Metodološki zvezki, Vol. 1, No. 1, 2004, 75-97

A Study of *k*-means Method where Starting Conditions are Changed: A Simulation Study

Rok Podgornik[1], Marijan Zafred[2], and Anja Pajtler[3]

Abstract

The goal of our work is to evaluate the performance of *k*-means clustering method. We created three (or two) groups of units, which compose one of our three different structures. First we created groups, which were clearly separated in three-dimensional space, and then we started to decrease the differences among them with increasing the data dispersion at keeping the centroids constant. We wanted to find out how long we can continue with this process until we still get satisfactory results at clustering of units into groups. Six different methods of clustering were used (with different initial cluster centers) and we evaluated every one of them. In comparison with initial clustering we are observing three criteria: percentage of correctly clustered units, centroids dislocation and value of criteria function. The results of clustering are presented graphically with added values of three chosen criteria for better possibility of evaluation. For the simulation of the data and graphical presentation we used program package R.

1 Introduction

Ordering or classifying of similar things (objects) into groups is a very old and intuitionally simple problem that represents a basic task of every research. The fundamental problem is to classify objects into groups by chosen criterion in a way that the objects inside groups are as internally cohesive (homogeneous) as possible and the objects in different groups are as externally isolated (separated) as possible. The groups that we get are then called classification or clustering. Although the problem is set very simple at the first sight, the determination of the right clustering is not an easy task. For solving the problem of clustering there is a

[1] Postgraduate study of Statistics, University of Ljubljana; employed in Henkel Slovenija, Ljubljana.

[2] Postgraduate study of Statistics, University of Ljubljana.

[3] Postgraduate study of Statistics, University of Ljubljana.

large number of methods, but in this article we focus only on *k*-means clustering method.

Our primary goal is to evaluate the performance of *k*-means clustering method. First of all we created three groups of units, which compose our structure, where we exactly know, in which group every single unit belongs. By doing that we can always check the correctness of the clustering. First we created groups, which were clearly separated in three-dimensional space, and then we started to decrease the differences among them. We wanted to find out how long we can continue with this process until we still get satisfactory results at clustering of units into groups. In comparison with initial clustering we are observing three criteria:

1. percentage of correctly clustered units,
2. centroids dislocation and
3. value of criteria function.

By creating groups, which are less and less isolated, we can choose between two methods: we can either bring centroids of the groups closer together at constant dispersion of the data or gradually enlarge the dispersion of the data at constant centroids. In this article we are focusing on only one of these methods – the increase of the data dispersion at constant centroids.

2 Used methods

2.1 Simulation of the data

In order to implement the research properly we first had to prepare the data, on which we could test the chosen clustering method. The advantage of usage of simulated data (in comparison with real data) is that we can choose the type of distribution and values of parameters – by doing that we remove the influence of eventual deviations. We usually use normal distribution, which is the presumption of most methods in multivariate data analysis.

We used program package R, with which we can create random normally distributed sets of data in multi-dimensional space. In order to stay in dimensions, which are still imaginable, we decided to create data in three-dimensional space, which means that we used three variables. The values at variables (x, y and z) were mostly set inside the interval (1, 5), which reminds us of the data sets from statistical research with ordinal variables (ranked from 1 to 5). All centroids were set near the point (3, 3, 3).

When we simulated random normally distributed numbers with a program, we had to define two basic parameters of distribution: vector of average values (centroid) and variance-covariance matrix ($\boldsymbol{\Sigma}$). In our case we changed only the variance-covariance matrix and defined the additional limitation that the average

values of generated numbers are exactly equal to initial centroids. At variance-covariance matrix, which also determines the shape of distribution in space, we chose only two simple cases: spherically symmetrical and ellipsoidal shape. In case of spherically symmetrical distribution $\mathbf{\Sigma}$ is a diagonal matrix with the same diagonal elements: $\Sigma_{ii} = \sigma^2_{xx} = \sigma^2_{yy} = \sigma^2_{zz}$. The shape of distribution is actually determined by only one parameter: σ^2_{ii} (dispersion).

In case of ellipsoidal distribution we also chose the simple example where the longer axis of ellipsoid is parallel with one of the diagonals of unit cube in first quadrant of the co-ordinate system. In that case $\mathbf{\Sigma}$ is symmetrical matrix with the same diagonal elements while the off-diagonal elements are different only by their signs: $\Sigma_{ij} = \pm \rho\sigma_i\sigma_j$, which determine the direction. So besides + or – sign there are two basic parameters: dispersion σ^2_{ii} and the correlation coefficient ρ, which determines the width of the ellipsoid.

2.2 The *k*-means clustering method

The *k*-means clustering method is one of the most widely used methods for data clustering in statistics. With its help we can cluster the units from the sample in any number of relatively homogenous groups. The method uses iterative algorithm, which finds the best group centroids (with iterations) and defines the ideal combination of clustering – this is the combination, at which the criterion function has one of its minimums (optimization).

The number of groups and initial cluster centers must be predefined. The k-means clustering method was started in five different ways (on this basis we later determined the performance of the method in comparison with initial cluster centers):

1. With random initial cluster centers (by default) set by the program, at which units are sorted the same way as they were simulated (by group membership – first 500 units from the first group, then 500 units from the second group etc.) – in continuation we are using the expression "ICC 1".
2. With random initial cluster centers (by default) set by the program, at which units are assorted with help of random numbers – in continuation we are using the expression "ICC 2".
3. With initial cluster centers, which we got from hierarchical clustering (Ward's method) – in continuation we are using the expression "ICC 3".
4. With initial cluster centers, which represent real group centroids – most favourable ICCs – in continuation we are using the expression "ICC 4".

5. With initial cluster centers, which lie exactly between the real group centroids – unfavourable ICCs – in continuation we are using the expression "ICC 5"[4].

When we repeat the same procedure with different initial cluster centers (ICCs) we can determine the best solution by simply observing which one has the minimum value of criteria function. First three ICCs are part of standard *k*-means clustering procedure while last two ICCs serve only for control purposes because they are only hypothetical for real data with unknown centroids.

At *k*-means clustering method the criterion function has following form:

$$P(\varphi) = \sum_{C \in \varphi} p(C)$$

where φ are clusters (C) from a set of possible clusters. The sum of squared Euclidean distances d of each unit X from the cluster center T_C:

$$p(C) = \sum_{X \in C} d^2(X, T_C)$$

represents a certain "error", which arises at union of units – it is also called "Error sum of squares of the cluster C" or "Inertia of the cluster C". The described criterion function can be found in literature under the name of "Ward's criterion function". We decided that at evaluation of clustering performance we also take into consideration generalized Ward's criterion function:

$$P(\varphi) = \sum_{C \in \varphi} \sum_{X \in C} d\ (X, T_C)$$

where d is in general a dissimilarity. Batagelj (1985) had proved generalization of Ward's criterion function from squared Euclidean distance to any dissimilarity. His mathematical proof goes beyond the extent of this paper. In our case generalized Ward's criterion function with Euclidean distance as dissimilarity was calculated for comparison and as an alternative criterion.

[4] Because at the last structure we have only two groups both initial cluster centers at ICC 5 should be in the same spot (between both real group centroids). Because of this fact we manually set two different initial cluster centers, which still lie in the middle of both group centroids but in different spots. Both points have the same distance to both real group centroids and therefore still represent unfavorable initial cluster centers.

2.3 Graphical representation of clustering

For graphical representation of data and clustering we used statistical program package R. The program enables the representation of three-dimensional groups in two dimensions with help of three-dimensional co-ordinate system. In co-ordinate system, in which every dimension represents one variable, we can then represent three-dimensional group with combination of dots, colours and shading (or with different black signs). The group consists of dots, which are relatively close to each other, and is surrounded with empty space or with single dots. The spinning of co-ordinate system helps us to identify the multi-dimensional form of the group (Ferligoj, 1988; Bogosavljević, 1988; Ferligoj, 1989).

2.4 Evaluation of performance of the clustering

After clustering was finished we had to evaluate the performance of clustering with adequate methods. When observing the results of clustering we especially focused on three quantities that we chose as indicators of performance of the method:

- percentage of correctly clustered units,
- centroids dislocation (dislocation of group means from real centroids) and
- values of criteria function.

The percentage of correctly clustered units was calculated with program Microsoft Excel where we used data from SPSS. We compared the real group membership (this was the variable that we created) with membership, which was set by *k*-means clustering method using different ICCs.

We measured centroids dislocation in Euclidean distance, calculated for i-th cluster with formula:

$$\Delta_i = \sqrt{(x_{T_{ci}} - x_{T_i})^2 + (y_{T_{ci}} - y_{T_i})^2 + (z_{T_{ci}} - z_{T_1})^2} ,$$

where T_i is the real (as simulated) centroid and T_{ci} the centroid of i-th cluster.

The value of generalized Ward's criterion function was calculated in SPSS as a sum of Euclidean distances between units and cluster centers T_{ci}:

$$P(\varphi) = \sum_{i=1}^{3} \sum_{j=1}^{N_i} \sqrt{(x_{ij} - x_{T_{Ci}})^2 + (y_{ij} - y_{T_{Ci}})^2 + (z_{ij} - z_{T_{Ci}})^2} .$$

The formula for standard Ward's criterion function is the same as previous, but without the square root.

3 Results

3.1 Preparation of the data

From variety of possible structures and different data we chose three different structures and for each of them we prepared four sets of data with different overlapping of the groups:

- first structure was named "spherical distribution" or "OOO structure" - it consists of three groups of spherically symmetrically distributed units with the same dispersion of the data;
- second structure was named "asymmetrical distribution, CCC structure" – it consists of three groups of ellipsoidally distributed units, which main axes are not parallel to each other. All three groups have the same dispersion and the same correlation coefficient $\rho = 0.8$, they differ only by their orientation, which is set by + or – sign of off-diagonal elements of symmetrical variance-covariance matrix:
  - for first group: $\Sigma_{xy} = \Sigma_{xz} = \Sigma_{yz} = + \rho\sigma^2$;
  - for second group: $\Sigma_{xy} = \Sigma_{yz} = - \rho\sigma^2$; $\Sigma_{xz} = + \rho\sigma^2$;
  - for third group: $\Sigma_{xz} = \Sigma_{yz} = - \rho\sigma^2$; $\Sigma_{xy} = + \rho\sigma^2$.

- third structure was named "asymmetrical distribution, BO structure" – it consists of two groups, from which one is spherically symmetrical and the other has curved form, which embraces the first group from one side. This (banana shaped) group was created with integration of nine spherically symmetrical subgroups.

Each group consists of 500 units. So at first and second structure together there are 1500 units while at the third structure there are only 1000 units. Centroids are fixed and have following values of co-ordinates:

- 1. centroid: (3, 2.5, 3)
- 2. centroid: (2, 2, 2.5)
- 3. centroid: (2.5, 3, 3.5)

The distances between centroids are: 0.87, 1.22 and 1.5.

For all three structures we created four data sets with following values of dispersion:

$\Sigma_{ii} = 0.04$ $(\sigma_{ii} = 0.2)$, for i = x, y, z;
$\Sigma_{ii} = 0.10$ $(\sigma_{ii} = 0.32)$,
$\Sigma_{ii} = 0.16$ $(\sigma_{ii} = 0.4)$,
$\Sigma_{ii} = 0.25$ $(\sigma_{ii} = 0.5)$.

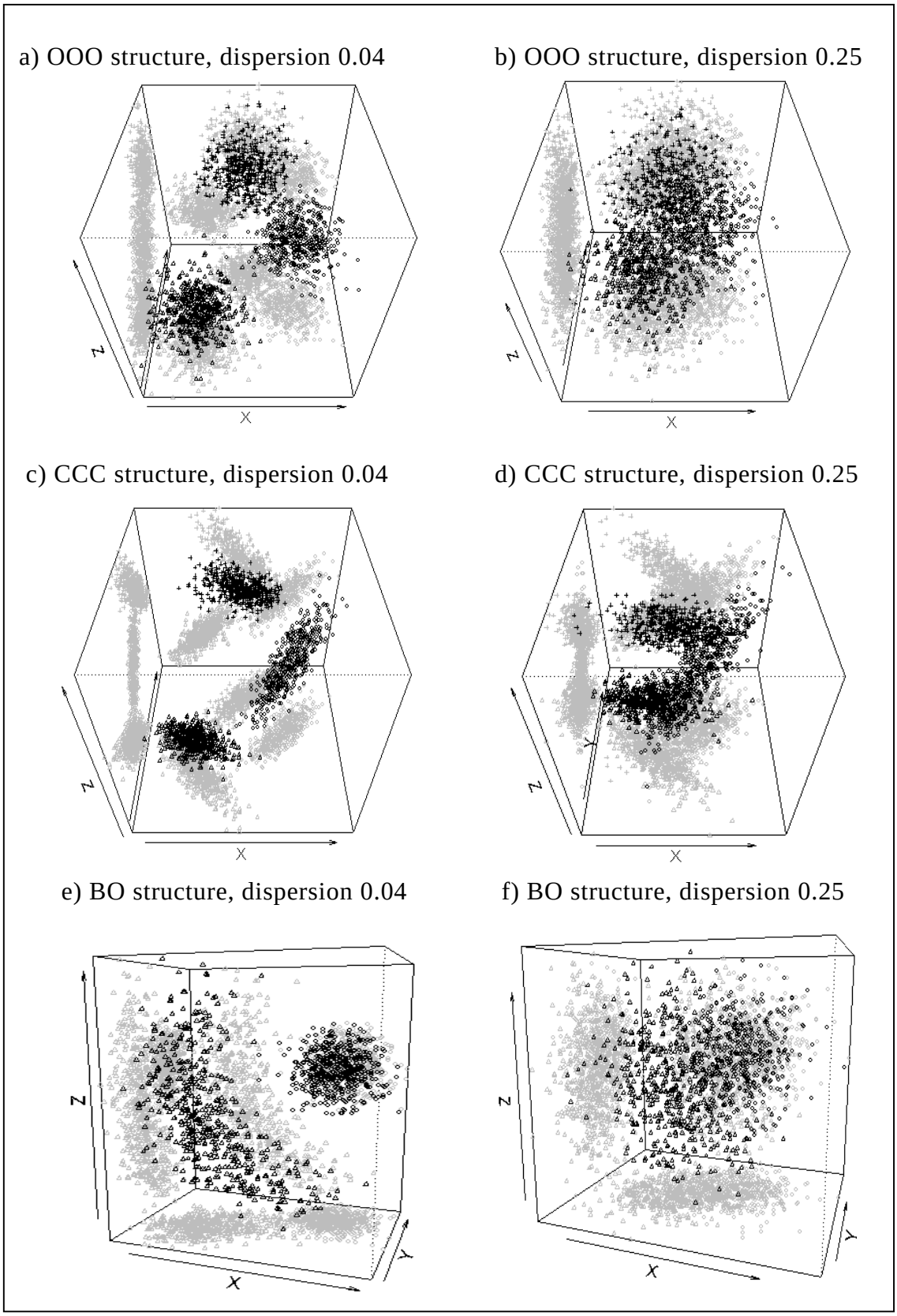

Figure 1: Simulated data for all three structures (at lowest and highest dispersion only).

At the lowest dispersion the groups are well separated while at dispersion value of 0.25 they are already very overlapping. At initial trials with dispersion values at spherical groups we found out that due to strong overlapping of groups the data with dispersion values over 0.25 is not appropriate for our analyses.

Simulated data is shown on Figure 1. For brief illustration we are here showing only simulated data with lowest and highest dispersion values.

3.2 Clustering

On each prepared set of data we did one hierarchical (Ward's method) and five different ways of *k*-means clustering in SPSS (with five different initial cluster centers - ICCs). The hierarchical clustering of data sets (Ward's method) serves for comparison and as the source for initial cluster centers for some modes at *k*-means clustering method. First three ICCs are part of standard *k*-means clustering procedure, while last two ICCs serve only for control purposes:

- With real centroids as the best possible ICCs the best possible solution should be found.
- Very unfavourable ICC 5 can indicate some instability of the method.

At the end of clustering we calculated and compared three already mentioned indicators. While the search of minimum of criteria function is already installed in the method, the centroids dislocation and the percentage of correctly clustered units is known only when we are familiar with real groups. In reality we are usually not and therefore we can observe only criteria function and differences in centroids positions between single results.

In general the results of five repetitions of clustering are equal or almost the same for each set of data. The values of indicators are the same or vary for some ‰ when observing percentage of correctly clustered units or even less when we observe value of criteria function. Such differences origin from a few units that change the cluster and we can consider these results as equivalent. The differences are significant only for some exceptional results, where groups are quite different. Rather than representing the tables, the values of indicators are added to the figures, which are more appropriate way for representation of the results for our three-dimensional case. Values of indicators, which are shown on figures, are properly rounded off and are represented with following symbols:

- η – the percentage of correctly clustered units;
- Σd – value of generalized Ward's criteria function;
- Σd^2 – value of standard (or Ward's) criteria function.

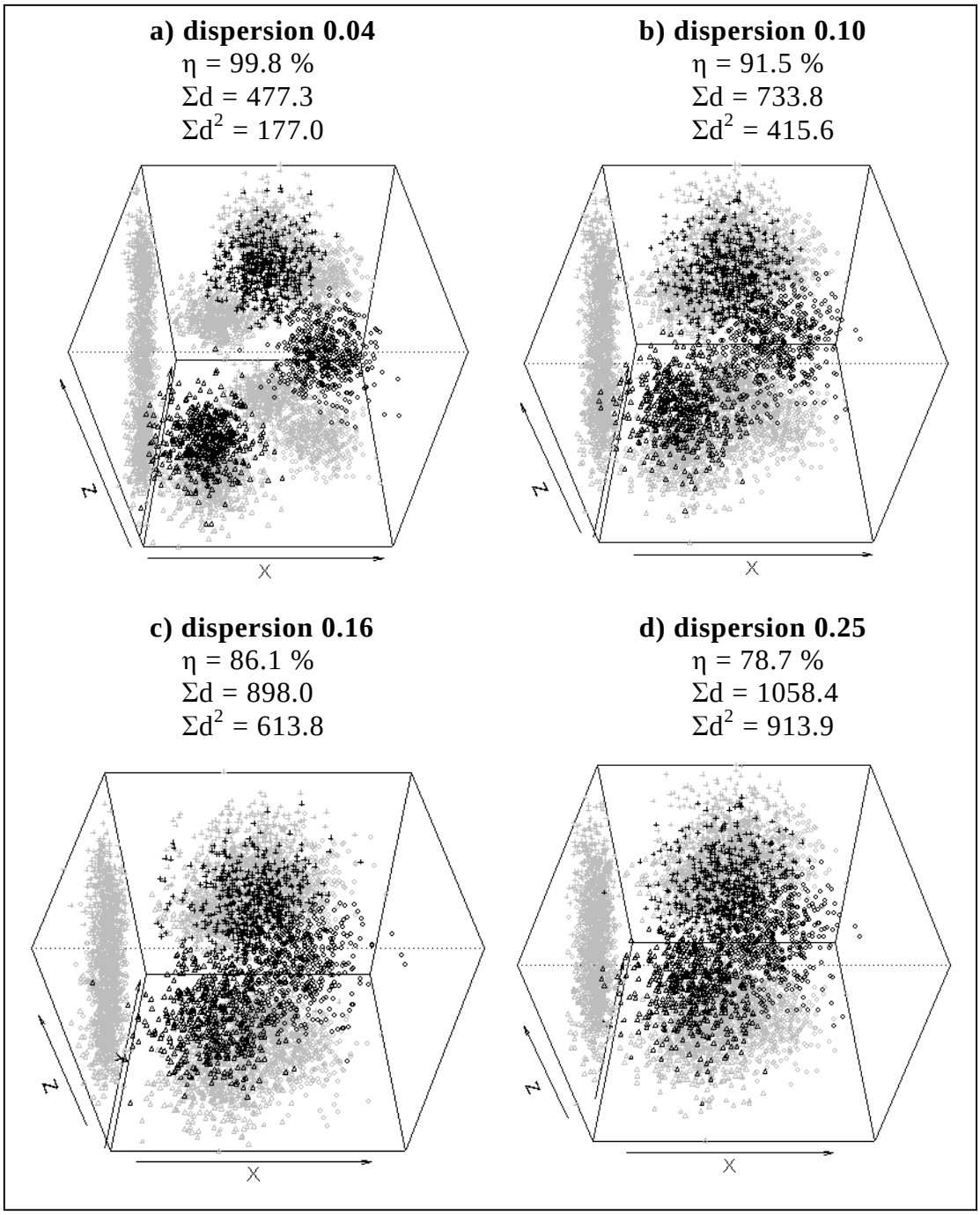

Figure 2: *K*-means clustering results at OOO structure (typical results with indicator values).

3.3 Results of clustering for spherically distributed data (OOO)

On graphical representations of clustering (Figure 2) we can recognize the basic characteristics of *k*-means clustering. The number of units in groups is nearly constant and units from single groups are not mixed between each other. Units from area between two groups are clustered into nearest group and the border between such groups is a plane. Incorrectly clustered units are lying around the two bordering planes between three groups (Figure 5b). With decreasing dispersion and overlapping of the groups the percentage of correctly clustered units is decreasing – for the greatest dispersion this value is only 78.7 %.

At the smallest dispersion (0.04) the first and the second group are still isolated while the first and the third group already partly overlap. For four different ICCs we get same clustering with only 18 incorrectly clustered units (1.2 %), which all lie in the area between the groups (Figure 3a). With unfavourable ICCs (ICC 5) we got completely wrong clustering (Figure 3b), at which two groups merged into one and the third group was divided into two new groups. In continuation we didn't get any further clustering results that would be so "bad".

On Figure 4 we can see significant difference between *k*-means and hierarchical clustering (Ward's method). Units in the border region of two neighbour groups are mixed and the percentage of correctly clustered units is lower. This happens at all dispersion values for OOO structure.

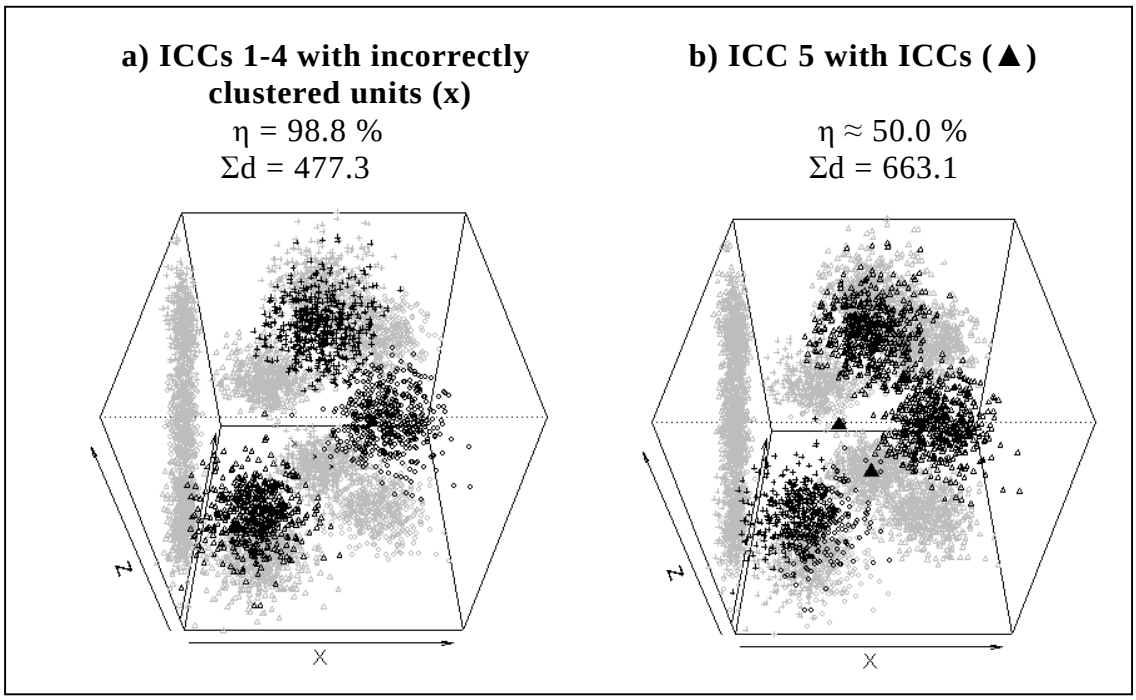

Figure 3: Comparison of typical and exceptional clustering at OOO structure for dispersion 0.04.

With ascending dispersion and overlapping of the groups (Figure 6) also the co-ordinates of centroids are changing - centroids are getting more and more distant. The largest deviation is at first centroid, which at the largest dispersion moved for 0.173 to the right. On Figure 6 deviations of the centroids are marked with sign Δ.

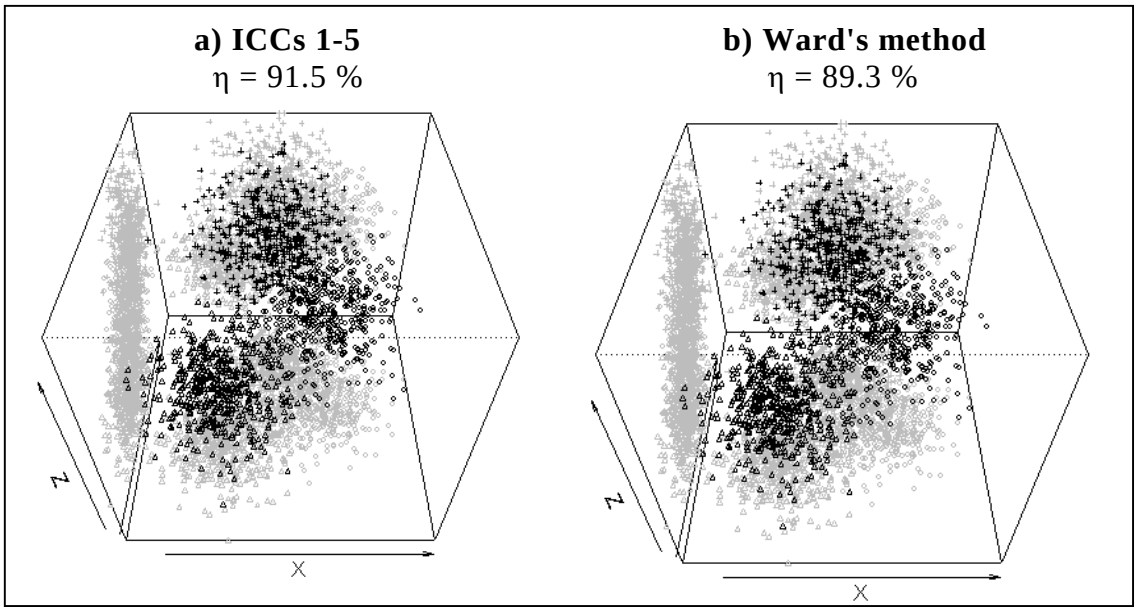

Figure 4: Comparison of two different clustering methods at OOO structure for dispersion 0.10

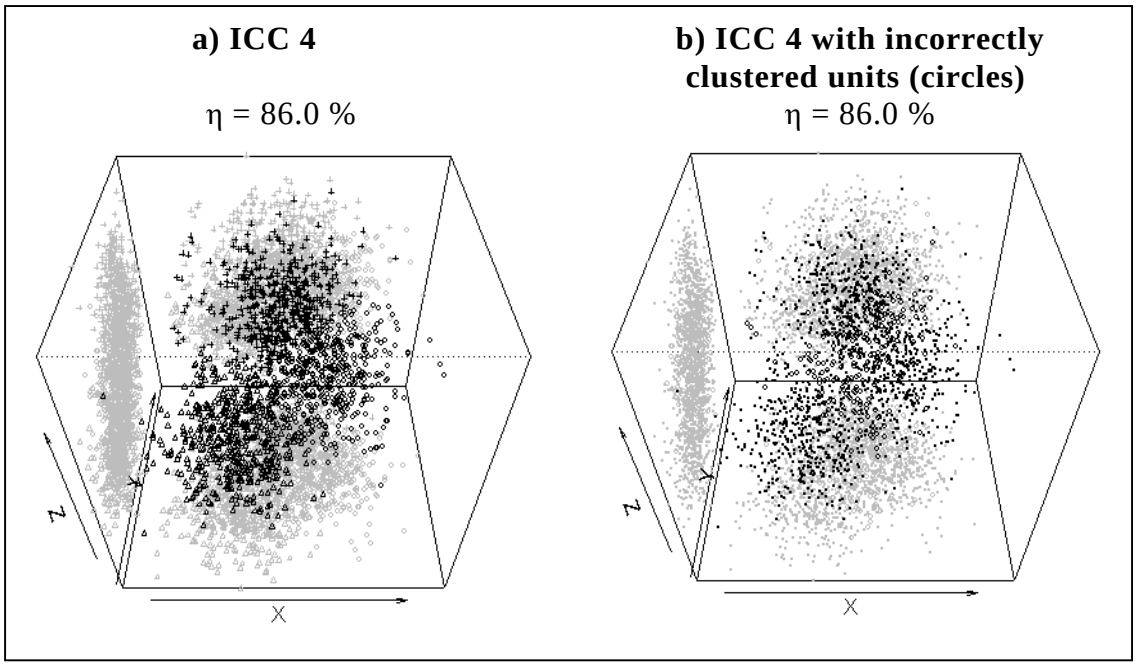

Figure 5: Results of clustering at OOO structure for dispersion 0.16.

With the values of criteria functions we got the same valuation of clustering as we got with help of indicator that measures the correctness of clustering. If we neglect the small differences between similar results, which we got with different ICCs, then all values of criteria functions are equal (at the same value of dispersion). When the dispersion values are growing the values of criteria function are increasing. For completely "missed" clustering result at dispersion 0.04 (ICC 5) we perceived very increased value of criteria function.

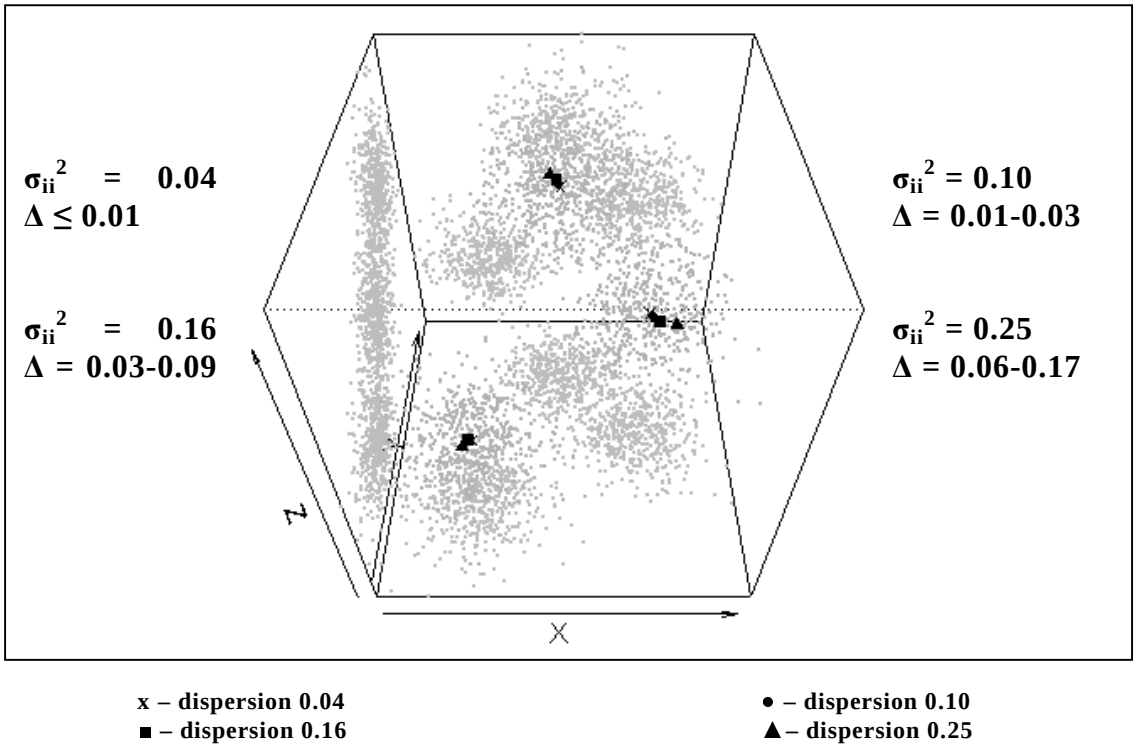

Figure 6: Centroids dislocation at OOO structure.

3.4 Results of clustering for asymmetrically distributed data (CCC)

Already a short sight on figures, which represent clustering for CCC structure (Figures 7-11), shows that the results of clustering are very different than for spherically distributed data. Groups are taking units from each other and we can not define a general rule for this process because at dispersion 0.16 it is completely different than by other dispersions. At the smallest dispersion we have only 12 incorrectly clustered units (0.8 %) while at higher dispersion values the performance of clustering is decreasing quicker than by spherically symmetrical distribution, which is consequence of very difficult configuration in this case.

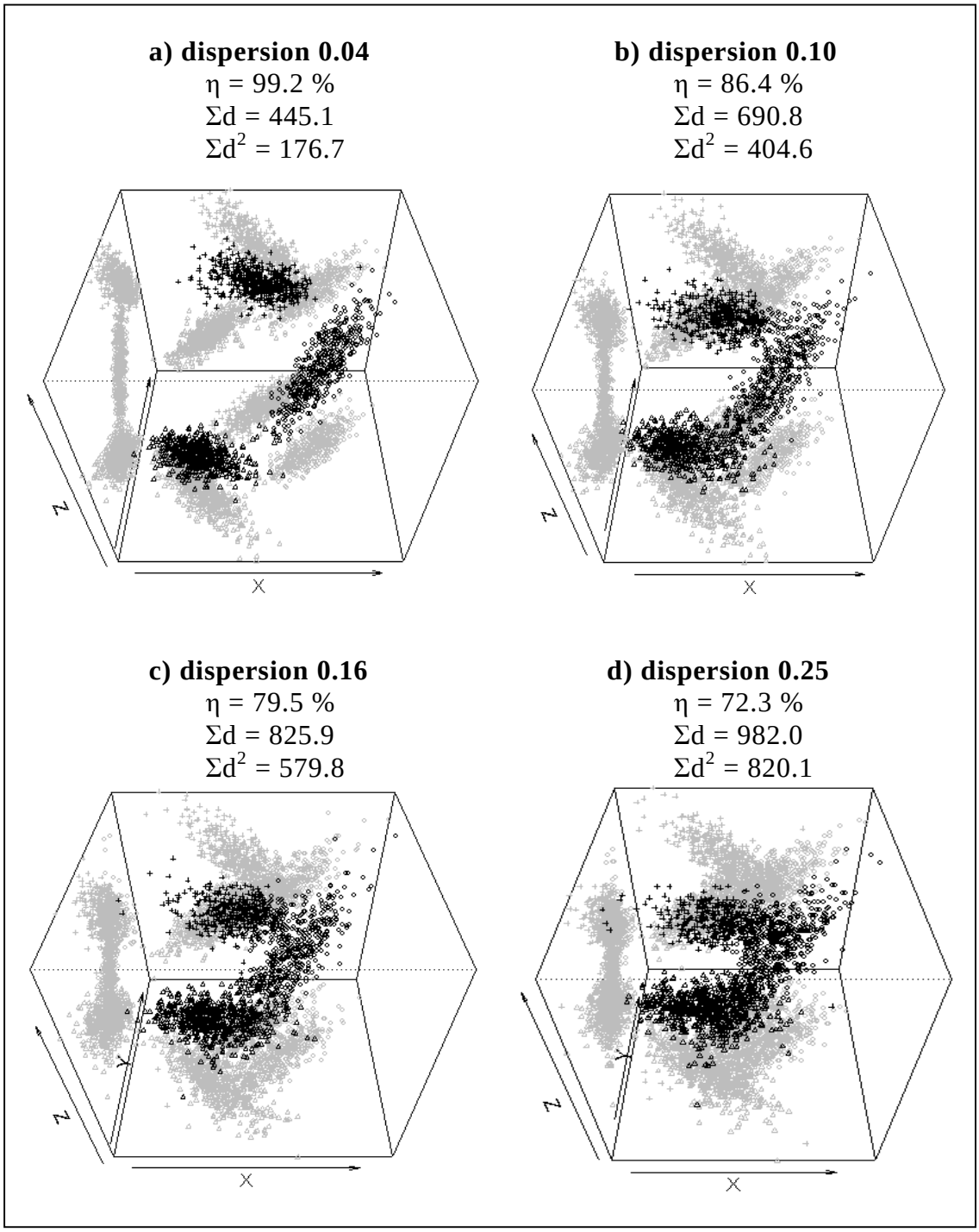

Figure 7: *K*-means clustering results at CCC structure (typical results with indicator values).

Already at the smallest dispersion (0.04) we can see the tendency of units take-over because second and third group already take-over some of the most exposed units from first group.

At dispersion value of 0.10 the take-over of the units is already strong. The border where the second group took over one part of units from first group is set much further from the area, on which there was actual mixing of the units from both groups (Figure 8a). It is harder to explain why the first group took over one part of units from the third group because on this area these groups are still relatively well isolated what can be seen by the top view (Figure 8c). On Figures 8c and 8d the incorrectly clustered units are marked with x.

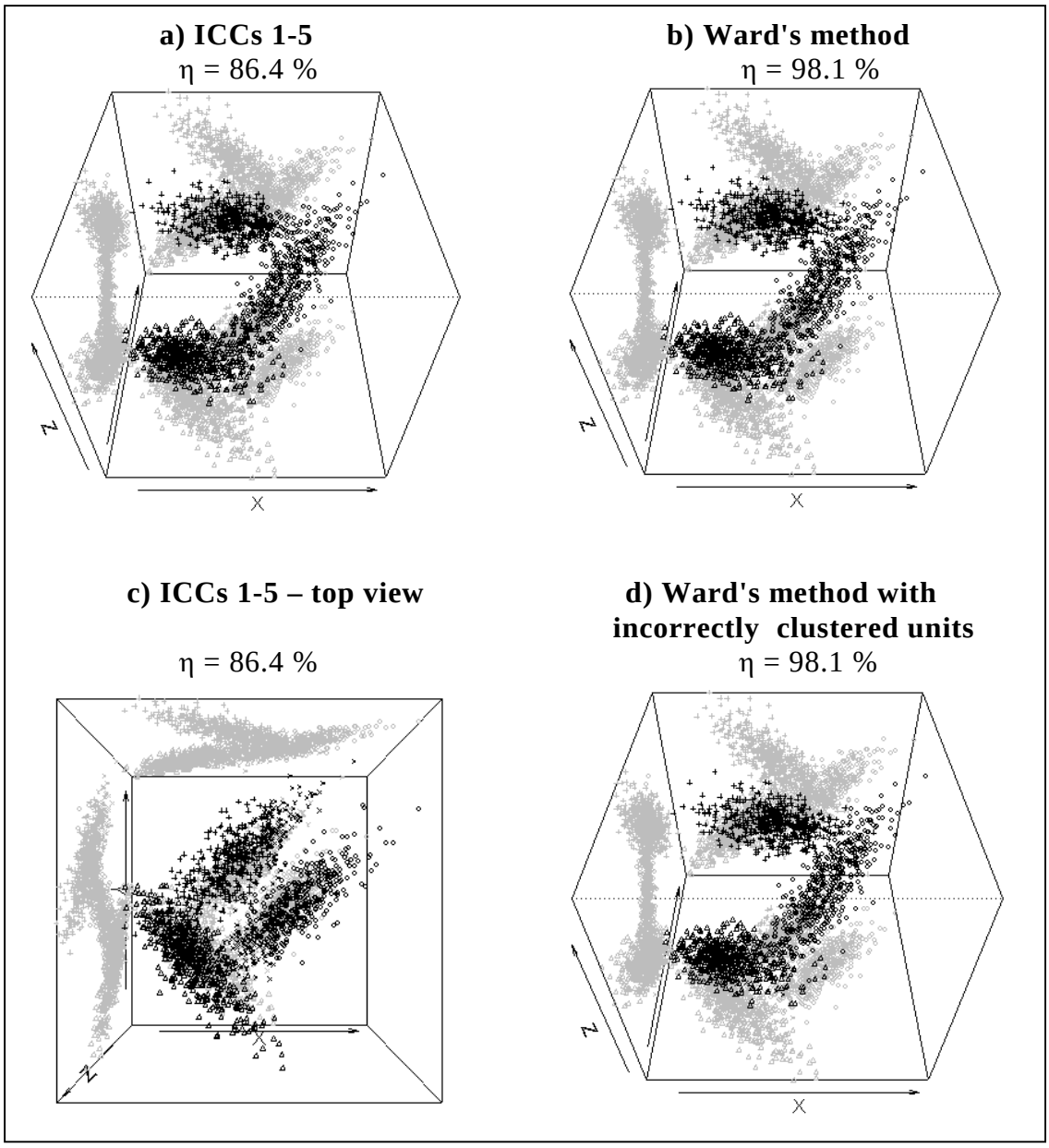

Figure 8: Comparison of typical *k*-means and hierarchical clustering (Ward's method) clustering at CCC structure for dispersion 0.10.

At dispersion value 0.16 where first and second group overlap quite heavily with their ends and first and third group come together we got two very different clustering results. Four times we got almost identical clustering, which is totally different from the previous case (Figure 9a). Now the third group took over half of the first one and the first took over almost half of the second group. Therefore we have one third of incorrectly clustered units (33.3 %). Only in one clustering (ICC 4) we got much better result, which is similar to previous case – in total we have 20.5 % of incorrectly clustered units (Figure 9b).

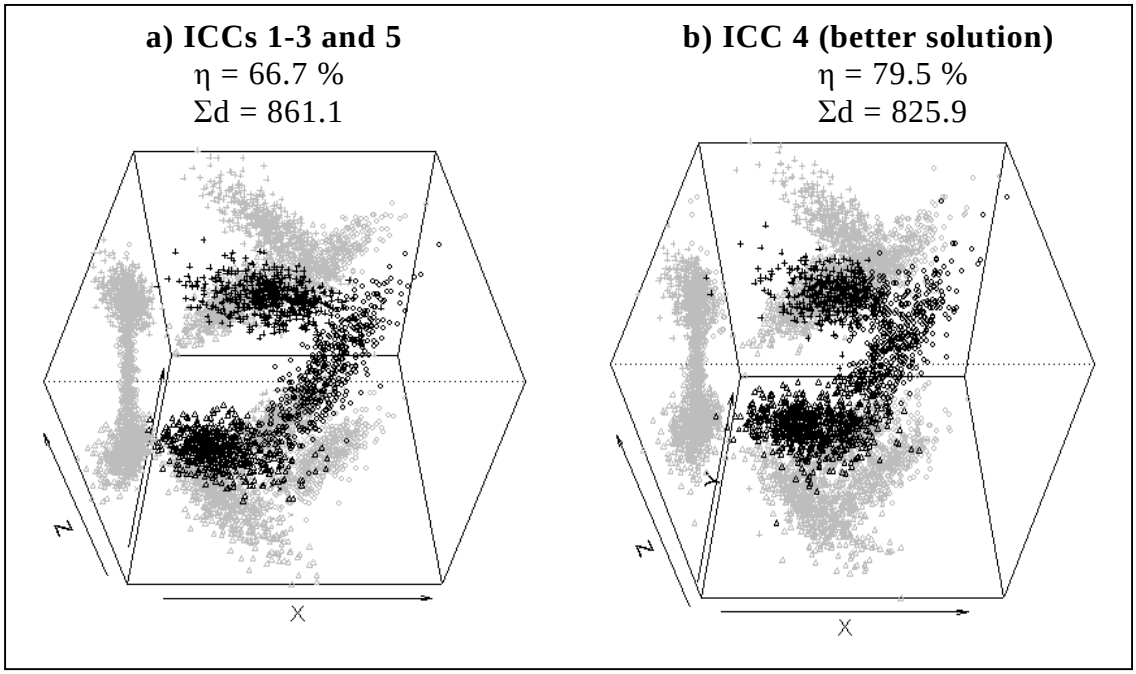

Figure 9: Comparison of typical and exceptional clustering at CCC structure for dispersion 0.16.

At the biggest dispersion value (0.25) we once again got better results at four ICCs and worse result at one of the ICCs, which is again similar to previous case (Figure 10).

In comparison with *k*-means clustering method the performance of hierarchical clustering (Ward's method) was better for lower dispersion values (Figure 8), while at higher dispersion values *k*-means method's performance is better.

At clustering where we had more take-overs of units and changes at numbers of units in single groups we noticed large centroid dislocations, which can be seen with assistance of graphical representation (Figure 11). At the smallest dispersion value the dislocation of the centroids is negligible but already at value of 0.10 this is not so. The centroids have moved in the counter-clockwise direction a little further away from the axes of single groups towards the areas where the take-overs took place. The strongest dislocation (one third of distance between the first and

the third centroid) was at centroid of the first group, which retained the least of its initial units. Deviations of the centroids are marked with sign Δ.

At dispersion value of 0.16 the dislocations are similar and slightly bigger only for one ICC. At worse results of clustering, which we got four times, the centroids moved in clockwise direction, which is in line with direction of take-over of the units. Dislocations are proportionately bigger according to number of incorrectly clustered units and the centroid of the first group is with dislocation length of 0.76 closer to the second centroid. At the largest dispersion value the dislocations at both results of clustering (at better and worse one) are larger in the same direction as by previous case.

Also at this structure the valuation of clustering on the basis of both criteria functions corresponds with correctness of clustering. The values of criteria function are increasing when there is an increase of dispersion. At much worse clustering results by dispersions 0.16 and 0.25 the values of criteria functions are heavily increased.

At this structure we perceived certain instability by clustering results for larger values of dispersion. This can be seen especially at dispersion value of 0.16, where we got the only better result because the chosen ICCs were very favourable. Visibly worse results represent local minimum of criteria function, which can not be easily evaded.

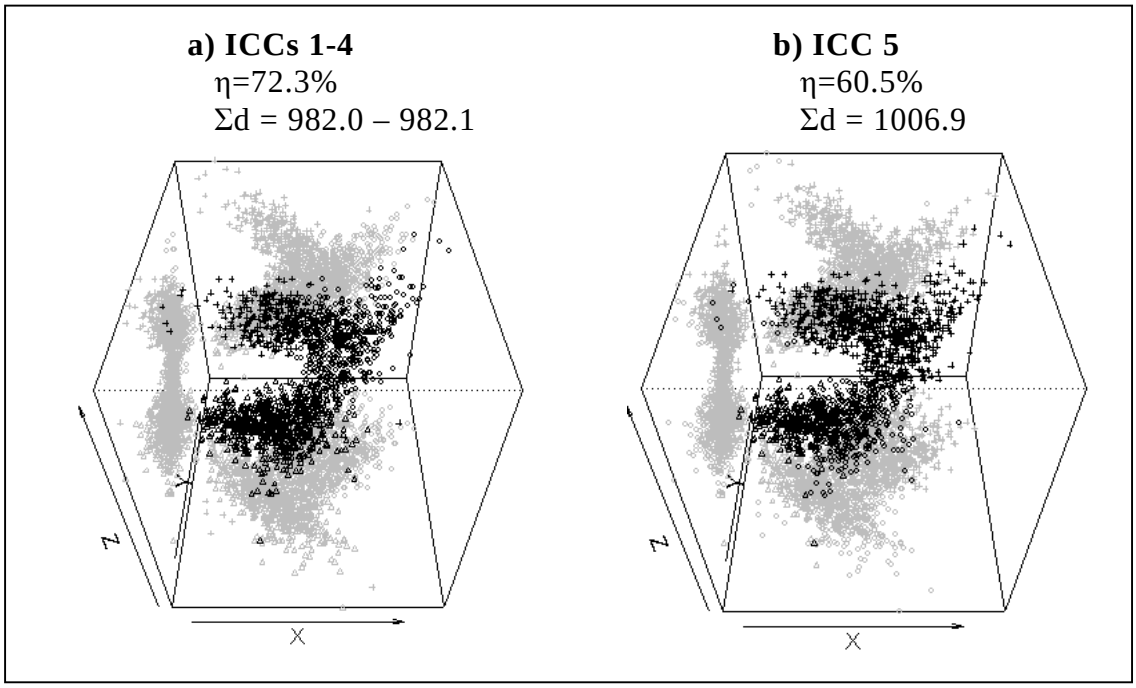

Figure 10: Comparison of typical and exceptional clustering at CCC structure for dispersion 0.25.

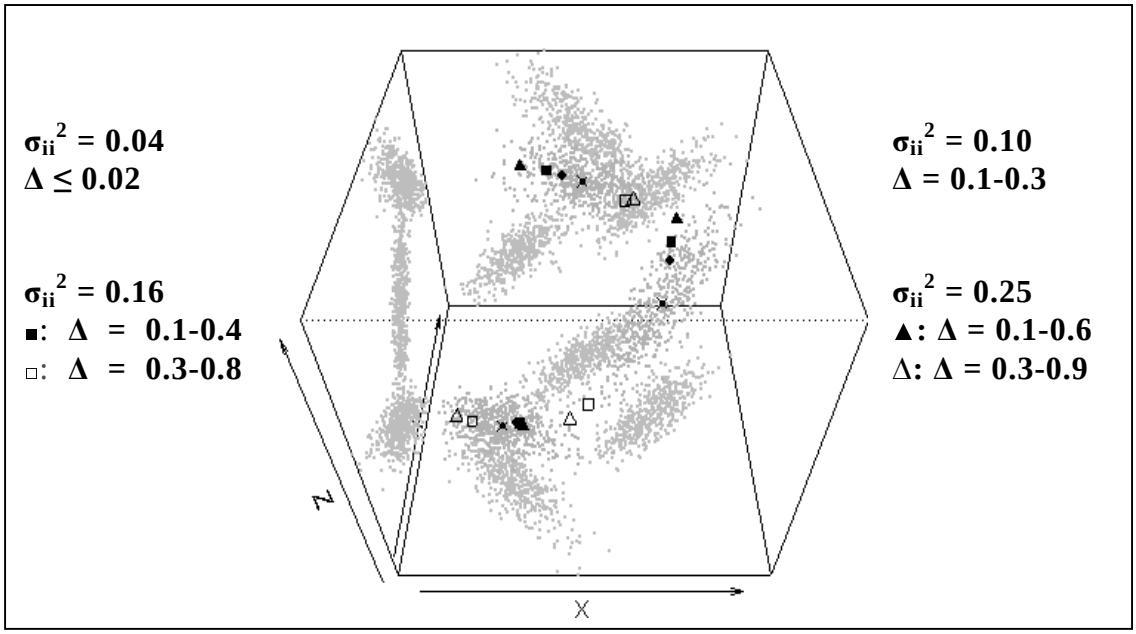

Figure 11: Centroids dislocation at CCC structure.

3.5 Results of clustering for asymmetrically distributed data (BO)

For BO structure the results of clustering (Figures 12-14) are similar to clustering results of spherically distributed groups. Again we got the demarcation plane between both groups, which we can prove with help of top view (Figure 14b). Smaller difference is visible only at position of incorrectly clustered units, which are mostly from spherically symmetrical group in the middle part and from curved (banana shaped) group at the edges. Percentage of correctly clustered units is decreasing from 98.8 % at the lowest dispersion (where both groups are still isolated) to 88.0 % at the largest dispersion. The percentage of correctly clustered units is in average higher than by spherical structure.

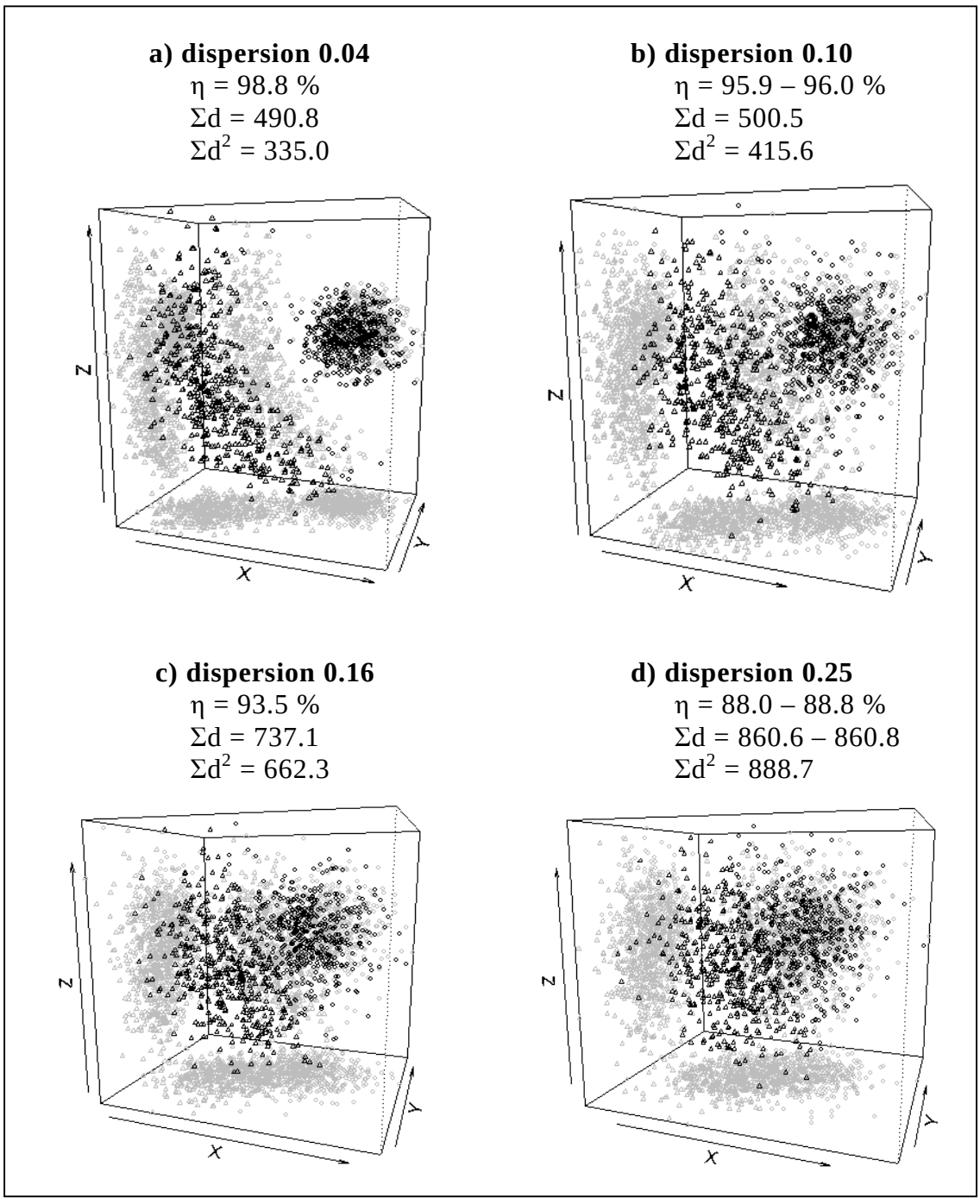

Figure 12: *K*-means clustering results at BO structure (typical results with indicator values).

Hierarchical clustering (Ward's method) proved itself as more successful at the lowest dispersion (100 % of correctly clustered units), at next two dispersion values it is roughly equal (Figure 13) and at the highest dispersion it is already much less successful than the *k*-means method.

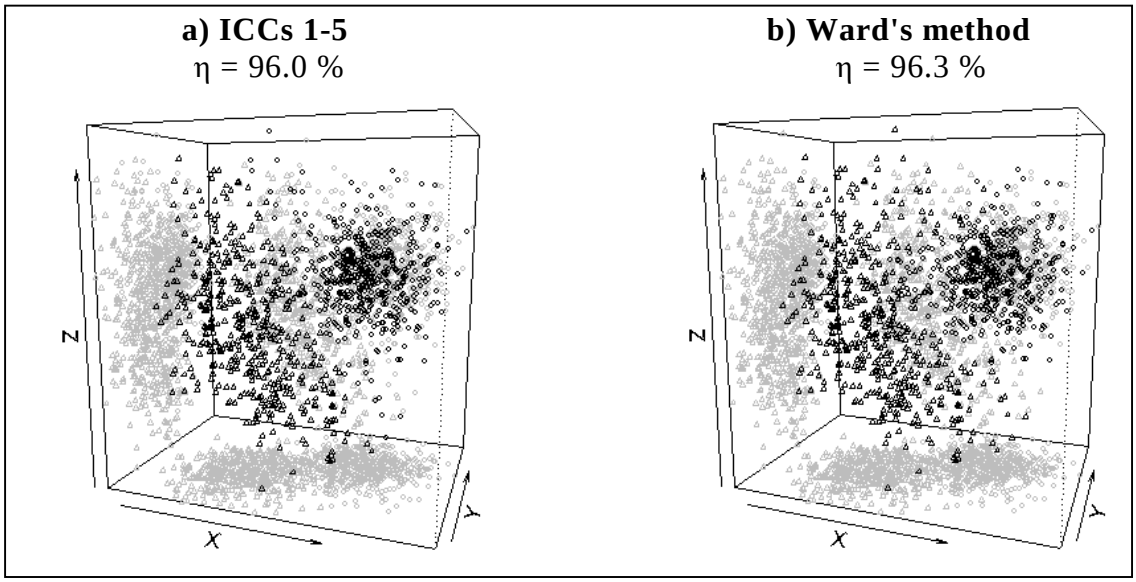

Figure 13: Comparison of typical *k*-means and hierarchical clustering (Ward's method) at BO structure for dispersion 0.10.

At BO structure the centroids dislocations are even smaller than by OOO structure in case we use the best clustering result. At the largest dispersion both centroids move only for 0.12. Values of criteria function are increasing when there is an increase of dispersion.

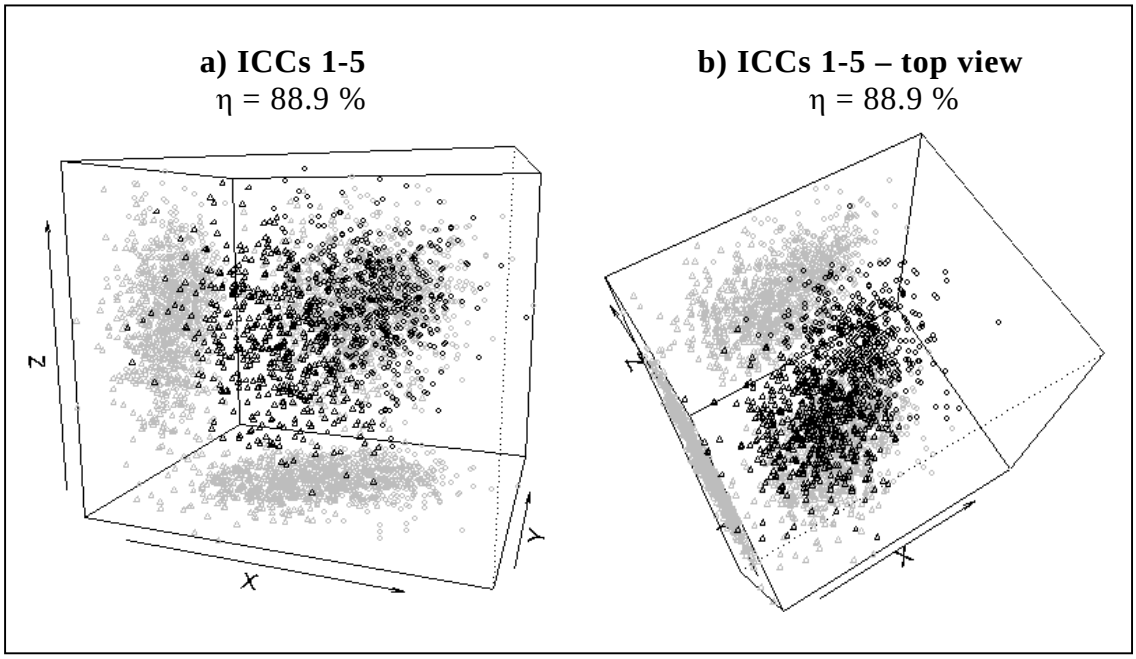

Figure 14: Typical *k*-means clustering at BO structure for dispersion 0.25.

3.6 Final evaluation of clustering results

Common characteristics of clustering of units into groups at spherically symmetrical and asymmetrical distribution (BO structure) are the distribution of units around the demarcation plane between two groups and smaller changes in number of units per group. The reason for such distribution is minimization of the criteria function, which value is always larger for the real clustering. The difference between its value and found minimums is rapidly increasing with increase of dispersion.

When the units from different groups begin to mix the best demarcation for clustering of these units is exactly between both centroids. By minimizing the criterion function the method finds a plane, which lies between both groups and divides units of one and other group. For any other method of clustering, for instance if we provide additional information and increase the percentage of correctly clustered units, the value of criterion function would be higher. We can also perceive that values of criterion functions are always higher or equal at actual distribution of the data than by any *k*-means clustering. All units that lie beyond the demarcation plane contribute more to the value of the criteria function, when they are taken into account with their own group as when they are clustered incorrectly to the nearest centroid. At hierarchical clustering (Ward's method) we got partly mixed clustering of units but the percentage of correctly clustered units, which is much lower, shows us that the method's performance is worse on the area where the units from different groups are mixing.

Similarity of clustering results at BO structure with spherically symmetrical structure is in its partial symmetry because there are two symmetrical planes that are rectangular on each other and are defined by the plane, which connects both centroids. One of the reasons for slightly better percentage of correctly clustered units at BO structure is in the way the second group was created – centers of subgroups are 1.4 units away from first centroid, which is 15 % more than the distance between centroids of first and second group.

At CCC structure of asymmetrical distribution the clustering with *k*-means method had lower performance. Besides low percentage of correctly clustered units we also got results that were very unstable – some results were diametrically opposed. Just like at spherically symmetrical groups also at CCC structure the criteria function is the reason for the shapes of demarcation lines. The criteria function tends toward clustering of groups, which are as much rounded off as possible. In the case of two distinct stretched parallel groups the clustering with *k*-means method will have smaller value of criteria function for solution where both ends of groups are joined together as the clustering, which would be closer to actual structure.

By both structures that contain stretched groups the performance of hierarchical clustering (Ward's method) is better than performance of the *k*-means method when we have small values of dispersion (when the groups are more

clearly isolated). But when the dispersion gets larger the performance of hierarchical clustering (Ward's method) starts to fall rapidly.

The comparison of results, which we got with different ICCs, confirms known rules that clustering must be done several times. It is recommendable that first clustering is done with one of the hierarchical methods because like this we get number of groups and ICCs. If we observe only results that we got with ICCs 1, 2 and 3, which are most commonly used in praxis, we can see that (with exception of one clustering) we always got results, which are very close to best clustering. Without additional testing on data (with varied parameters) we can not say why there was an exception in clustering results at dispersion value of 0.16 at CCC structure. In that case ICC 4 was very useful because with its help we found a better solution. With ICC 5, which usually provided worst results, we illustrated possible unstableness of the method when ICCs are picked very unskilfully.

Because ICCs 4 and 5 are merely hypothetical and also the percentage of correctly clustered units is merely hypothetical tool the only indicators for measuring performance of clustering methods, which can be used at real problems, are criteria function and centroids dislocation.

Because at real problems we never know real centroids we can only monitor the differences in their positions by single clustering results. While the differences in position of centroids are only couple of percents at similar clustering results we can see that by worse results these differences can increase up to several ten percents, which is already comparable with average distance between centroids of the groups.

Values of criteria functions at similar clustering results are very similar and differ 0.2 % at the most while this difference is bigger than 2.5 % for exceptions. There aren't any bigger divergences at both criteria functions that we monitored (Σd^2 in Σd). Both show the same trend of growth (with increase of dispersion) and big increase at "worse" results. Only by some similar clustering results the minimal differences at values of criteria functions can be seen but these differences are mostly only on second or third spot behind the decimal point.

4 Conclusion

Our goal in this article was to evaluate the performance of the *k*-means clustering method on data with known structure and distribution. Although because of temporal and other limitations we tested only a couple of structures and parameter values we determined following characteristics of *k*-means clustering method:

- performance of clustering is decreasing in line with increase of dispersion;
- performance strongly depends on structure and distribution of the data;
- minimum of criteria function (either standard Σd^2 or generalized Ward's Σd) corresponds to the best clustering result.

For evaluation of clustering performance we used two indicators, the percentage of correctly clustered units and dislocation of centroids; these are parameters, which values are not known for actual data in real clustering problems.

On actual data we can monitor only value of criteria function and position of the centroids. Therefore small differences in values of criteria functions at single modes of clustering and minimal differences in positions of group centroids tell us that we are close to the best achievable result. Clustering exceptions (larger values of criteria function or large differences at centroid positions) indicate the instability of clustering and warn us that we are not necessarily close to ideal clustering. Therefore we have to take into consideration significant rule that the clustering should be repeated several times with carefully chosen ICCs. It is recommendable that we also include others, especially hierarchical methods of clustering (we can determine number of groups and ICCs), because in this research the performance of *k*-means clustering method is not always better than performance of other used methods.

The *k*-means method is most successful at spherically distributed structures. The performance is worse at asymmetrical stretched structures when the aspiration for rounded groups (which is dictated by minimization of criteria function) can bring to worse and unstable results of clustering. In such cases the hierarchical clustering (such as Ward's method) has better performance.

References

[1] Batagelj, V. (1985): Razvrščanje v skupine – nehierarhični postopki. Magistrsko delo. Ljubljana: FE.

[2] Batagelj, V. (1988a): Razvrščanje in optimizacija. *Zbornik radova*. Majski skup `87. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 19 – 27.

[3] Batagelj, V. (1988b): Local optimization method for the generalized Ward clustering problem. *Zbornik radova*, **2**. Majski skup `88. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 154 – 159.

[4] Bogosavljevič, S. (1988): Grafičko prikazivanje rezultata multivarijacione analize. *Metodološki zvezki*, **3**, Jugoslovansko združenje za sociologijo – sekcija za metodologijo in statistiko. Ljubljana: RI FSPN, 58 – 70.

[5] Dillon, W.R. and Goldstein, M. (1984): *Multivariate Analysis: Methods and Applications*. New York: Wiley, 157 – 208.

[6] Ferligoj, A. and Batagelj, V. (1983): *Razvrščanje v skupine – izbrane teme.* Ljubljana: RI FSPN.

[7] Ferligoj, A. (1986): *Metode za multivariatno analizo podatkov.* Ljubljana: RI FSPN.

[8] Ferligoj, A. (1988a): Razvoj in perspektive razvrščanja v skupine z omejitvami. *Zbornik radova.* Majski skup `87. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 44 – 51.

[9] Ferligoj, A. (1988b): Metode za grafično predstavitev multivariatnih podatkov. *Metodološki zvezki,* **3.** Jugoslovansko združenje za sociologijo – sekcija za metodologijo in statistiko. Ljubljana: RI FSPN, 47 – 57.

[10] Ferligoj, A. (1989): Razvrščanje v skupine. Teorija in uporaba v družboslovju. *Metodološki zvezki,* **4.** Jugoslovansko združenje za sociologijo – sekcija za metodologijo in statistiko. Ljubljana: RI FSPN.

[11] Ferligoj, A. (1998): Razvrščanje v skupine. Zapiski predavanj, Ljubljana: FDV.

[12] Hubert, L. and Arabie, P. (1985): Comparing partitions. *Journal of Classification*, **2**, 193 – 218.

[13] Jambu, M. and Lebeaux, M-O. (1983): *Cluster Analysis and Data analysis.* North – Holland Publishing Company, Amsterdam – New York – Oxford.

[14] Jerman, M. (1989): Vloga koeficientov v Lance-Williamsovi formuli pri razvrščanju v skupine. *Zbornik radova*, **3**. Majski skup `89. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 88 – 103.

[15] Jug, J. (1988): Metoda voditeljev v programskih paketih CLUSE in SPSS/PC+. *Metodološki zvezki*, **3**. Jugoslovansko združenje za sociologijo – sekcija za metodologijo in statistiko. Ljubljana: RI FSPN, 92 – 105.

[16] Jug, J. (1989a): Programi za razvrščanje v skupine v programskih paketih CLUSE, SPSS in SAS. *Zbornik radova,* **3**. Majski skup `89. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 126 – 153.

[17] Jug, J. (1989b): *Razvrščanje v skupine s SPSS/PC+*. Ljubljana: RI FSPN.

[18] Jug, J. (1989c): Razvrščanje binarnih enot s podprogramom CLUSTER v SPSS/PC+. *Metodološki zvezki,* **5**. Ljubljana: RI FSPN, 118 – 125.

[19] Koprivnik, S. et al. (1993): Comparative Studies Methodology – Clustering Approaches (The Case of Motivation Items in USA, Japan and Slovenia). *Metodološki zvezki*, **9.** Ljubljana: FDV, 149 – 163.

[20] Košmelj, K. (1985): *Razvrščanje enot opisanih s kratkimi časovnimi vrstami – II. del.* Ljubljana: RI FSPN.

[21] Vehovar, V. (1988): Tipologija raziskovalnih organizacij. *Metodološki zvezki*, **3**. Jugoslovansko združenje za sociologijo – sekcija za metodologijo in statistiko. Ljubljana: RI FSPN, 169 – 183.

[22] Zupan J. (1988): Hierarhično grupiranje velikih množic podatkov. *Zbornik radova.* Majski skup `87. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 165 – 173.