Metodološki zvezki, Vol. 2, No. 2, 2005, 173-193

# Hierarchical Clustering with Concave Data Sets

Matej Francetič, Mateja Nagode, and Bojan Nastav[1]

**Abstract**

Clustering methods are among the most widely used methods in multivariate analysis. Two main groups of clustering methods can be distinguished: hierarchical and non-hierarchical. Due to the nature of the problem examined, this paper focuses on hierarchical methods such as the nearest neighbour, the furthest neighbour, Ward's method, between-groups linkage, within-groups linkage, centroid and median clustering.

The goal is to assess the performance of different clustering methods when using concave sets of data, and also to figure out in which types of different data structures can these methods reveal and correctly assign group membership. The simulations were run in a two- and three-dimensional space. Using different standard deviations of points around the skeleton further modified each of the two original shapes. In this manner various shapes of sets with different inter-cluster distances were generated. Generating the data sets provides the essential knowledge of cluster membership for comparing the clustering methods' performances.

Conclusions are important and interesting since real life data seldom follow the simple convex-shaped structure, but need further work, such as the bootstrap application, the inclusion of the dendrogram-based analysis or other data structures. Therefore this paper can serve as a basis for further study of hierarchical clustering performance with concave sets.

## 1 Introduction

Clustering methods represent one of the most widely used multivariate techniques in practice. Essentially, there are two main groups of these methods: hierarchical and non-hierarchical. Statistics mainly uses the latter; this paper, on the other hand, primarily deals only with hierarchical clustering methods. The idea behind this decision is that the non-hierarchical methods, the most widely used being the k-means method, do not perform well with concave sets of data, since the centroids' usage results in wrong classifications. Although several hierarchical clustering methods exist this paper focuses on the methods implemented in the statistical software SPSS. These are: the nearest neighbour, the furthest neighbour, centroid

[1] University of Ljubljana, Slovenia

method, median clustering, linkage between groups and within linkage groups, and Ward's method. Due to high discrepancies in naming the methods, we will follow the SPSS wording.

The primary aim of this paper is to assess the performance of different clustering methods when using concave sets of data and also to figure out in which types of different data structures these methods can reveal and correctly assign group membership. Sets of points differing in shape (skeleton) and inter-point distance were used in the analysis: the simulations were run in a two and three-dimensional space with different standard deviations of points around the skeleton. Generating the sets of points gave an advantage, since the knowledge of cluster membership is essential in comparing the performances of clustering methods (perhaps better "in comparing"). In this manner various shapes of sets with different inter-cluster distances were generated. Certain limitations were imposed since the used parameters can lead to a vast number of generated sets. Applying different hierarchical clustering methods to these generated sets was the basis for assessing clustering accuracy and by this the performance of different hierarchical clustering methods for concave sets of data. We have not come across any studies dealing with hierarchical clustering on concave data; the latter are however, mainly dealt with other, "modern" methods, such as the fuzzy or wave clustering (see section five).

The paper consists of two parts. First we introduce the research methodology and briefly outline the methods used (section two). In the second part, a report of generating data sets is presented in section three and presentation of the results of successfulness of different clustering methods performed on the generated data in section four. Section five concludes the work and presents suggestions for possible further research.

# 2 Clustering

Clustering is believed to be one of the mental activities that human beings have been using for centuries. Classifying objects into classes has improved the control over objects classified and deepened the understanding of different classes. Gathering and accumulation of knowledge would be of no practical use without clustering (perhaps better "without clustering"). Besides the spread of knowledge base time has also brought an advance in clustering methods. Nowadays, despite the mathematical and statistical primacy over the methods, other fields, especially medicine and marketing, find clustering as a very useful tool.

The goal of clustering is to merge objects (units or variables) with regard to their characteristics and, by doing so, obtain internal homogeneity and external heterogeneity (isolation) of classes (clusters) produced. Characteristics, or better, the criteria according to which the objects are clustered, are usually, depending on the method used, based on the proximity matrix, which measures the objects'

distances or similarities. Our work was based on hierarchical clustering in SPSS (see description of methods below) and distance was the tool used for measuring the similarity among objects. To be more precise, we have used Euclidian distance, which can be presented by the following equation (or graphically presented in Figure 1):

$$D_{ij} = \left( \sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|^{2} \right)^{1/2} \tag{2.1}$$

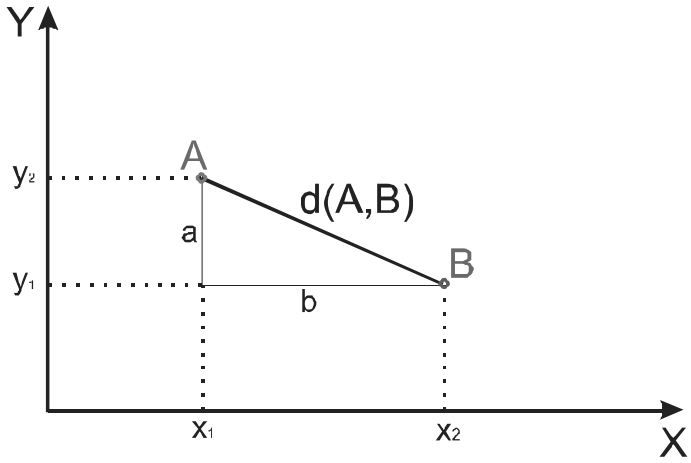



**Figure 1:** Euclidian distance for two points A and B.

Euclidian distance measures the actual (spatial) distance between two objects (Figure 1 gives a presentation of two-dimensional space, but can be applied to *p* variables using Equation (2.1)). Other distances could be used as well, such as Mahalonobis distance, which is a statistical distance (taking into account the standardization of units) between two points and includes covariances or correlations between variables (Sharma, 1996: 44). Applying this distance to our concave data sets would blur the picture: data would be "stretched" along some line in space and due to interlaced data its use would not lead to better results. The data used were generated (see chapter three) with no additional standardization and the Euclidian distance was applied. Choosing the tool to measure the distance is the first step. The second step is defined by the method and refers to the way this tool (distance) is applied among objects.

## 2.1 Hierarchical clustering

Hierarchical clustering is an iterative procedure for clustering objects. The starting point of hierarchical cluster analysis is a data matrix containing proximities between separate objects. At each hierarchical step, the two objects that are most similar, given certain criteria, depending on the method applied, are joined. A

joined pair is again called an object or a cluster. This means that at any hierarchical step (1) two single items may be clustered to form one new cluster, (2) a single item may be added to an existing cluster of items, or (3) two clusters may be combined into a single larger cluster. This process continues until all items are joined into only one cluster (Abswoude et al, 2004:337)

### 2.1.1 Nearest neighbour[2]

The nearest neighbour method measures distance between clusters as the distance between two points in the clusters nearest to each other. It tends to cause clusters to merge, even when they are naturally distinct, as long as proximity between their outliers is short (Wolfson et al, 2004: 610). The effect of the algorithm that it tends to merge clusters is sometimes undesirable because it prevents the detection of clusters that are not well separated. On the other hand, the criteria might be useful to detect outliers in the data set (Mucha and Sofyan, 2003). This method turns out to be unsuitable when the clusters are not clearly separated but it is very useful when detecting chaining structured data (chaining effect).

### 2.1.2 Furthest neighbour[3]

This method proceeds like the nearest neighbour method except that at the crucial step of revising the distance matrix, the maximum instead of the minimum distance is used to look for the new item (Mucha and Sofyan, 2003). That means that this method measures the distance between clusters through the distance between the two points in the clusters furthest from one another. Furthest neighbour results in separate clusters, even if the clusters fit together naturally, by maintaining clusters where outliers are far apart (Wolfson et al, 2004: 610). This method tends to produce very tight clusters of similar cases.

### 2.1.3 Centroid method

The centroid is defined as the centre of a cloud of points (Joining Clusters: Clustering Algorithms). Centroid linkage techniques attempt to determine the ‘centre’ of the cluster. One issue is that the centre will move as clusters are merged. As a result, the distance between merged clusters may actually decrease between steps, making the analysis of results problematic. This is not the issue with single and complete linkage methods (Wolfson et al, 2004: 610). A problem with the centroid method is that some switching and reversal may take place, for

[2] Also called Single Linkage Method or Minimum Distance Method.
[3] Also called Complete Linkage or Maximum Distance Method.

example as the agglomeration proceeds some cases may need to be switched from their original clusters (Joining Clusters: Clustering Algorithms).

### 2.1.4 Median method

This method is similar to the previous one. If the sizes of two groups are very different, then the centroid of the new group will be very close to that of the larger group and may remain within that group. This is the disadvantage of the centroid method. For that reason, Gover (1967) suggests an alternative strategy, called the median method, because this method could be made suitable for both similarity and distance measures (Mucha and Sofyan, 2003). This method takes into consideration the size of a cluster, rather than a simple mean (Schnittker, 2000: 3).

### 2.1.5 Linkage between groups[4]

The distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. This method is also very efficient when the objects form naturally distinct 'clumps', however, it performs equally well with elongated, 'chain' type clusters (Cluster Analysis).

### 2.1.6 Linkage within groups[5]

This method is identical to the previous one, except that in the computations the size of the respective clusters (i.e. the number of objects contained in them) is used as a weight. Thus, this method should be used when the cluster sizes are suspected to be greatly uneven (Cluster Analysis).

### 2.1.7 Ward's method

The main difference between this method and the linkage methods is in the unification procedure. This method does not join groups with the smallest distance, but it rather joins groups that do not increase a given measure of heterogeneity by too much. The aim of Ward's method is to unify the groups such that variation inside these groups does not increase too drastically. This results in clusters that are as homogenous as possible (Mucha and Sofyan, 2003). Ward's method is based on the sum-of-squares approach and tends to create clusters of similar size. The only method to rely on analysis of variance, its underlying basis

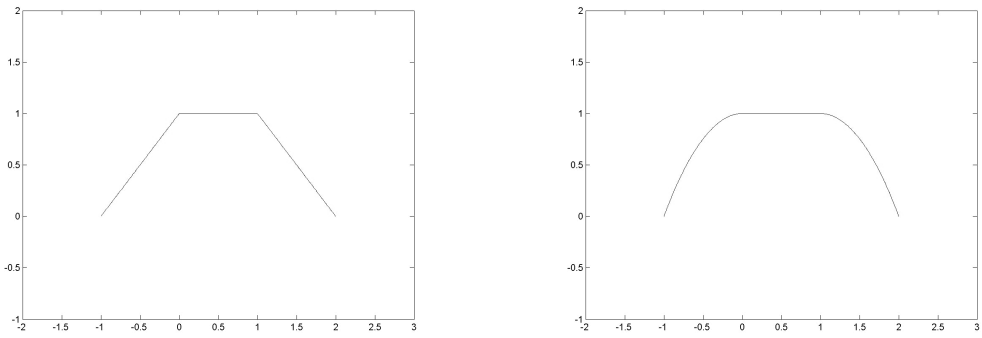
[4] Unweighted Pair-Groups Method Average (UPGMA).
[5] Weighted Pair-Groups Method Average (WPGMA).

is closer to regression analysis than the other methods. It tends to produce clearly defined clusters (Wolfson et al, 2004: 610).

# 3 Data generation

The analysis is based on concave sets of points. For the purposes of this paper the data are generated in two and three-dimensional space, however, it is easy to extend this process to a more dimensional space. This process consists of three steps.

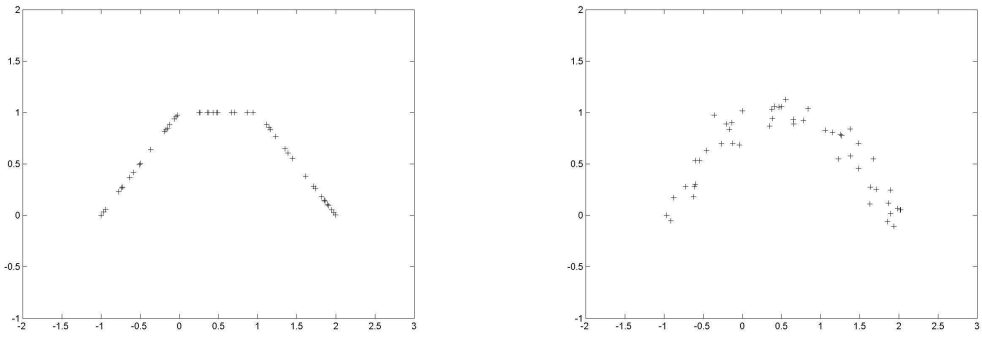
The first step is the construction of the skeleton. The skeleton is an arbitrary curve in a more dimensional space. The curve is represented as the finite set of ordered points. The points that lie on the curve between these selected points can be approximated with linear or cubic interpolation.



***(a)*** ***(b)***

**Figure 2:** Linear (a) and cubic (b) interpolation.

The use of linear interpolation in this paper is due to simplification of the calculations that are needed for further analysis. A better approximation of the target curve can also be achieved with a larger set of ordered points.



***(a)*** ***(b)***

**Figure 3:** Chosen points before (a) and after (b) shifting.

The second step is choosing the sample points. In order to do this we have normalized the length of the curve so that an arbitrary point on the curve is given by S(t), where 0<t<1. Sample points are then chosen as $T_i'=S(t_i)$, where $\{t_i; i=1,\ldots,k\}$ are independent uniformly distributed random numbers on interval [0,1].

The third step is moving (shifting) the chosen points by error vectors. We have decided that error vectors will have multivariate normal distribution with expectation 0 and covariance matrix $\sigma^2 I$. After generating independent identically distributed random vectors $E_i$ (i=1,…,k; one for each point) the final points are obtained as $T_i=T_i'+E_i$.

Before applying hierarchical clustering methods to two sets of point generated in the way described above, some definitions of certain parameters have to be introduced. The most basic definition is the definition of distance between two skeletons. This distance is called minimal distance (mr):

$$mr(S_1, S_2) = \min_{s,t} d(S_1(s) - S_2(t)) \tag{3.1}$$

where $S_1(s)$ and $S_2(t)$ stand for parameterization of two chosen skeletons, $d(S_1(s)-S_2(t))$ is Euclidian distance between two points.

The variation of points around the skeleton is another parameter that has to be clearly defined. It would be appropriate to define variation of points as the expected distance of these points from the skeleton. Taking into account the fact that the skeleton is section-linear (like a broken line), deriving the expected distance will be limited to the straight line. Let us imagine the following situation: the point on the straight line (with directed vector $p$) is shifted by an arbitrary n-dimensional normal vector $X$ with expectation 0 and covariance matrix $\sigma^2 I$. We would like to obtain the expected distance of this point from the straight line. For simplicity, assume that the straight line is vertical and that the shift $x_1$ (first component of a random vector) is collinear with the directed vector of the straight line. To be sure: there is an orthogonal matrix $A$ ($A^TA=I$), which is used to multiply $X$ to get $X'=AX$, such that $x_1'=Ax_1$ is collinear with directed vector of the straight line, $p$; $X'$ is distributed normally with expectation 0 and covariance matrix $\sigma^2 I$. In this case, Euclidian distance between the shifted point and the straight line equals $\sqrt{x_2^2 + x_3^2 + \ldots + x_n^2}$. Let $h = \left(x_2^2 + x_3^2 + \ldots + x_n^2\right)/\sigma^2$. This means that $h$ is distributed in $\chi^2$ with $n-1$ degrees of freedom, thus $E(h)=n-1$ and $V(h)=2(n-1)$ ($x_2, x_3,\ldots,x_n$ are independent and identically distributed ~ $N(0, \sigma^2)$). Let us approximate $\sqrt{h}$ using the development of Taylor series around $E(h)$:

$$\sqrt{h} \cong \sqrt{E(h)} - \frac{1}{2\sqrt{E(h)}}(h - E(h)) + \frac{1}{4\sqrt{E(h)^3}}(h - E(h))^2 \tag{3.2}$$

Using this $E\left(\sqrt{x_2^2 + x_3^2 + ... + x_n^2}\right) = \sigma$. $E\left(\sqrt{h}\right)$ is approximated with:

$$\sigma E\left(\sqrt{h}\right) \cong \sigma\left(\sqrt{n-1} + \frac{1}{2\sqrt{n-1}}\right) \tag{3.3}$$

The dispersion of points around the skeleton, *r*, is defined

$$r_S = \sigma_S\left(\sqrt{n-1} + \frac{1}{2\sqrt{n-1}}\right) = \sigma_S \frac{2n-1}{2\sqrt{n-1}} \tag{3.4}$$

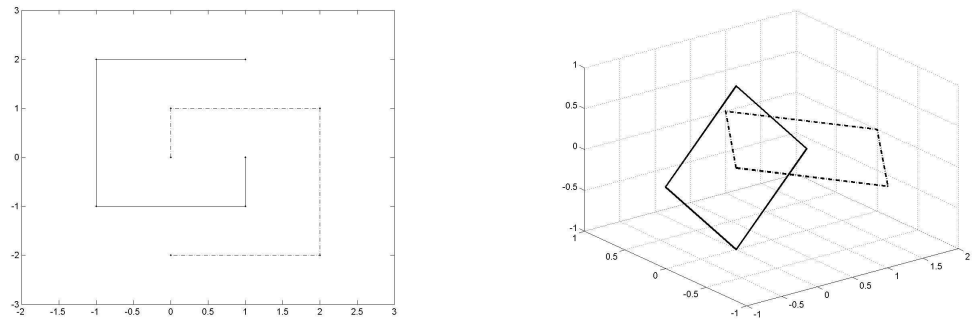
where $\sigma_S$ is standard deviation used in generating the points and *n* is the dimension of space, where skeleton *S* is situated.

Lastly, some number is needed to represent the degree of separation of data separation. Regarding this number, each case could be assigned the troublesome (or not) classification of objects into clusters. Furthermore, this number will serve as a degree of admissibility with each of the clustering methods used, when, if at all, some of the methods will have higher tolerance or will be able to separate data and correctly assign objects to classes. This number will be called the degree of separation of data, marked by *SLP*.

$$slp(S_1, S_2) = \frac{mr(S_1, S_2)}{r_{S_1} + r_{S_2}} \tag{3.5}$$

where $S_1$ and $S_2$ are skeletons of groups of points, *mr* is their minimal distance, $r_{S1}$ and $r_{S2}$ are respective dispersion of points.

The paper tests hierarchical clustering methods on ten examples of concave data sets. Each of the examples has two groups of points and examples are different with regard to skeleton and used standard deviations of points around the skeleton. We have limited our work to two different pairs of skeletons. One pair is defined in two, the other in three-dimensional space (see Figure 4). Furthermore, to briefly test the stability of the methods applied, some of the examples have been re-run using subsampling (of 50%).



***(a)*** ***(b)***

**Figure 4:** Pairs of two (a) and three-dimensional (b) skeletons.

For further analysis, the minimal distance between the skeletons in the above examples has to be known. In case a), this is equal to 1. It is interesting that in this case the minimal distance is achieved in almost all of the points on the skeleton. In case b) the minimal distance equals $1/\sqrt{3}$ (approximately 0.58) and is achieved only four times (from 0.58 to 1). From this structure of the skeletons one might expect that clustering would, in the first case, be harder along the whole skeleton, whereas in the second case this harder clustering would occur only at the place where rings reach their minimal distance between the skeletons.

In generating the data in each case the same standard deviation around the skeletons has been used. This was defined in a way to suit previously determined degree of separation of the data (SLP), which were chosen from the 1 – 2.5 interval. The standard deviations used with each pair of the skeletons are given in the following table.

With degree of separation of the data 2.5, the structures are well separated (there is practically no probability, that any of the points from one group would be closer to other skeleton); with SLP 1, the interlacing of the data is extreme. The following figure (Figure 5) shows the data with SLP 2 and 1.2.

# 4 Results

The estimation and successfulness of the methods included is possible on the basis of the percentage of correctly assigned group membership that is known in advance. This, however, is almost never the case in real life. Therefore, in real data the performance of the method is seldom accurately assessed. This paper deals with generated data, which means that the real situation is known and measuring performances of different methods is therefore a rather easy task. Using the percentage of correctly assigned units (to the correct cluster) is the key indicator used in the analysis. Further on, when performing the methods with statistical package SPSS classification into two groups was chosen. This produced the results with percentages from 50% (all cases were correctly classified just in

one group and the other group remained 'empty') to 100% (all cases were correctly classified in both groups) of correctly assigned group membership. Tables separated for two- and three-dimensional space present the results. The marginal examples (SLP=2.5 and SLP=1) together with the central (SLP=1.5) have been additionally subsampled and used on the methods more times (50) in order to obtain the notion of the methods' stability. Thus, the method in this context appears to be stable if the average percentage of correctly assigned group membership of all subsamples does not vary much among different degrees of separation of the data.

**Table 1:** Used standard deviations.

| SLP value | 2.5 | 2 | 1.5 | 1.2 | 1 |
|---|---|---|---|---|---|
| a) | 0.11 | 0.17 | 0.22 | 0.28 | 0.33 |
| b) | 0.07 | 0.08 | 0.11 | 0.13 | 0.16 |

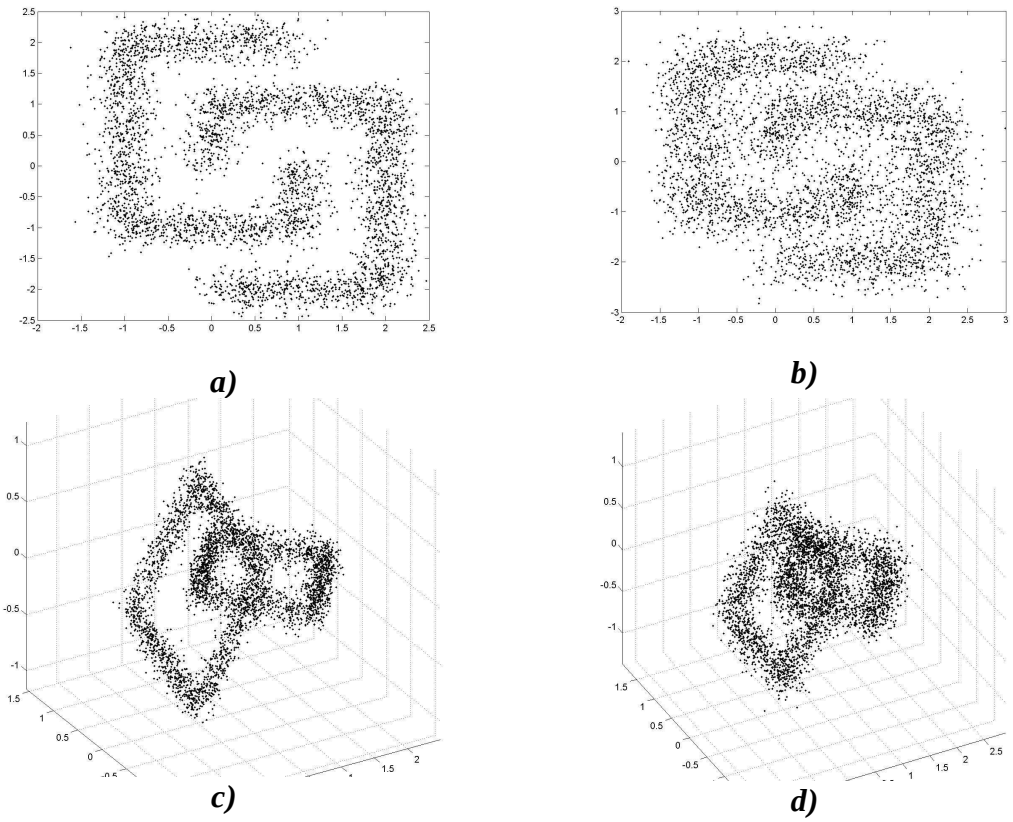



**Figure 5:** Pictures of the data: a) 2D, SLP=2; b) 2D, SLP=1.2; c) 3D, SLP=2; d) 3D, SLP=1.2.

The results of two-dimensional data gave the following conclusions. When the points are clearly separated, the only method that classifies all the cases correctly is the nearest neighbour. This method is the only method that is suitable for these

types of data. Even when the groups are not (internally) cohesive but are isolated from each other the nearest neighbour method can produce correct results. Other methods do not fulfil the criterion of correctly assigned group membership with such a high percentage. Nevertheless, the percentages are not so low and this is the consequence of great degree of separation of data.

**Table 2:** Results of clustering with two-dimensional data.

| **Correctly classified units (%)** | SLP=2,5 | | | SLP=2 | | | SLP=1,5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **Total** | **1** | **2** | **Total** | **1** | **2** | **Total** |
| **METHOD** | | | | | | | | | |
| Between | **76,4** | **69,4** | 72,9 | **81,1** | **76,8** | 78,95 | **71,8** | **72,1** | 71,95 |
| Within | **69,9** | **80,5** | 75,2 | **65,2** | **56,6** | 60,9 | **61,9** | **50,5** | 56,2 |
| Nearest | **100** | **100** | 100 | **100** | **0,1** | 50,05 | **100** | **0,1** | 50,05 |
| Furthest | **73,3** | **76,9** | 75,1 | **73,5** | **44,9** | 59,2 | **62,9** | **70,1** | 66,5 |
| Centroid | **76,3** | **66,5** | 71,4 | **54,2** | **78,1** | 66,15 | **54,9** | **56,9** | 55,9 |
| Median | **31,6** | **100** | 65,8 | **51,6** | **52,2** | 51,9 | **91,0** | **66,7** | 78,85 |
| Ward's | **72,5** | **48** | 60,25 | **77,6** | **68,2** | 72,9 | **73,2** | **78,6** | 75,9 |
| K-means | **61,5** | **57,8** | 59,65 | **59** | **57** | 58 | **62,3** | **62,6** | 62,45 |

| **Correctly classified units (%)** | SLP=1,2 | | | SLP=1 | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **Total** | **1** | **2** | **Total** |
| **METHOD** | | | | | | |
| Between | **74,2** | **73,6** | 73,9 | **50,7** | **50,6** | 50,65 |
| Within | **49,7** | **61,9** | 55,8 | **77,9** | **69,6** | 73,75 |
| Nearest | **100** | **0,1** | 50,05 | **100** | **0,1** | 50,05 |
| Furthest | **41,2** | **92,7** | 66,95 | **66,5** | **86,4** | 76,45 |
| Centroid | **100** | **0,1** | 50,05 | **100** | **0,1** | 50,05 |
| Median | **90,7** | **19,9** | 55,3 | **100** | **0,1** | 50,05 |
| Ward's | **67,9** | **74,1** | 71 | **68,4** | **77,6** | 73 |
| K-means | **59,9** | **59,9** | 59,9 | **66,5** | **63,6** | 65,05 |

**Table 3:** Results of clustering with three-dimensional data.

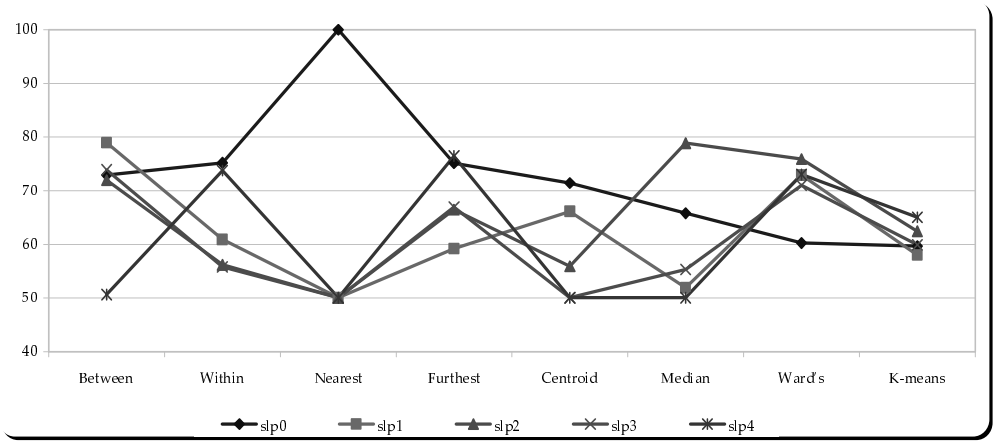
| Correctly classified units (%) | SLP=2,5 | | | SLP=2 | | | SLP=1,5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **Total** | **1** | **2** | **Total** | **1** | **2** | **Total** |
| **METHOD** | | | | | | | | | |
| Between | **100** | **59,3** | 79,65 | **100** | **74,2** | 87,1 | **100** | **66,9** | 83,45 |
| Within | **100** | **53,3** | 76,65 | **55,8** | **100** | 77,9 | **99,9** | **60,5** | 80,2 |
| Nearest | **100** | **100** | 100 | **100** | **0,1** | 50,05 | **100** | **0,1** | 50,05 |
| Furthest | **100** | **75** | 87,5 | **100** | **63,6** | 81,8 | **55,2** | **94,1** | 74,65 |
| Centroid | **100** | **52,1** | 76,05 | **100** | **51,6** | 75,8 | **100** | **0,1** | 50,05 |
| Median | **100** | **31,3** | 65,65 | **100** | **44,5** | 72,25 | **100** | **2,4** | 51,2 |
| Ward's | **100** | **51,9** | 75,95 | **100** | **53,1** | 76,55 | **100** | **42,8** | 71,4 |
| K-means | **75** | **74,1** | 74,55 | **71** | **77,7** | 74,35 | **74,7** | **75,5** | 75,1 |

| Correctly classified units (%) | SLP=1,2 | | | SLP=1 | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **Total** | **1** | **2** | **Total** |
| **METHOD** | | | | | | |
| Between | **100** | **49,3** | 74,65 | **42,5** | **99,4** | 70,95 |
| Within | **100** | **50,6** | 75,3 | **56,5** | **89,2** | 72,85 |
| Nearest | **100** | **0,1** | 50,05 | **100** | **0,1** | 50,05 |
| Furthest | **45,7** | **83,7** | 64,7 | **77,4** | **98,2** | 87,8 |
| Centroid | **100** | **0,1** | 50,05 | **100** | **23,2** | 61,6 |
| Median | **100** | **65,4** | 82,7 | **100** | **0,1** | 50,05 |
| Ward's | **100** | **49,7** | 74,85 | **100** | **42,2** | 71,1 |
| K-means | **72,9** | **75,6** | 74,25 | **73,4** | **75,5** | 74,45 |

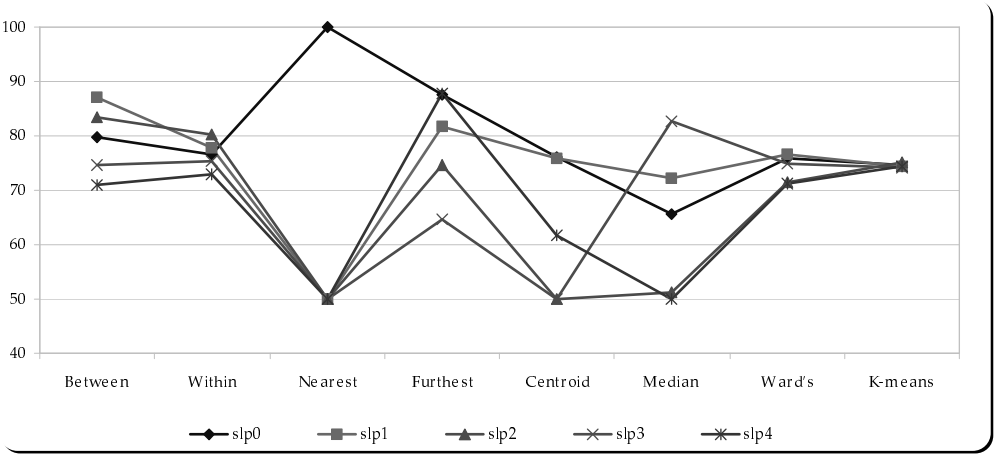
**Note: With SLP3 and SLP4, 75% and 60% samples, respectively, were used, due to technical problems with the software.**

Results obtained by using methods with next degree of separation are different from results obtained with the highest degree of separation. In this step the nearest neighbour method is completely unsuccessful. The average method performs with a quite high percentage of correctly assigned units. For all further degrees of separation of the data the results are quite similar, except when observing the data with SLP=1. In that case there is no internal cohesion and no isolation between groups. With the smallest degree of separation of the data in a two-dimensional space the best performance is assigned to the furthest neighbour method, average

method (between groups) and Ward's method. Ward's method seems to be the most stable method among the observed hierarchical methods. In almost all cases (except when the groups are clearly separated, but not cohesive) the successfulness of correctly assigned group membership is around 70%. This is not surprising since it is known that this method performs well and presents a compromise between chaining and compact data. However, the clusters formed are not “in line” with the skeletons generated, but rather compact clusters.




*(a) 2-dimensional space*




*(b) 3-dimensional space*

**Figure 6:** Results of all the methods used.

Similar conclusions can be drawn when observing data and methods in a three-dimensional space. Again, certainly the best method with the highest degree of data separation is nearest neighbour and we can conclude that it is the only method suitable for that kind of data. However, this is only the case when the data are clearly separated. When the degree of separation decreases, the method is no longer successful. The opposite holds for the lowest degree of data separation, where its performance turns out to be completely unsuccessful. In this case the average method (between and within groups) performs better. Ward's method gives similar results and consequently similar conclusions as in a two-dimensional

space. In a three-dimensional space the furthest neighbour method gives the highest percentage (87.8%) of correctly assigned group membership with the lowest degree of data separation. With very similar percentage it performs also in the case of the highest degree of data separation. In both extreme cases the furthest neighbour performs with similar successfulness. This is to some extent surprising, since this method is believed to be the most successful with compact clusters and not with skeletons like this. At the same time, the results reveal that the clusters this method (and others) tend to produce are compact and do not follow the skeletons generated.

# 5 Conclusions

This paper is based on generated data and software (SPSS) usage. The latter was used to run and test all of the seven hierarchical clustering methods implemented in this software on previously generated data. These were obtained by dispersing points around two basic shapes of skeleton, both in two and three-dimensional space. In the paper, only one data set was used and in some examples further subsampling of these data was constructed in order to obtain some idea about the methods' stability. The skeletons used represent rather hard-to-separate shapes and we have intentionally used such skeletons – if the method can work and be successful with such structures, it can be successful elsewhere as well. Unfortunately, very seldom can we come across less "tricky" data in real life problems. Despite this, using the generated data has given us the power to see which methods perform well and which do not and, relying on this, some conclusions can be drawn.

Results in the form of tables (previous section) and figures (appendix) speak for themselves. Nevertheless, Figure 6 sums up the results of clustering.

When implementing some of the variability obtained from the subsampling and rerunning the methods several (50) times, the methods appear to be more stable within the given degrees of separation of the data. Based on these results, it is hard to recommend instructions on how to deal with such data, since no method used performs particularly well with the generated data. Some, i.e. the centroid, and the median methods appear to be more variable within the given SLP, whereas the nearest neighbour method completely fails when the data are more interlaced (but it is stable given the other levels of SLP). Other methods are thus more appropriate regarding stability. Bearing in mind the fact that real life does not follow simple, homogeneous and isolated groups, only brief outlines can be put forward at this point.

1. *Check the data skeleton*. This task seems rather simple for two or three-dimensional data, but is otherwise virtually impossible. With several variables one can first use a data-reduction technique, such as the principal

component analysis, and then, using fewer variables, the task of determining the proper skeleton to the data turns out to be much easier. An example of too-many-variables problems can be found in Gordon (1999: 24, 25). If the skeleton(s) can be determined, take the following step.

2. *Compare these skeletons* with the skeletons used in this assignment and determine the most similar skeletons, and choose the most appropriate method, given the skeleton and dispersion of points.

Our findings can be further backed up by the following:

3. Performance of clustering methods decreases with increased dispersion of the data, which is expected. In case when the criteria, such as the degree of separation of the data (SLP) used in this paper, are high enough (dispersion is low), the method to be used is the nearest neighbour. However, when the data are not so well isolated additional attention needs to be given to choosing the right method.
4. Using three variables (a three-dimensional space), as opposed to a two-dimensional space gives much better results (by several percentage points), meaning that three variables can better determine the proper data structure. However, one should keep in mind the previously mentioned too-many-variables problems (Point 1), which we can come across in reality. Using (at least) three variables is thus advised.

Separate from this paper's main focus, i.e. the hierarchical clustering methods, new methods are being developed and used. They apply to different fields and are usually custom made for each type of analysis. Intuitively speaking, they would follow point one of the above mentioned points and (algebraically) determine the skeleton; furthermore, by applying the Euclidian distance, objects would be classified. Phrases such as "fuzzy clustering" are used, describing the situation where some objects are with certainty classified to one group, while the others could be in one or the other group. Similarly, overlapping groups where objects are in one and the other groups need special attention. Wave clustering, introduced by Sheikholeslami, Chatterjec and Zhang (1998) is a method, which presumably works well with concave data sets. Our task was not to focus on such methods, but we cannot avoid mentioning them at this point. Additional reading can be found in Gordon (1999: 111-130) and a series of scientific articles using clustering methods in practise, mainly medical studies.

Further improvements to this paper could be made by generating several random data sets with given parameters (skeleton, degree of separation of data) and running the methods on these data. The described process would allow to test the methods more broadly for their variability among different degrees of separation of data. We believe that observing the obtained dendrograms is a good tool to be used with clustering since to a certain degree, they can reveal the real

structure of the data. However, their usage is again limited – figures in the appendix show that only with the highest degrees of separation of the data, SLP=1 in Figure 7, they can be used to determine the proper clustering. However, dendrograms were not under close inspection in this analysis and therefore, along with other data structure, this represents a point to be studied further in details.

# References

[1] van Abswoude, A.A.H, Vermunt, J.K., Hemker, B.T., and van der Ark, L.A. (2004): Mokken Scale Analysis Using Hierarchical Clustering Procedures: *Applied Psychological Measurement*, **5**, 332-354.

[2] Aldenderfer, S.M. and Blashfield, K.R. (1984): Cluster Analysis. Series: *Quantitative Applications in the Social Sciences*, **44**. Sage Publications.

[3] CL Algorithm Details. http://ei.cs.vt.edu/~cs5604/f95/cs5604cnCL/CL-alg-details.html, 19.08.2004.

[4] Cluster Analysis. http://www.statsoft.com/textbook/stcluan.html, 19.08.2004.

[5] Ferligoj, A. (1989): Razvrščanje v skupine. Teorija in uporaba v družboslovju. *Metodološki zvezki*, **4**, Ljubljana.

[6] Field, A. (2000): Cluster Analysis.

[7] http://www.sussex.ac.uk/Users/andyf/teaching/pg/cluster.pdf, 19.08.2004.

[8] Gordon, A.D. (1999): *Classification*. New York: Chapman&Hall/CRC,.

[9] Joining Clusters: Clustering Algorithms. http://149.170.199.144/multivar/ca_alg.htm, 19.08.2004.

[10] Mucha, H-J. and Sofyan H. (2003): Cluster Analysis. http://www.xplore-stat.de/tutorials/clustnode3.html, 15.08.2004.

[11] Schnittker, J. (2000): Cluster Analysis Presentation. http://www.indiana.edu/~socsrp/cluster_analysis.pdf, 15.8.2004.

[12] Sharma, S. (1996): *Applied Multivariate Techniques*. John Wiley&Sons, Inc., New York.

[13] Sheikholeslami G., Chatterjec S., and Zhang A. (1998): WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. http://www.cs.sfu.ca/CourseCentral/459/han/papers/sheikholeslami98.pdf, 05.10.2004.

[14] SPSS Statistical Algorithms. SPSS Inc., 1985.

[15] Wolfson, M., Madjd-Sadjadi, Z., and James, P. (2004): Identifying national types: A cluster analysis of politics, economics and conflict. *Journal of Peace Research*, **5**, 607-623.

# Appendix

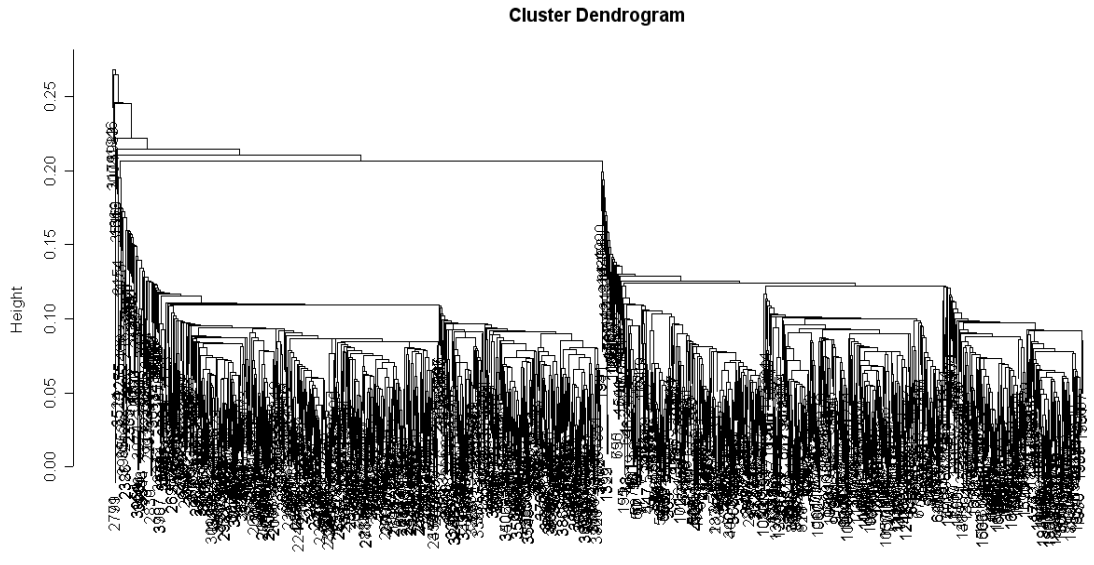



**Figure 7:** Dendrogram for SLP=2. Two groups can be clearly distinguished.

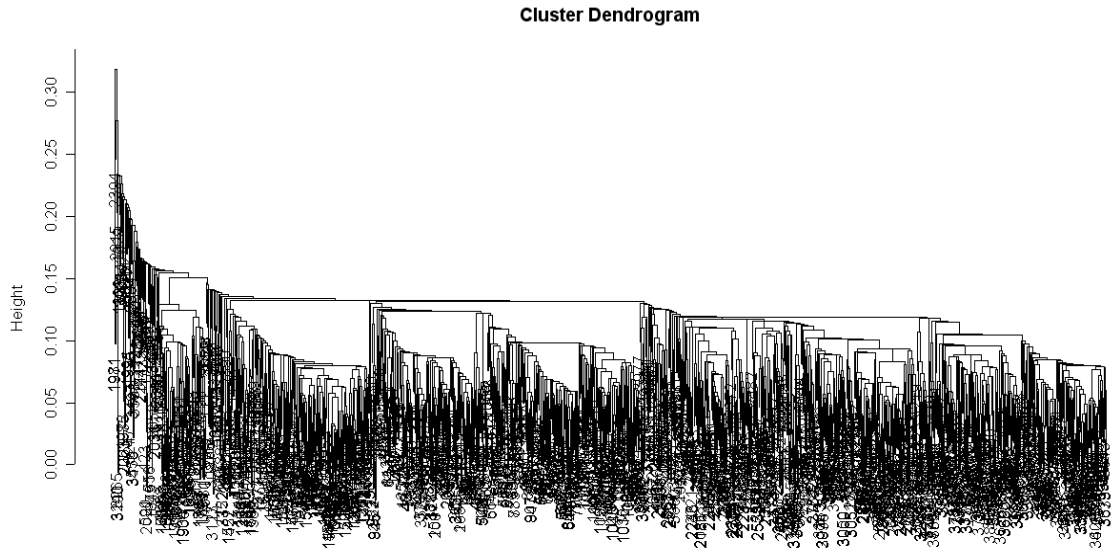



**Figure 8:** Dendrogram for SLP=1.5. Number of groups is not clearly seen.

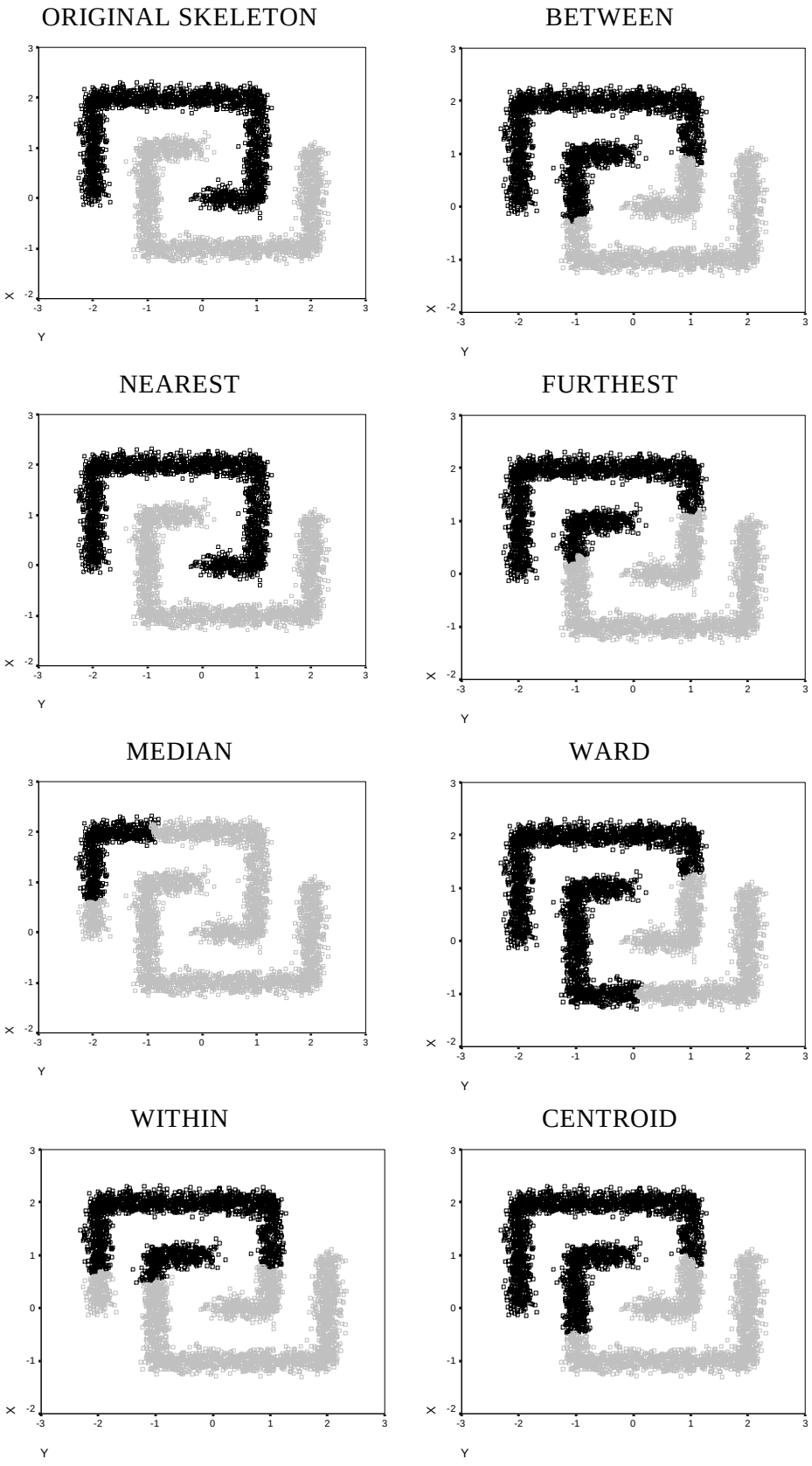



**Figure 9:** Graphical representation of results for 2D with SLP=2.5.

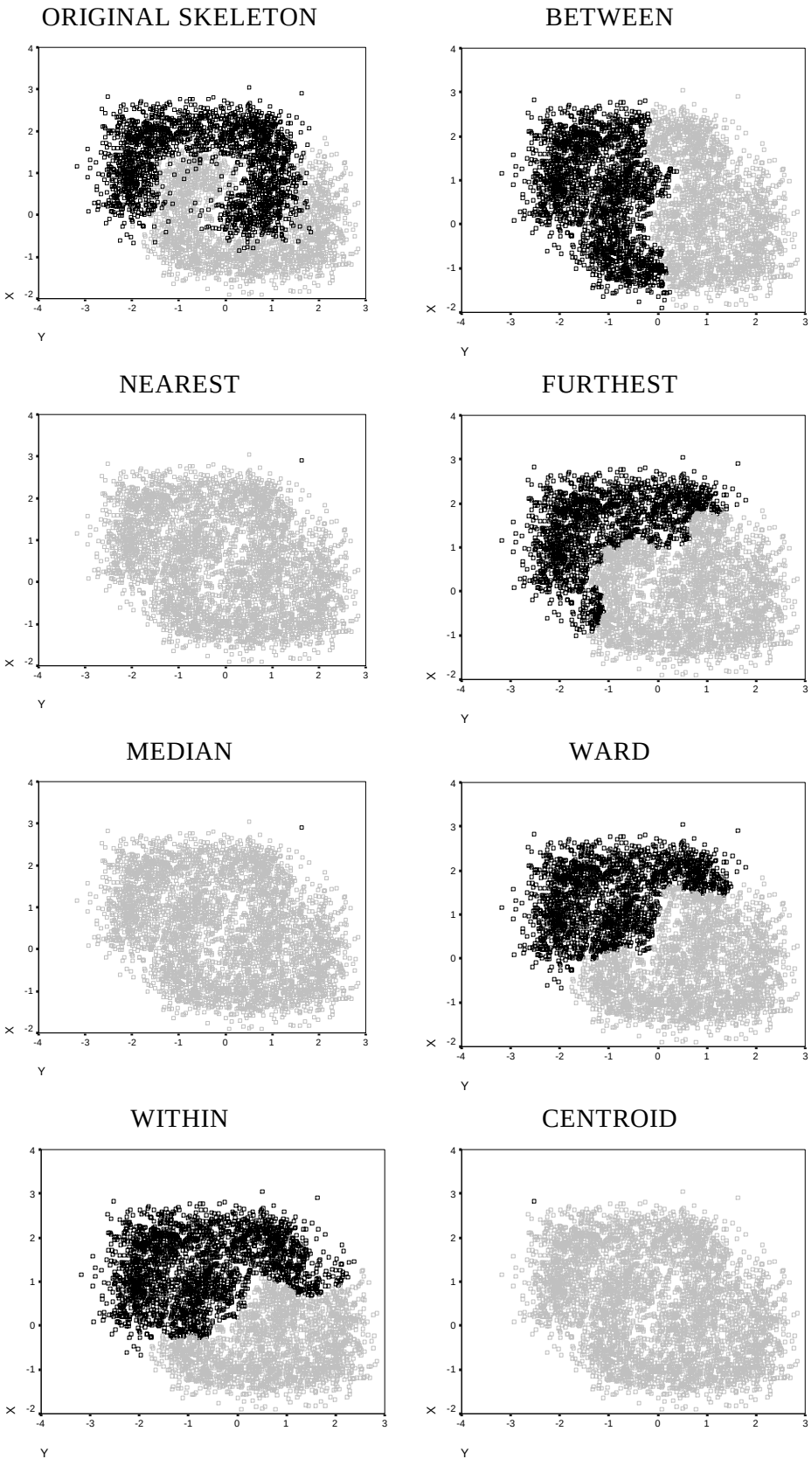



**Figure 10:** Graphical representation of results for 2D with SLP=1.

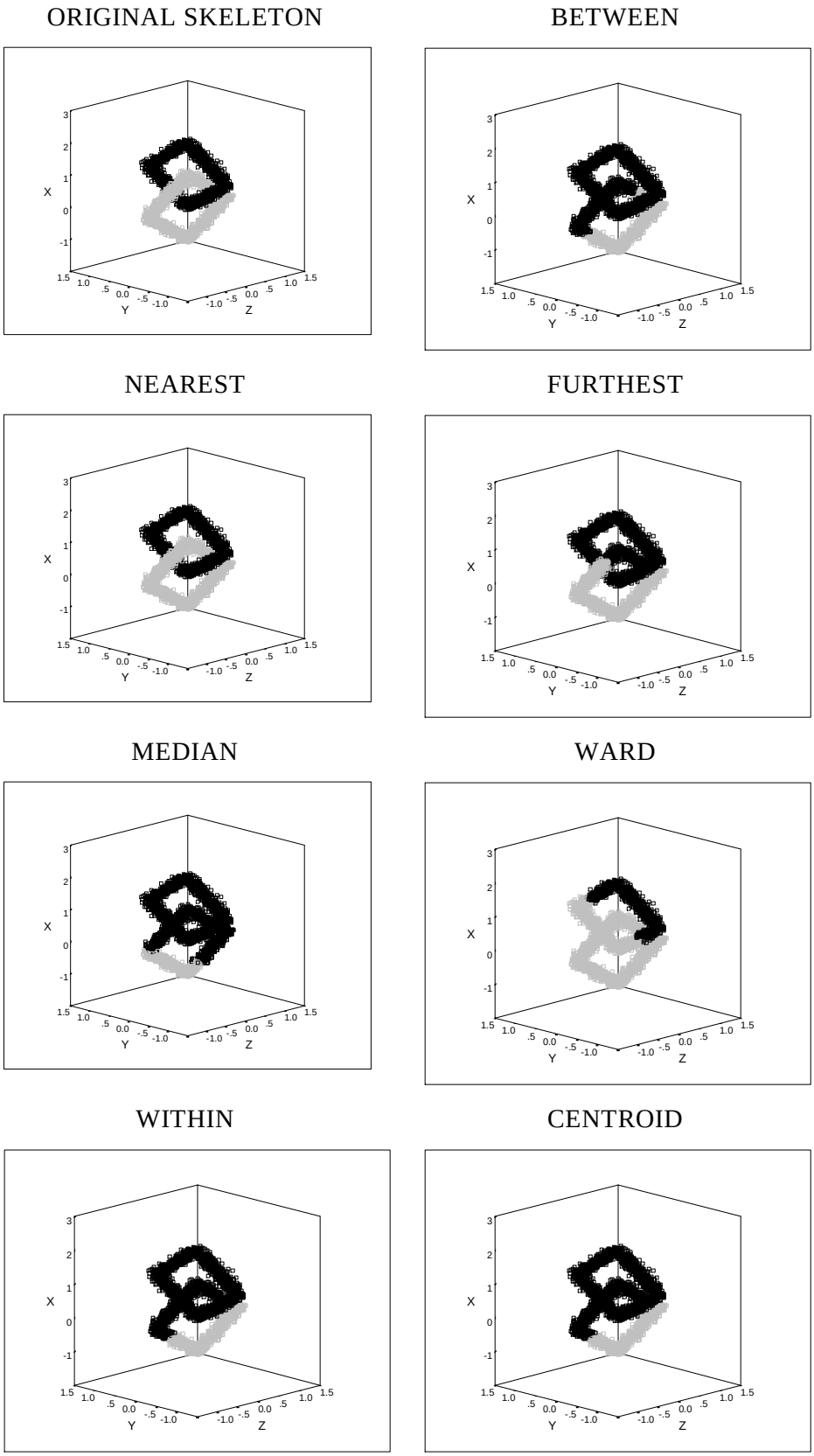



**Figure 11:** Graphical representation of results for 3D with SLP=2.5.

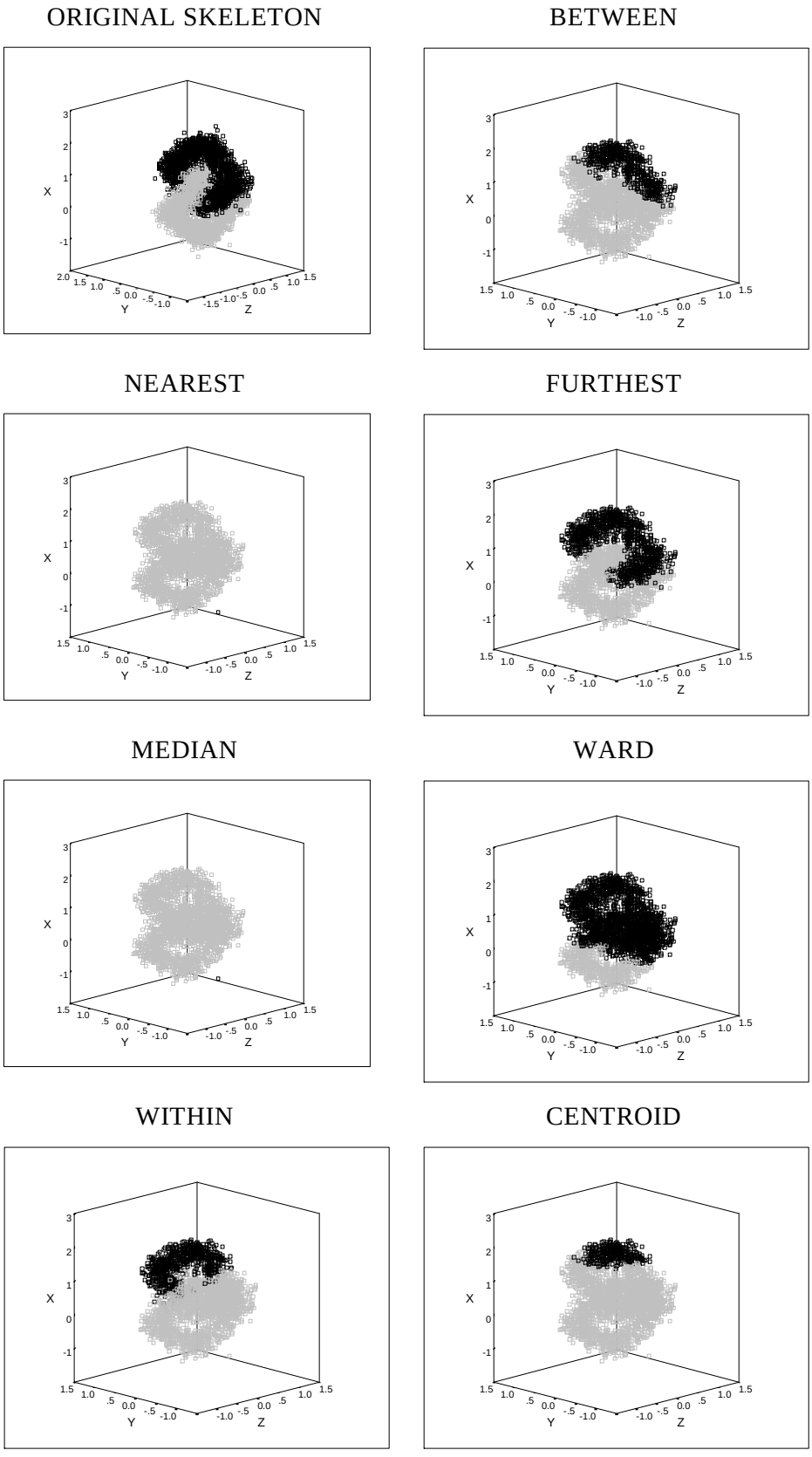



**Figure 12:** Graphical representation of results for 3D with SLP=1.