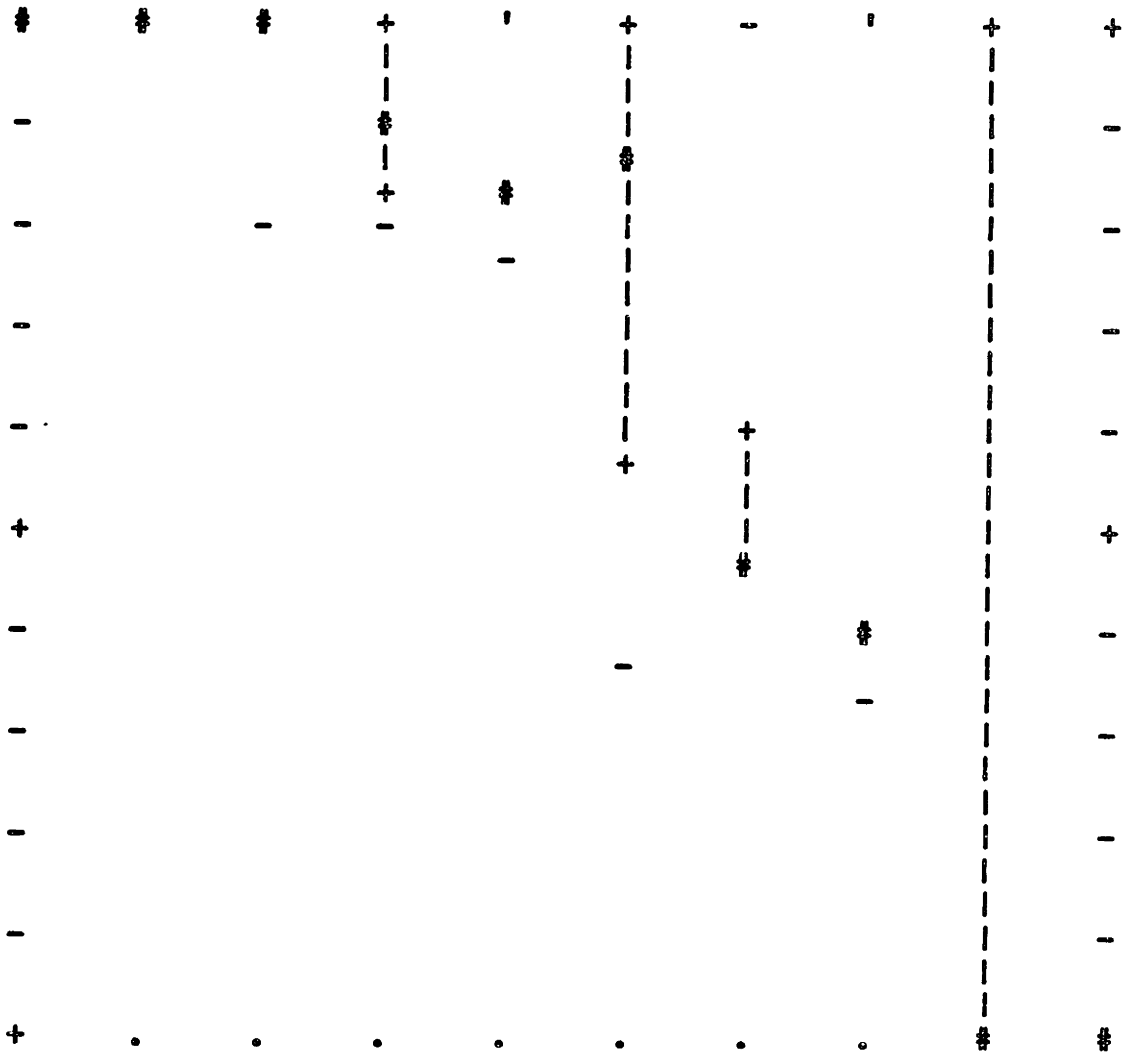
	A1	A2	A3	B1	B2	B3	C1	C2	C3	C4
A1 -		.20	.20	.25	.25	.25	.50	.50	.50	.50
A2 -			.20	.25	.25	.25	.50	.50	.50	.50
A3 -				.25	.25	.25	.50	.50	.50	.50
B1 -					.15	.15	.50	.50	.50	.50
B2 -						.15	.50	.50	.50	.50
B3 -							.50	.50	.50	.50
C1 -								.10	.10	.10
C2 -									.10	.10
C3 -										.10

Figure 3.1.5.2: Tree for Data Set A2



In figure 3.1.5.3, notice that fewer Bk's for k = 3 are perfectly equal to one. Complete linkage is not getting the same 3 clusters consistently this time because the A's and B's are more easily confused. Also, Bk was slightly higher for k = 4.

Figure 3.1.5.3: Plot of Bk versus k
Set A2, Complete Linkage



The stability of clusters with this data not only depends on the difference in distances, e.g. .20 to .25 in set A2, but also the variation that sampling will produce. The maximum variance in a binomial random variable occurs when the probability is .50. So for data set A3, pictured below, the real difference between links at .50 and .55 is less than those in set A2 at .20 and .25, even though the

arithmetic difference, .05, is the same. This property is something that a data analyst could easily forget while simply looking at a single tree, especially when the variation is not well known. As expected, the Bk plot for set A3 with complete linkage has even less stablility at 3 clusters.

Figure 3.1.5.4: Tree for Data Set A3

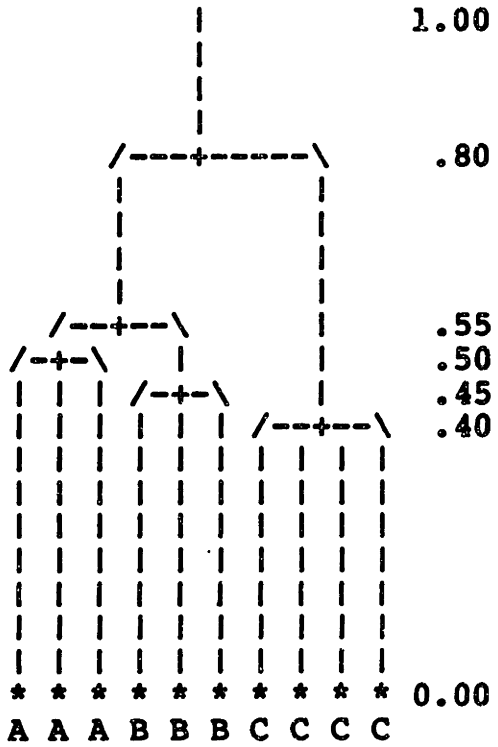
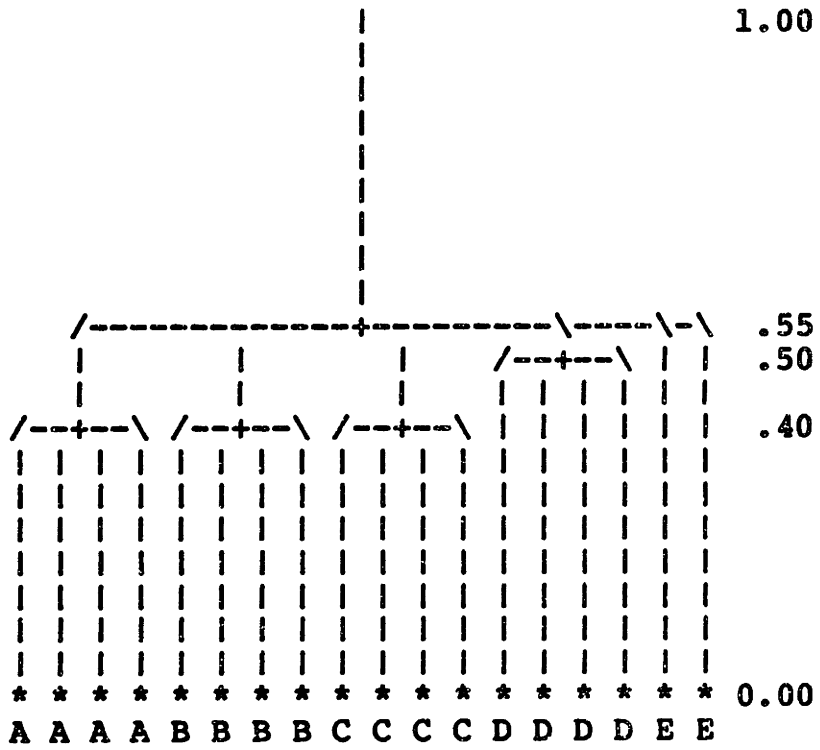


Figure 3.1.5.5: Plot of Bk versus k
Set A3, Complete Linkage

Because we can not define a true null distribution, it is difficult to say exactly how one could tell whether a peak is significant or not. It would be easy to make an arbitrary cut off point at, say, .90 or .85; but as k increases, this may not be appropriate since the Bk plot must eventually decay. Perhaps this question of significant peaks can only be answered after a few hundred plots of experience.

3.1.6. Data Set B2

Figure 3.1.6.1: Data Set B2



Data set B2 is a less stable version of B1. Average and complete linkage find stable clusters at k = 9 while single linkage does not. However, these 9 clusters are not as stable as those is Set B1, as indicated by the lower values of Bk.

Figure 3.1.6.2: Plot of Bk versus k
Set B2, Complete Linkage

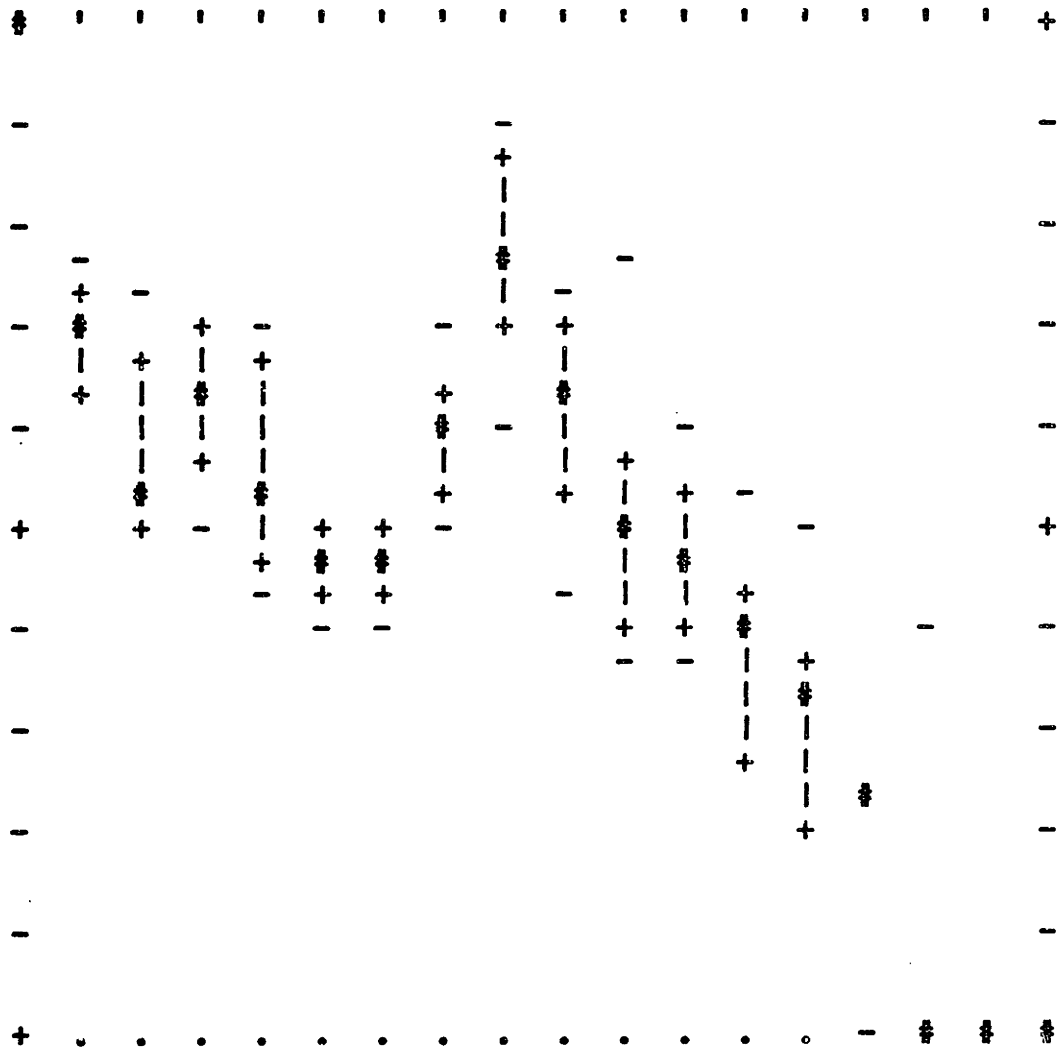
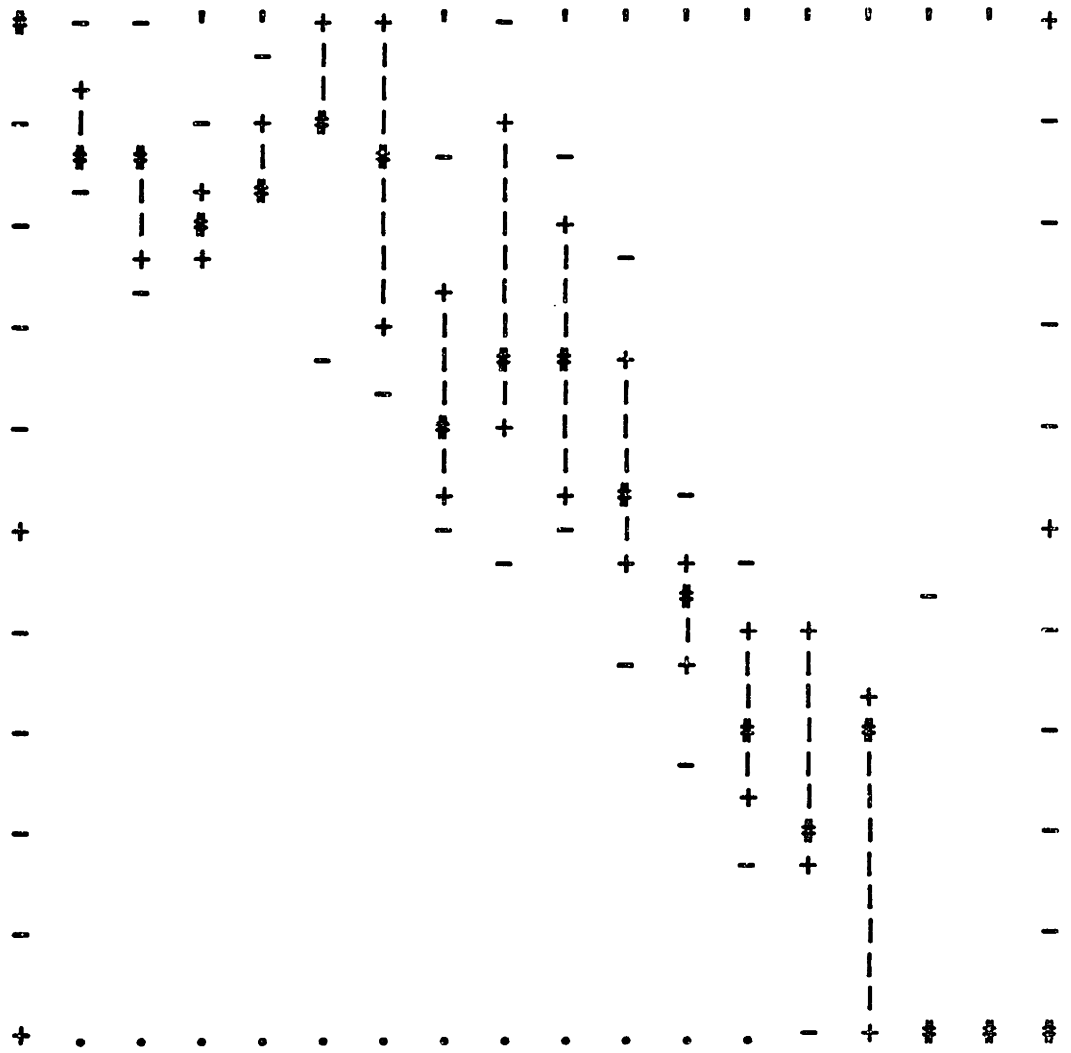
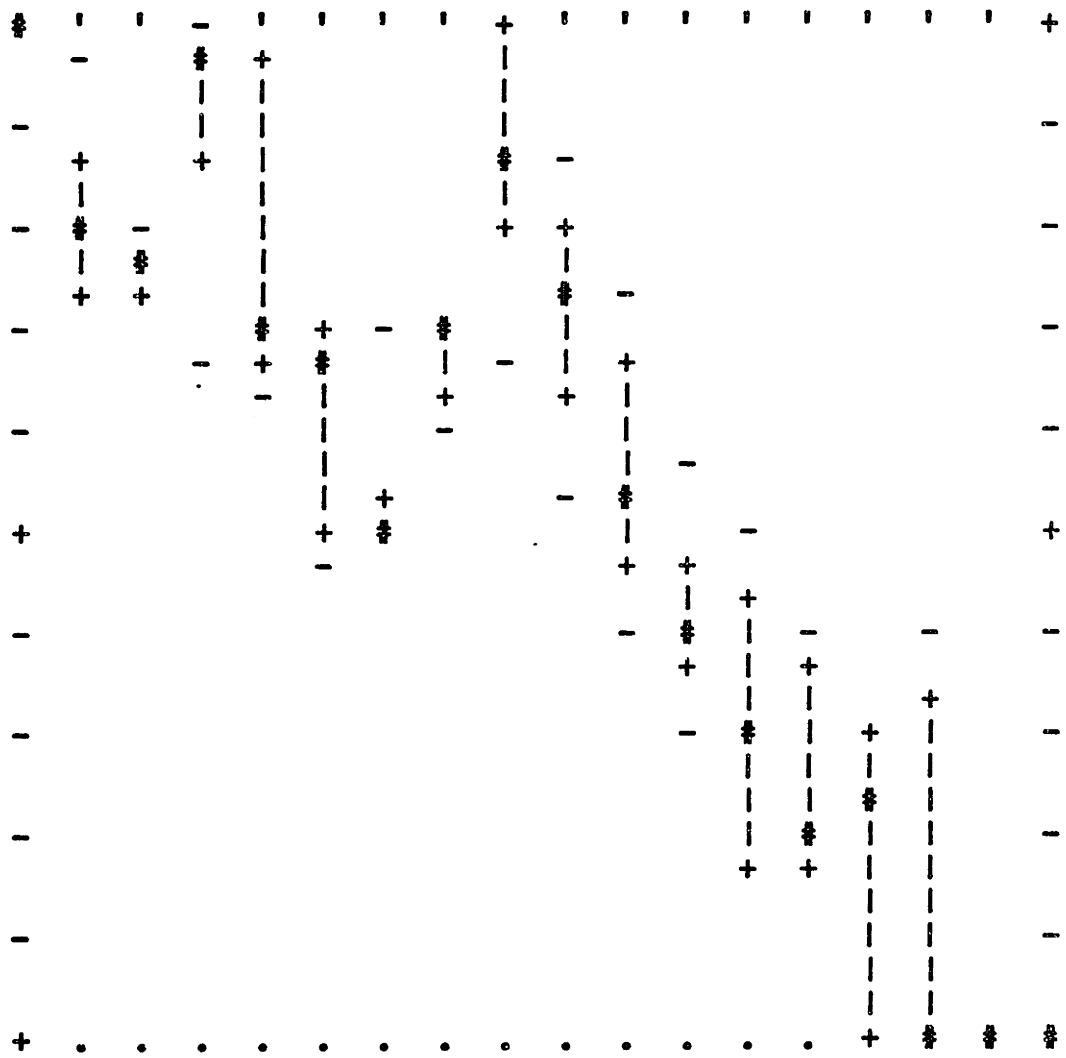


Figure 3.1.6.3: Plot of Bk versus k
Set B2, Single Linkage



Notice that average linkage (figure 3.1.6.4) has stable clusters at 4 which is odd because set B2 does not. It also has stable clusters at k = 9.

Figure 3.1.6.4: Plot of Bk versus k
Set B2, Average Linkage



So we have shown that the Bk plot is useful in finding stable and relevant clusters. The problem is that we were taking many samples from a known truth while in practice we get only one sample and no truth. Now the question is "Can one find a reliable estimator of the true Bk between the sample and population trees?"

3.2. THE DISTRIBUTION OF BK

Since we do not know the population in practice, we must use subsampling techniques like the bootstrap to simulate sampling from the true distribution. As stated above, the bootstrap uses the sample as an approximation of the population.

In order to see if the distribution of the sample-to-bootstrap Bk was actually close to that of the true-to-sample Bk, 50 samples were taken from the true distances and some of these samples were chosen and 50 bootstrap-samples were drawn from each of them. Then the sample Bk could be compared with the bootstrapped Bk's to see if the bootstrap reproduced the original situation well. To compare the distributions, a standard chi-squared goodness of fit test was used. The null hypothesis is that the distributions are the same and a low value of the chi-squared statistic, relative to the degrees of freedom, will indicate that the null hypothesis is acceptable.

Binning was based on the distribution of the true-to-sample Bk's. Those Bk's from the bootstraps were binned with the closest sample Bk. For complete linkage and 50 bootstraps on each of 6 samples from set A1, the results are in Table 3.2.1.

Table 3.2.1: Results of Chi-squared Tests
Set A1, Complete Linkage, Samples 1 through 6

k	df	chisq
4	2	61.879
4	2	16.425
4	2	6.901
4	2	32.516
4	2	5.014
4	2	5.209

95th percentile = 5.991

5	4	64.729
5	4	241.342
5	4	195.578
5	4	147.015
5	4	306.318
5	4	71.061

95th percentile = 9.488

6	1	7.219
6	1	.357
6	1	.357
6	1	7.219
6	1	2.228
6	1	7.219

95th percentile = 3.841

7	3	106.452
7	3	105.390
7	3	20.121
7	3	93.056
7	3	122.238
7	3	2.400

95th percentile = 7.814

8	1	8.000
8	1	18.000
8	1	5.120
8	1	.720
8	1	11.520
8	1	6.480

95th percentile = 3.841

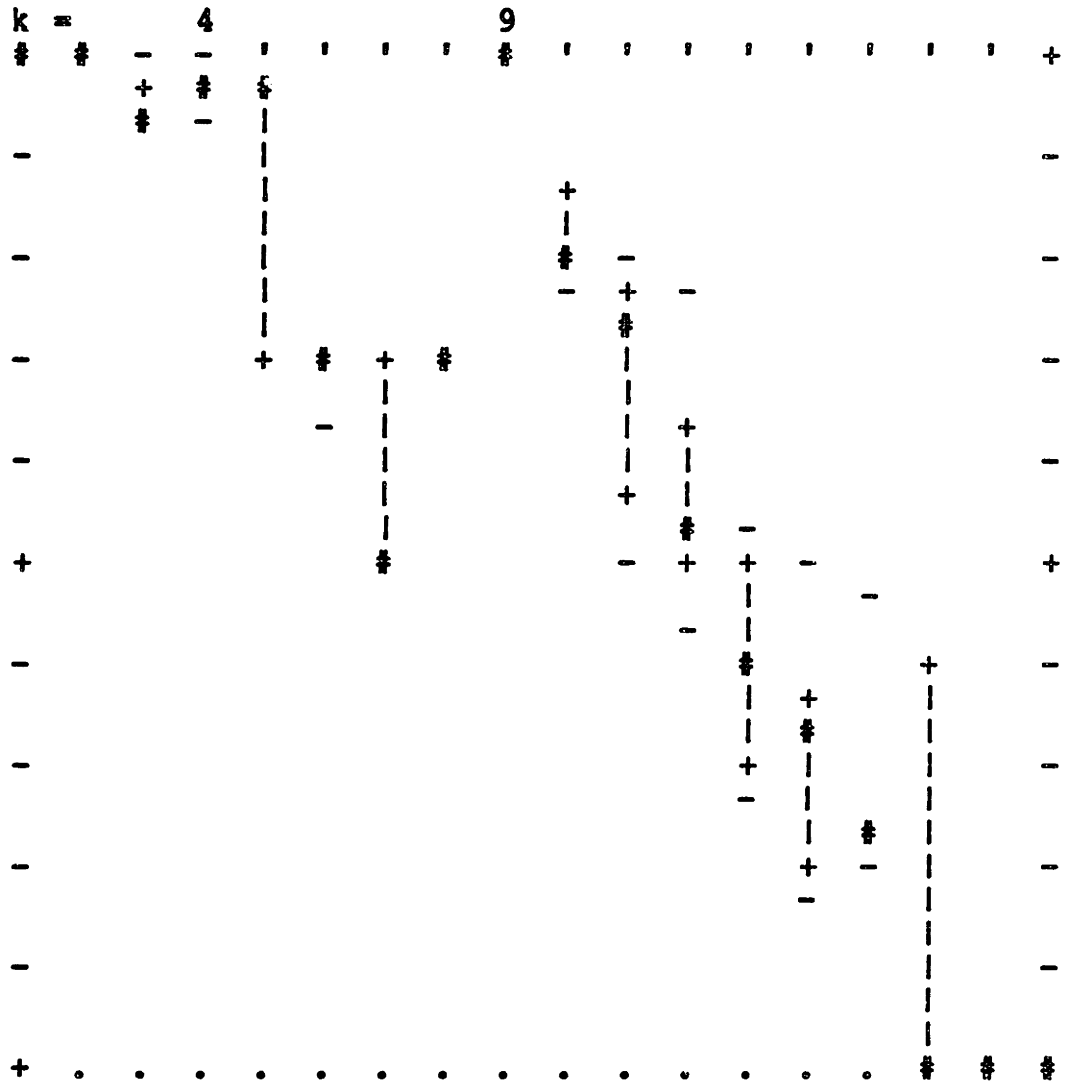
9	1	.750
9	1	4.083
9	1	21.333
9	1	.750
9	1	.000

9 1 33.333
95th percentile = 3.841

The results show that the distribution of the sample Bk is not close to that of the true. The cases when k = 2 or 3 could not be tested because the sample distributions were degenerate. Even with different binning schemes, the results were still the same: the distributions are different. This was also true for all three clustering methods. The difference was that the bootstrapped Bk's were nearly always shifted from the sampled. Usually, it was shifted to the lower side, which is what one would expect.

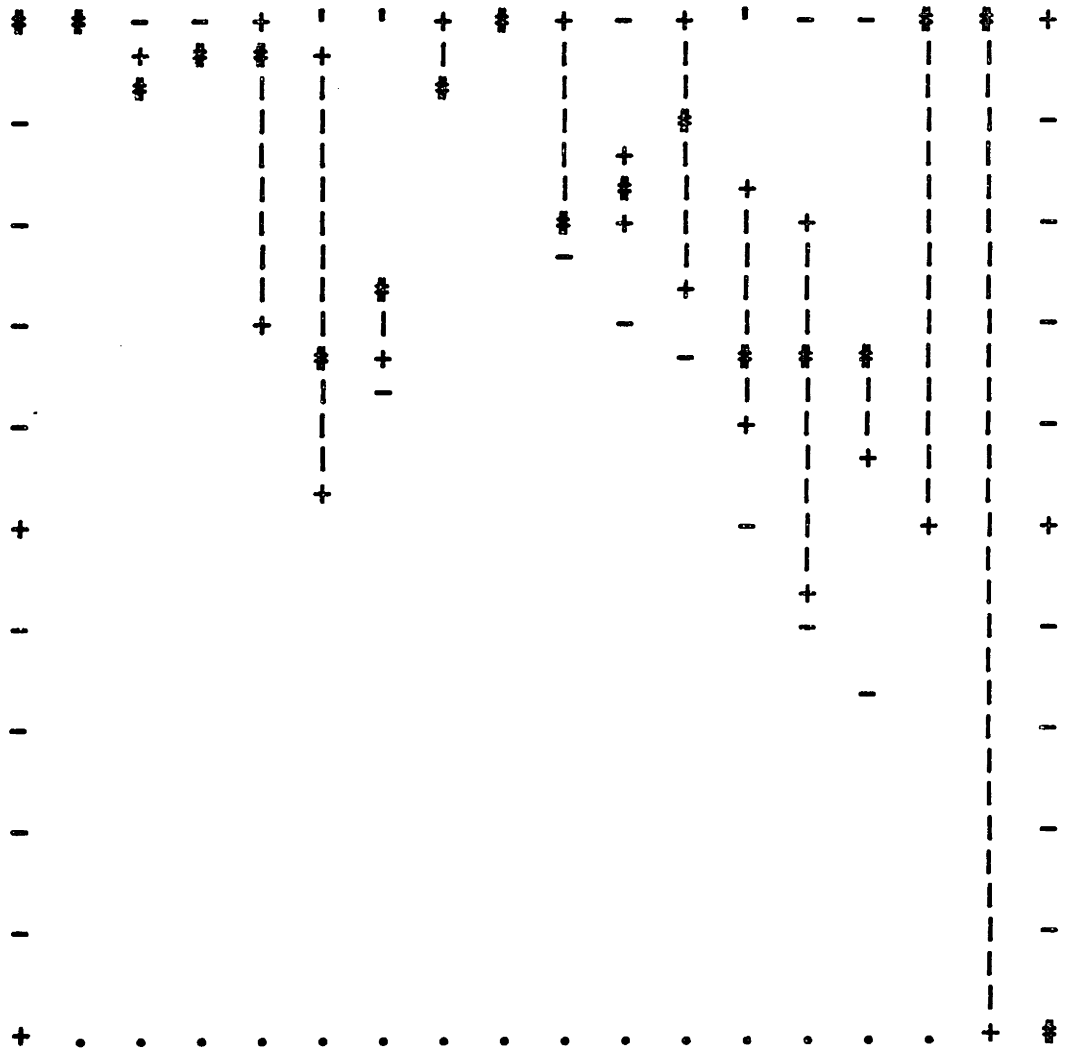
This difference is made clear by the two plots shown below. The Bk plot in figure 3.2.3 compares the 50 bootstrap samples (based on sample number 33) with sample number 33 itself, while the one in figure 3.2.2 compares the 50 original samples with the population.

Figure 3.2.2: Plot of Bk versus k
Set B1, Average Linkage, 50 Samples



For data sets with 18 objects, e.g. set B1, the situation was the same. So the problem was not simply caused by the small number of objects. Even average linkage, the least sensitive method, had bootstrap Bk's that were very different from the sample.

Figure 3.2.3: Plot of Bk versus k
Set B1, Average Linkage, 50 Bootstrap Samples



These results are not particularly good news. It means that the empirical distributions defined by the samples can be very different from the true distribution. In these cases, it was always different. However, all is not lost. Notice that the median of the bootstrap Bk's is equal to the median of the sample Bk's for k = 2, 3, 4, 5, 9, and 10 above. Even though the distributions of Bk's are not the

same, some statistics like the median may be very close to the original. The purpose of section 3.3 is to examine how close it really is.

3.3. ESTIMATORS FOR BK

It may be sufficient to estimate the sample-true Bk rather than its distribution. It is proposed here that estimators can be calculated from bootstrapped Bk's. We looked at 3 natural estimators, the mean, median, and mode, for 50 bootstraps of one sample. The values are given in Table 3.3.1.

Table 3.3.1: Estimators of Bk
Set A1, Complete Linkage

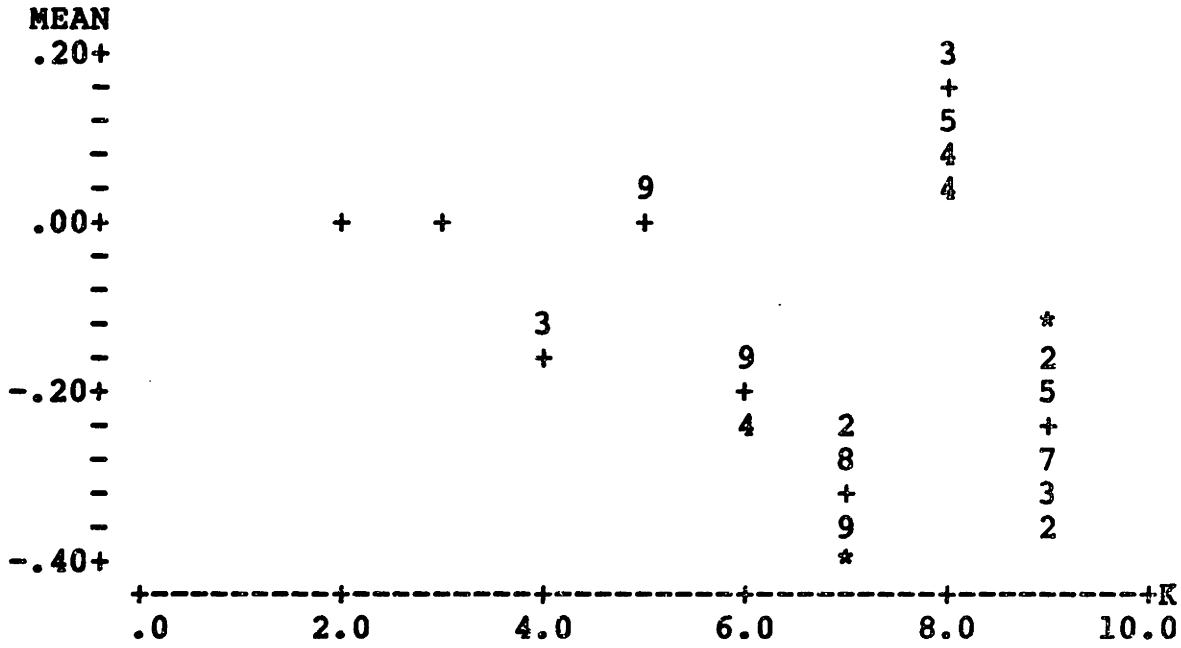
k	true	mean	median	mode(s)
Sample 3				
2	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000
4	.778	.866	.825	.825
5	.857	.816	.857	.857
6	.730	.794	.800	.800
7	.750	.818	1.000	1.000
8	.408	.563	.500	.500
9	.000	.340	.000	.000
Sample 4				
2	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000
4	1.000	.839	.825	.825
5	.857	.774	.772	.772
6	.730	.740	.800	.800
7	1.000	.740	.750	.750
8	1.000	.740	1.000	1.000
9	1.000	.720	1.000	1.000

Most of the time, the median equaled the mode. After examining many trials it appears that the median and mode are either exactly equal to the true Bk, or they are further from it than the mean.

3.3.1. Data Set A1, Complete Linkage

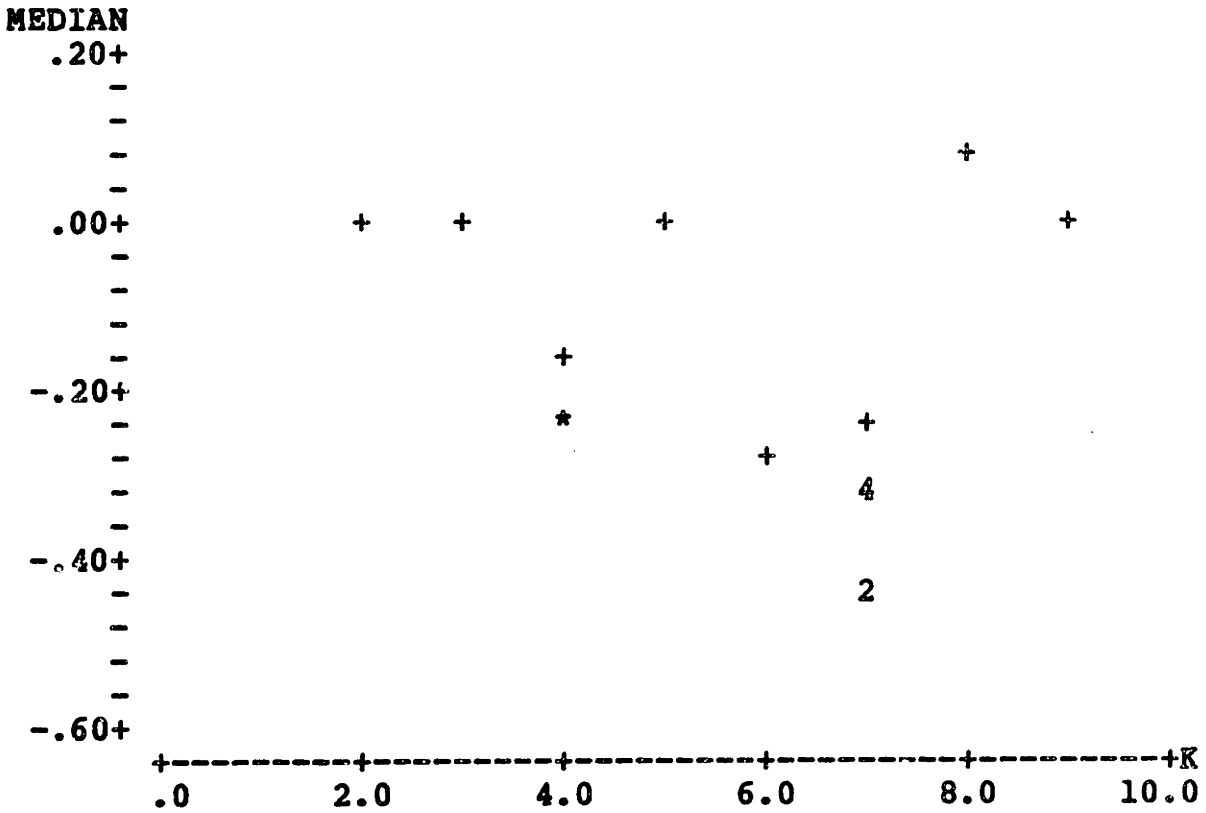
In order to get a better idea on how accurate estimators are, the bootstrap can be used again to estimate the distribution of these statistics. One sample from the true distribution was taken and 30 sets of 50 bootstrap samples from the sample distribution were created. For each set, the mean median, and mode of the 50 boots were calculated and the true Bk was subtracted.

Figure 3.3.1.1: Deviation of Esimator from the True Bk
Data Set A1, Complete Linkage, Sample Number 10



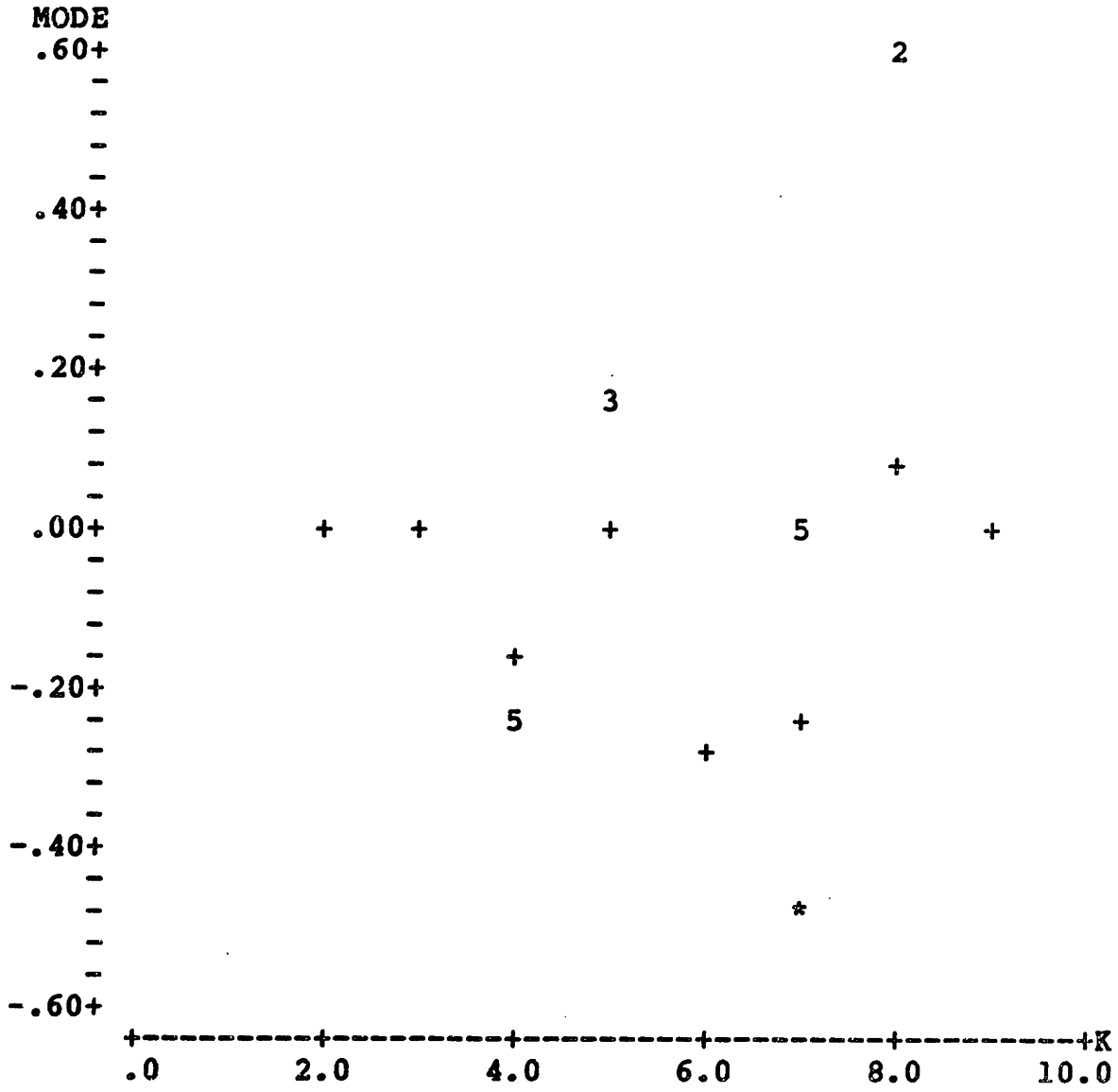
The plots here are of the difference between the bootstrap-estimated Bk and the actual Bk for the three estimators using complete linkage. The perfect result would be when the estimators all line up beside 0.00. These plots were made with the help of the MINITAB statistics package. The plus symbols indicate 10 or more points, while asterisks mean a single point.

Figure 3.3.1.2: Deviation of Estimator from True Bk
Set A1, Complete Linkage, Sample 10



The median appeared better than the mode and mean as an estimator, if only slightly. So we will present only the median in the remainder of the study.

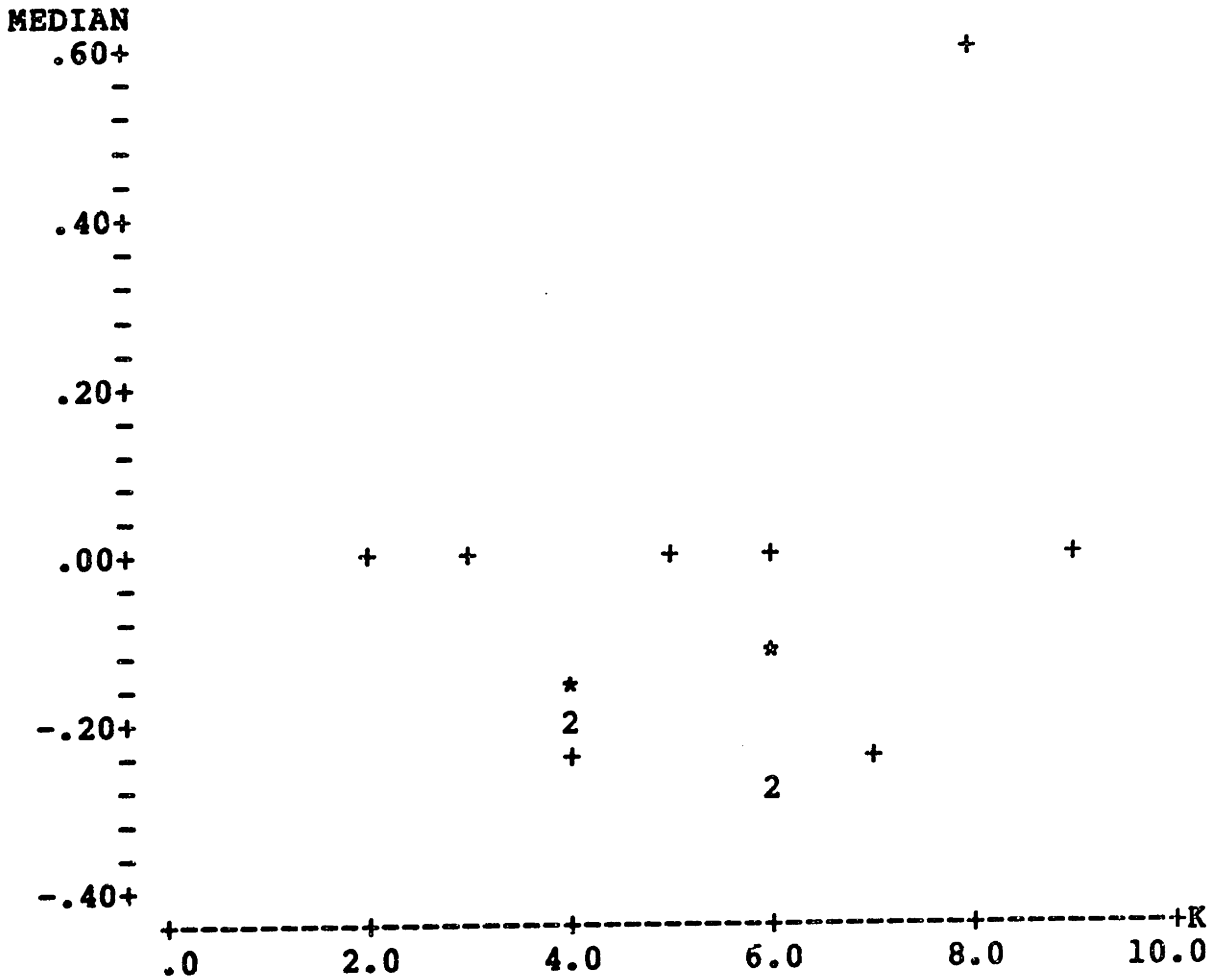
Figure 3.3.1.3: Deviation of Estimator from True Bk
Set A1, Complete Linkage, Sample 10



3.3.2. Data Set A1, Other Methods

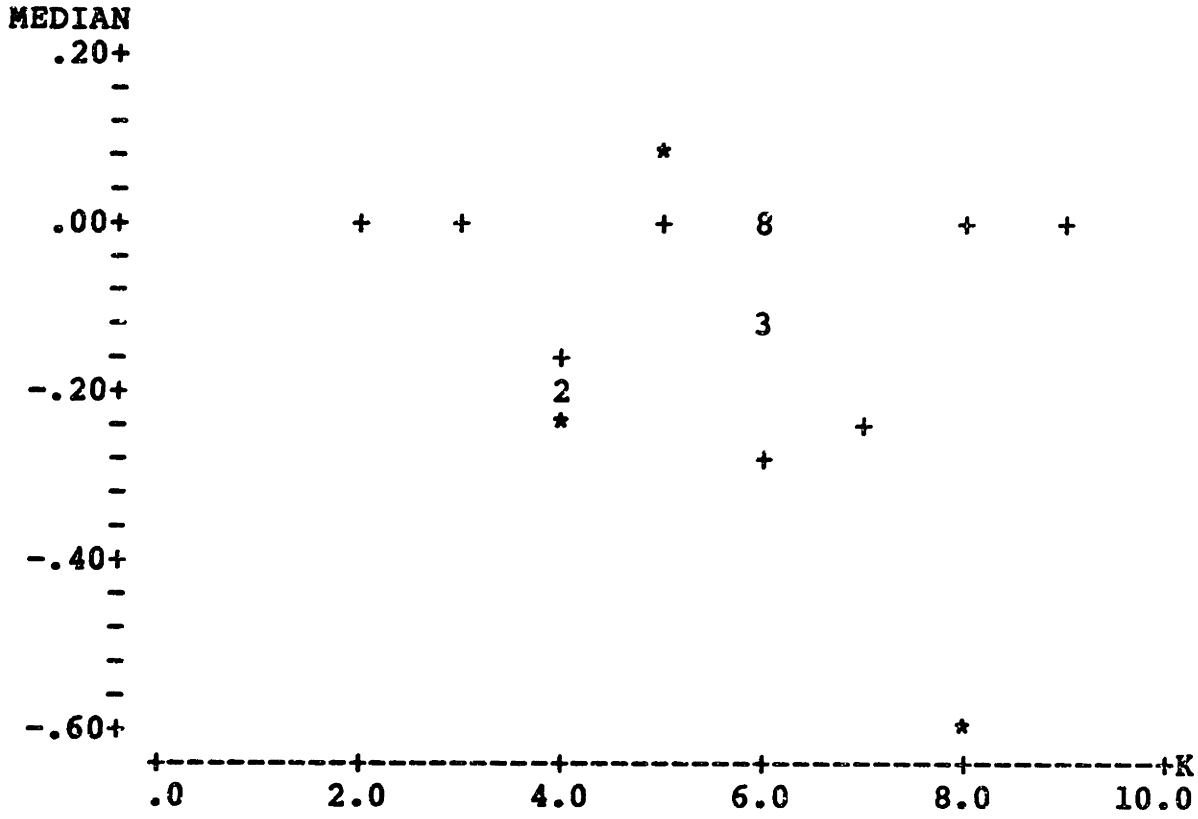
Here we present the results for the same data set using single and average linkage. The median does slightly worse with single linkage than complete while it does slightly better with average linkage.

Figure 3.3.2.1: Deviation of Estimator from True Bk
Set A1, Single Linkage, Sample 10



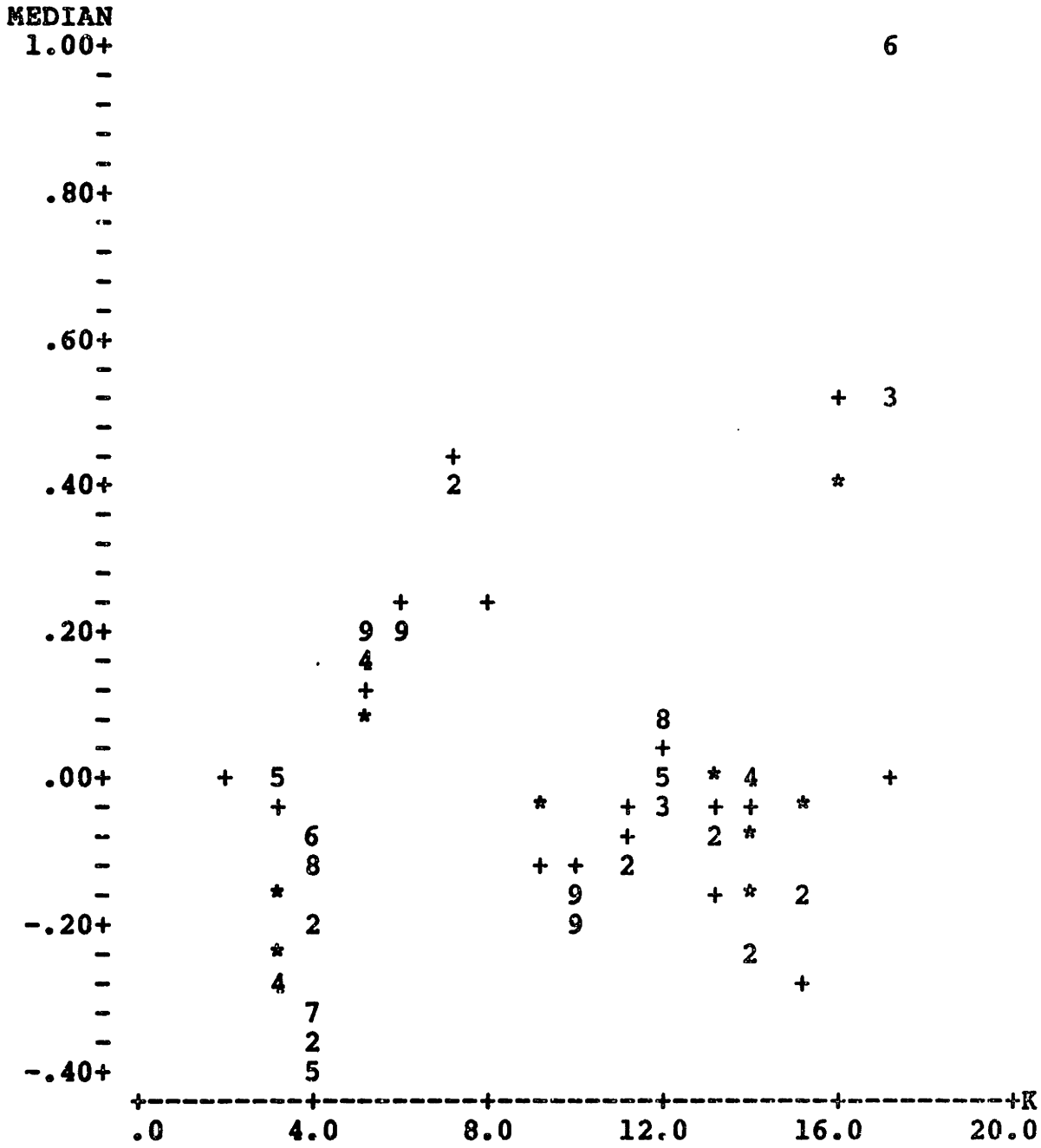
All three methods did perfectly for k = 2, 3, 5, and 9. In this context, it does not mean those are stable clusters. It only means that we accurately estimated the true Bk.

Figure 3.3.2.2: Deviation of Estimator from True Bk
Set A1, Average Linkage, Sample 10



3.3.3. Data Set B1, All Three Methods

Figure 3.3.3.1: Deviation of Estimator from True Bk
Set B1, Complete Linkage, Sample 10



With more objects, estimation becomes less accurate. However, when the true-sample Bk was high, the bootstrap-sample estimator was high also. Recall that for set A1, the estimators were perfect for k equal to 2 and 3. Here, the estimate of Bk for k = 9 is usually accurate.

Figure 3.3.3.2: Deviation of Estimator from True Bk
Set B1, Single Linkage, Sample 10

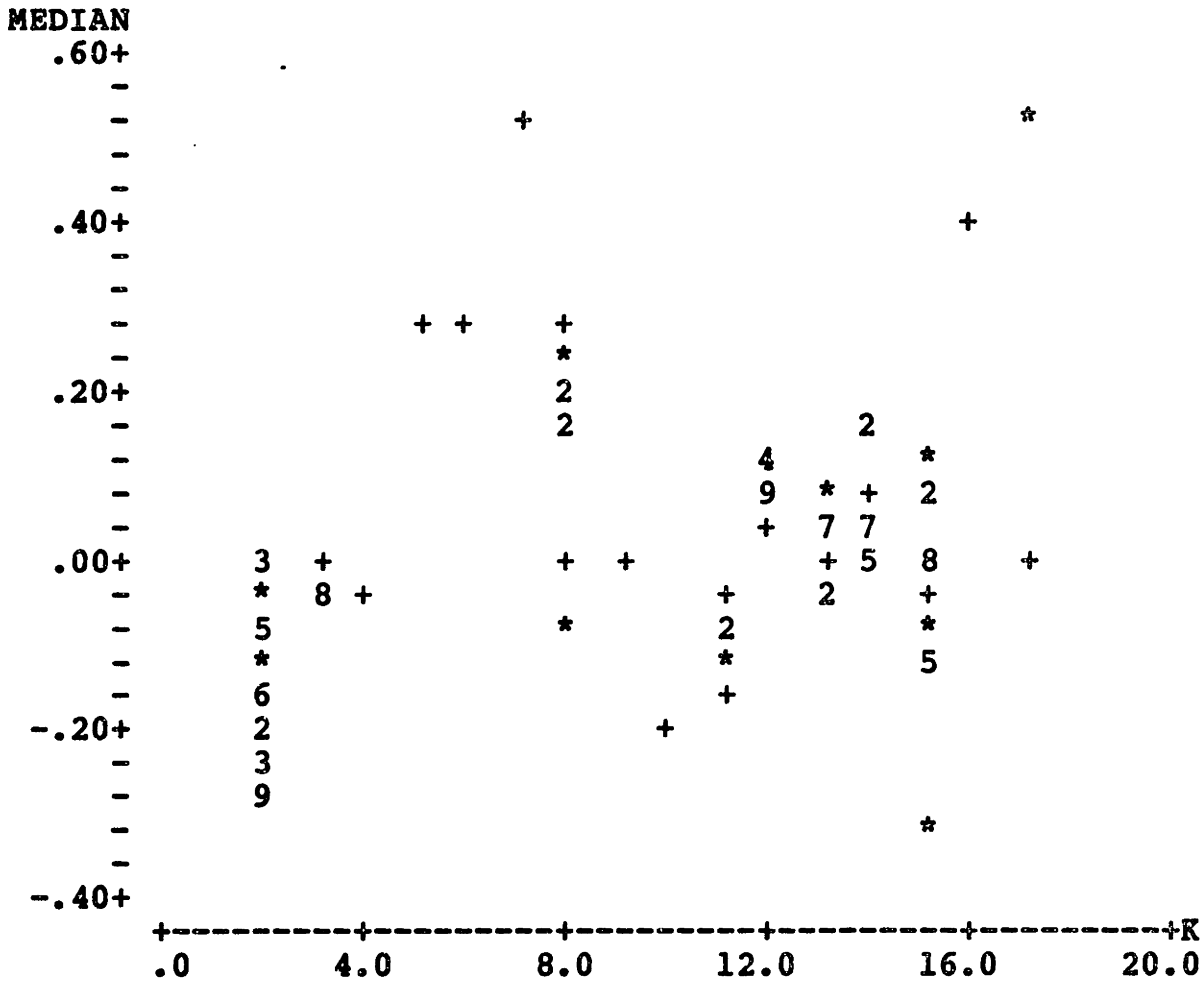
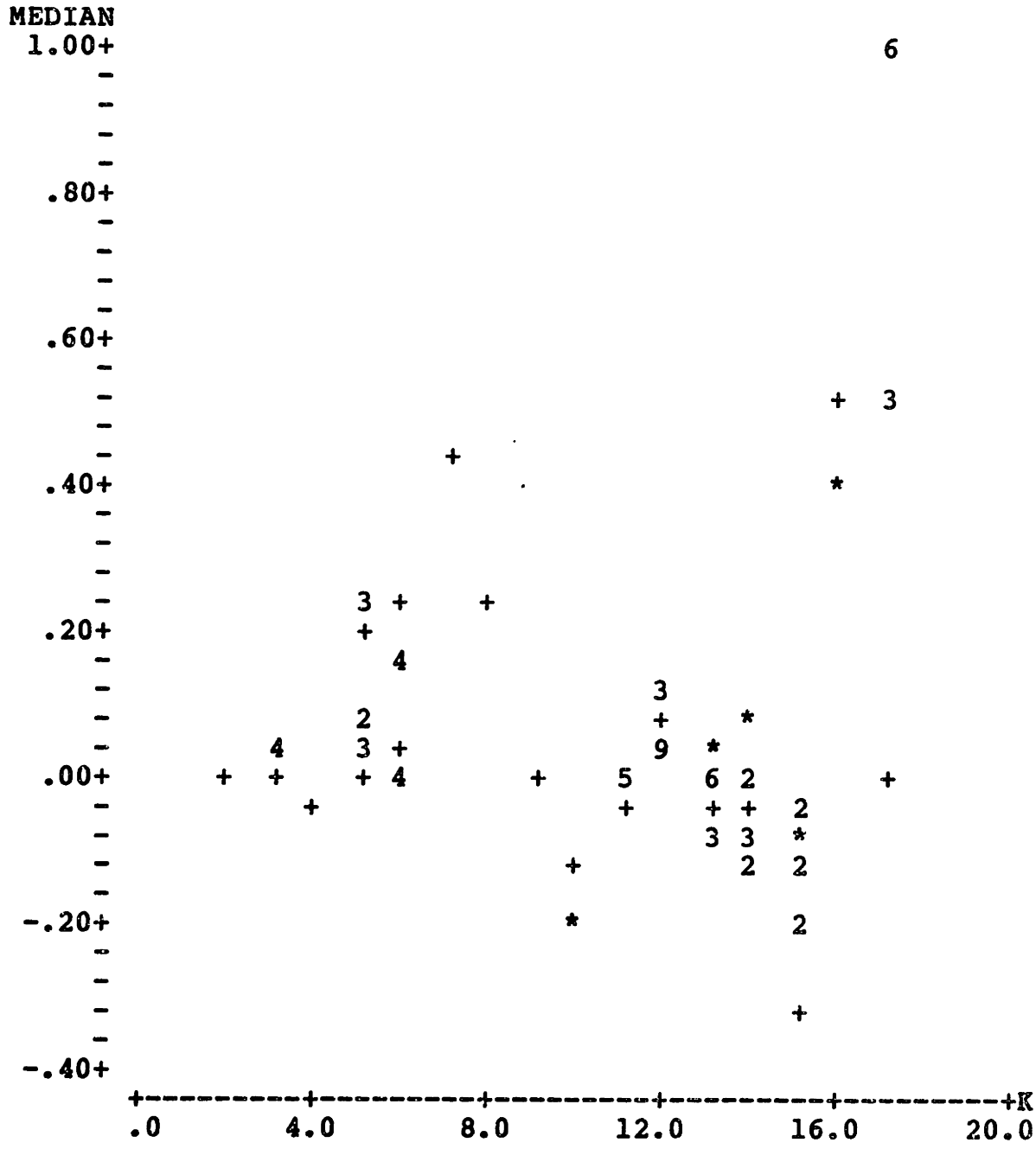


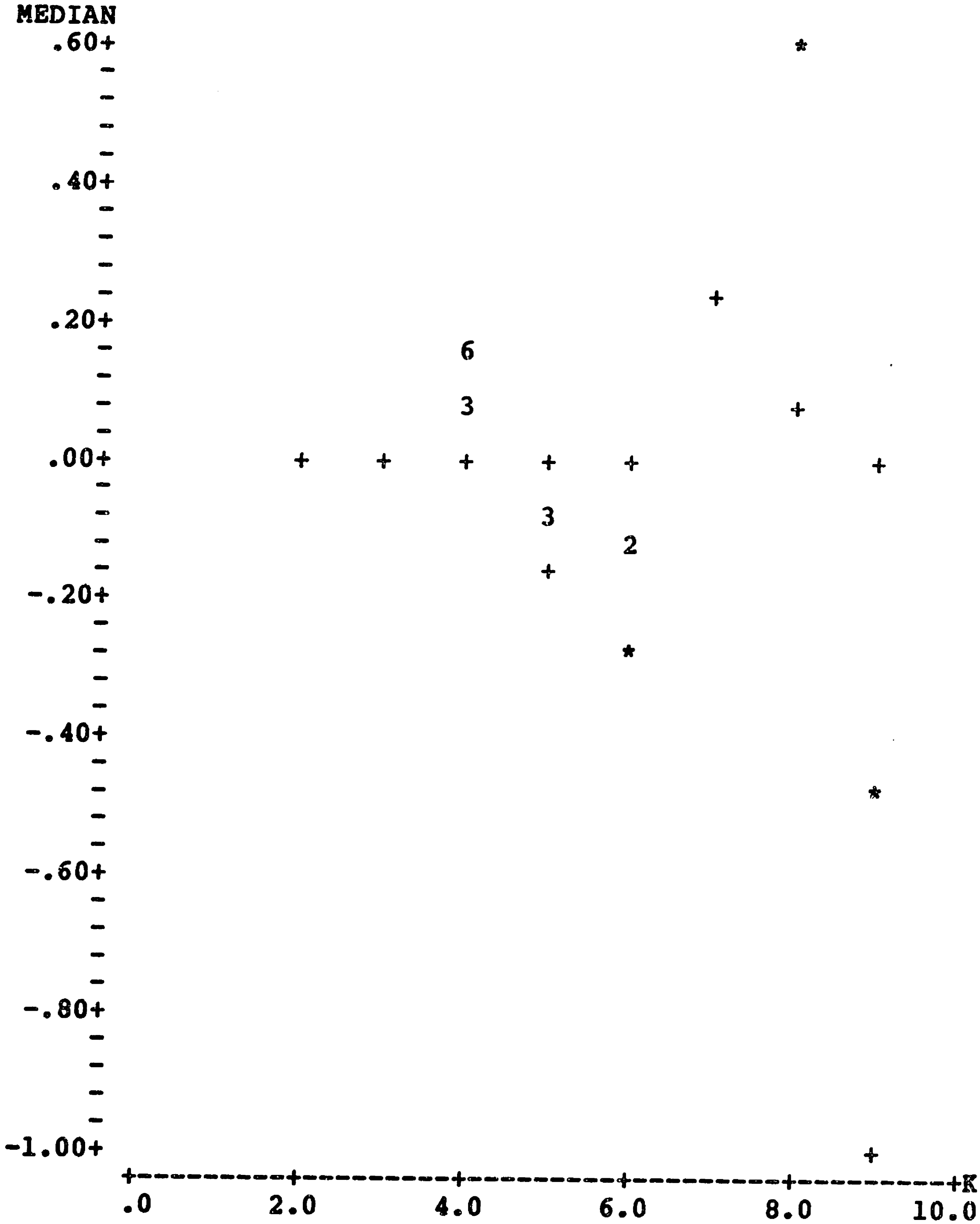
Figure 3.3.3.3: Deviation of Estimator from True Bk
Set B1, Average Linkage, Sample 10



3.3.4. Data Set A1, Other Samples

Most of the results above are based on one sample, the tenth. In order to make sure that the sample is not atypical, the same procedure was run on other samples, numbers 11 through 13.

Figure 3.3.4.1: Deviation of Estimator from True Bk
Set Al, Average Linkage, Sample 11



These plots show that certain statistics on bootstrap samples will be reasonable estimators of the true Bk. Different samples will give different results, however. In any situation where the bootstrap is used, it would be advisable to create and examine many samples from possible models. That way the reliability of the bootstrap in that particular situation would be known.

Figure 3.3.4.2: Deviation of Estimator from True Bk
Set A1, Average Linkage, Sample 12

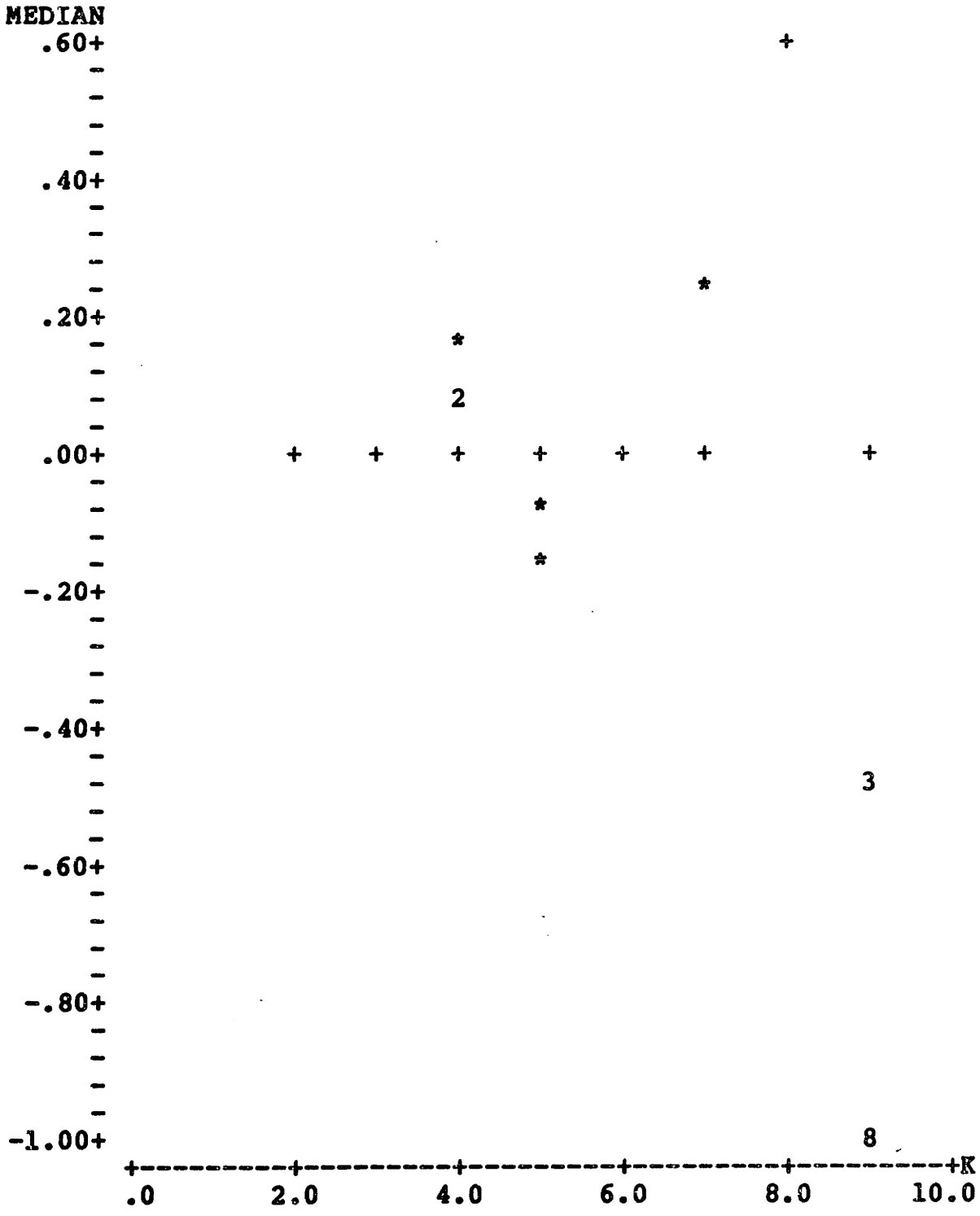
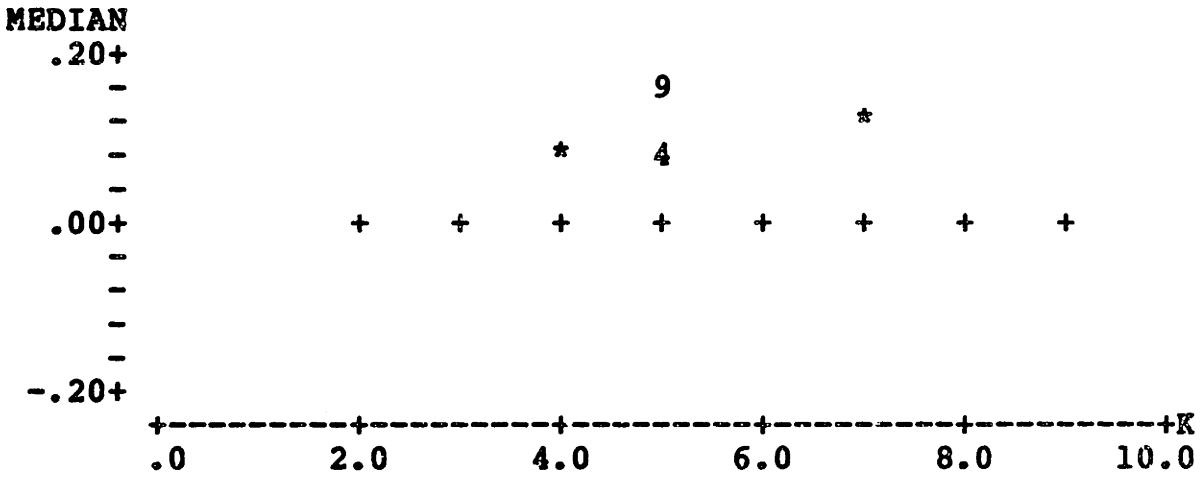
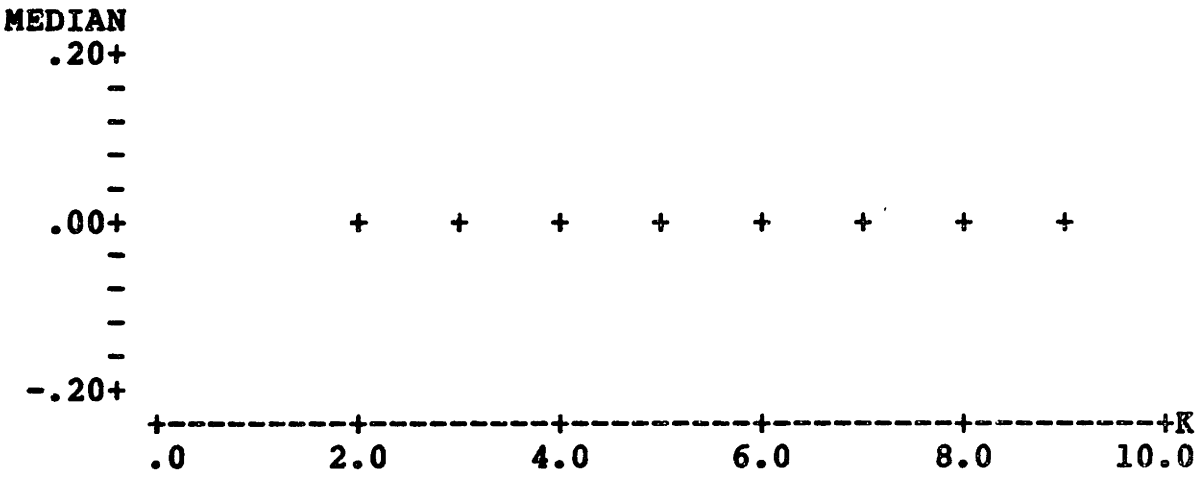


Figure 3.3.4.3: Deviation of Estimator from True Bk
Set A, Average Linkage, Sample 13



Complete linkage worked perfectly on sample number 13.

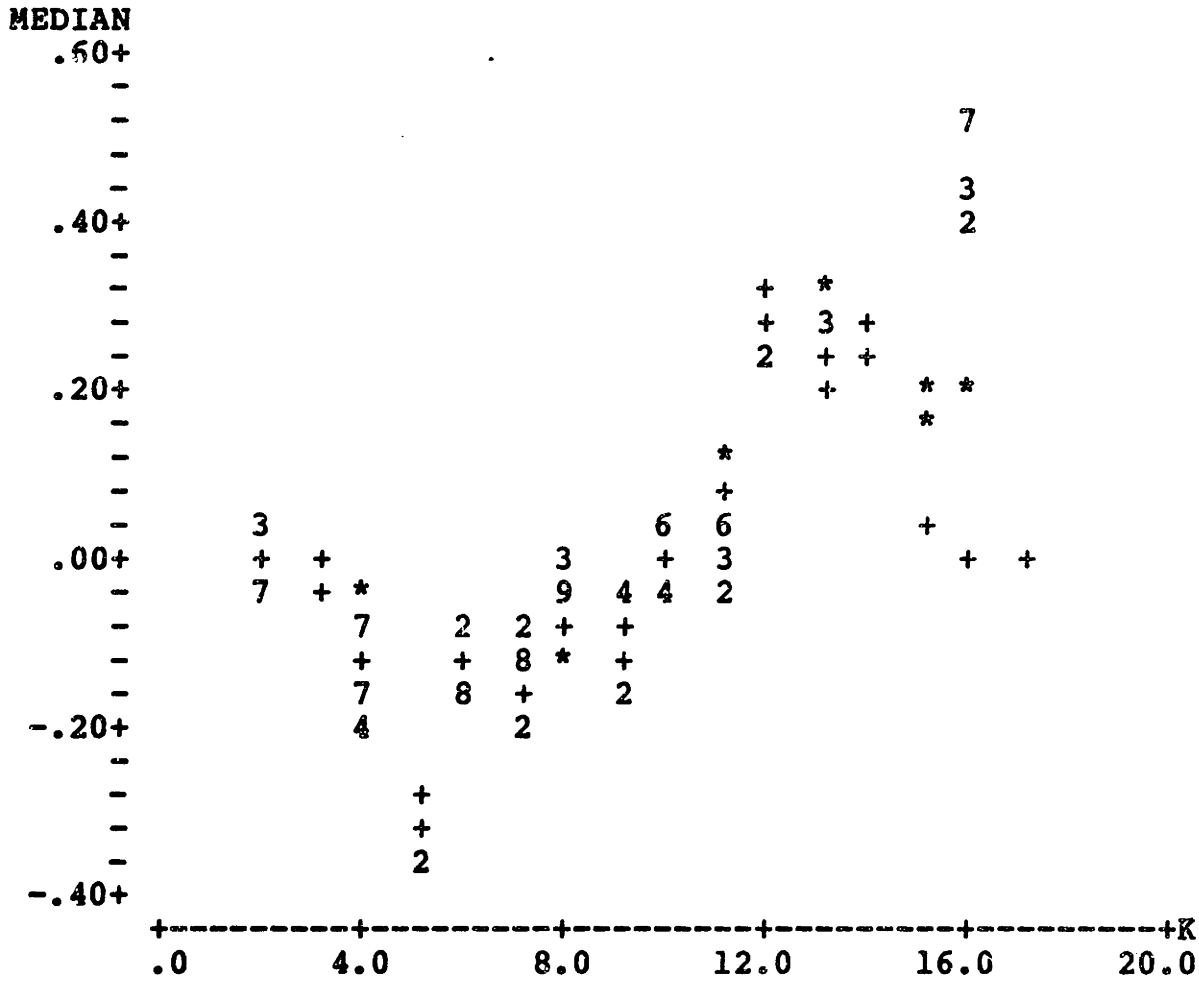
Figure 3.3.4.4: Deviation of Estimator from True Bk
Set A1, Complete Linkage, Sample 13



3.3.5. Data Set B2, Average Linkage

This last plot shows how the technique works on a set with less distinct clusters. The results are not significantly different from the other experiments.

Figure 3.3.5.1: Deviation of Estimator from True Bk
Set B2, Average Linkage, Sample 13



4. CONCLUSIONS AND SUMMARY

In this study we propose to use the bootstrap and the Bk statistic in combination to perform two tasks:

1) We can find the more stable and important clusters by looking at those clusterings which tend to produce peaks in the Bk plot. This information about the variability of the data set will be very useful when interpreting the clustering tree.

2) We can estimate the sample-to-true Bk by examining the bootstrap-to-sample Bk's. The best results in this study came from using the median of bootstrapped Bk's and either average or complete linkage.

We have documented the results of a simulation study here and these results indicate the usefulness of the proposed procedure. The model used in the experiments was appropriate for the common example described above. However, it is not known whether the variation in other situations is comparable. It could be much more, with worse results, or much less, with better results. The methods presented here show promise but, because of the limited scope of the study, have yet to be tested in real applications.

BIBLIOGRAPHY

Baker, F. B. (1974), "Stability of Two Hierarchical Grouping Techniques Case I: Sensitivity to Data Errors". JASA 69, pp. 440-465.

Baker, F. B. and Hubert, L. J. (1975), "Hierarchical Clustering and the Concept of Power". JASA 70, pp. 31-38.

Efron, B. (1979a), "Computers and the Theory of Statistics: Thinking the Unthinkable". SIAM Review 21, pp. 460-480.

Efron, B. (1979b), "Bootstrap Methods: Another Look at the Jackknife". Annals of Statistics 7, pp. 1-26.

Efron, B and G. Gong. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation". American Statistician 37 #1, pp. 36-48.

Fowlkes, E. B. and C. L. Mallows. (1983), "A Method for Comparing Two Heirarchical Clusterings" JASA (to appear in September).

Hartigan, J. A. (1969), "Using Subsample Values as Typical Values" JASA 64, pp. 1303-1317.

Hartigan, J. A. (1971), "Error Analysis by Replaced Samples", J. Royal Stat. Soc. Series B 33, pp. 98-110.

Hartigan, J. A. (1975), Clustering Algorithms. Wiley.

Hartigan, J. A. (1981), "Consistency of Single Linkage for High-Density Clusters", JASA 76, pp. 388-394.

Hubert, L. (1974), "Approximate Evaluation Techniques for the Single Link and Complete Link Hierarchical Clustering Procedures", JASA 69, pp. 698-704.

Johnson. S. C. (1967), "Hierarchical Clustering Schemes", Psychometrica , pp. 241-254.

Ling. R. F. (1973), "A Probability Theory of Cluster Analysis", JASA 68, pp. 159-164.

Rand, W. M. (1971), "Objective Criteria for Evaluation of Clustering Methods". JASA 66, pp 846-850.

Wong, M. A. (1982), "A Hybrid Clustering Algorithm for Identifying High-Density Clusters", JASA 77, pp 841-847.

Wong, M. A., and T. Lane (1983), "A Kth Nearest Neighbor Clustering Procedure", J. Royal Stat. Soc. Series B (to appear).