Department of Economics and Society, Dalarna University

A Classification of Belarusian Regions upon Economic Development Using Cluster Analysis

Sanja Kukolj, Dalarna University
Yu shu Li, Dalarna University
Viachaslau Naurotski, Dalarna University

Supervisor: Johan Bring, Dalarna University

C-level essay in Statistics, Fall 2006

Abstract

One of the main goals of this paper work is to classify regions of Belarus based on six variables indicating economic development of these regions taken for a period of five years from 2000 to 2004. The research covers 71 regions including 64 administration districts and 7 largest towns located within their borders. The aim of this work is not to classify regions for every of these years in particular but to make a data pool for all five years of observation in order to get a proper understanding of processes and tendencies taking place in these regions due to changes in their economic activities. As nowadays there exist many methods and measures available for conducting a cluster analysis a special procedure is proposed to define a best classification for the given partitions for different methods and metrics. As a result, in the research a new approach is suggested to determine the quality of different classifications. It was also uncovered that a choice of software may cause different classification results and particularly the dissimilarity between R and SPSS is discussed concerning their clusterization features.

Key words: cluster analysis, regions classification, Belarus regions, economic classification.

1 Introduction

The main aim of this paper work is to classify regions of Belarus according to values of their economic variables for a period of five years from 2000 to 2004. The paper deals with 71 provinces of Homiel, Mahileu and Minsk regions (64 administrative regions and 7 largest cities). Cluster analysis is chosen as one of the methods for classifying a number of objects into distinctive groups. Calculations and evaluating procedure are held using SPSS and R statistical software that suggest several methods of hierarchical clustering process along with different distance measures for each method. That is why the problem of deciding which of these methods and measures is more acceptable in our case arises and how to find the rule to compare these classifications.

Table 1: A list of Belarusian provinces included into the research.

Mahileu Region		Minsk and Minsk Region		Homiel Region	
1	Mahileu city	**24**	Minsk city	**49**	Homiel city
2	Babrujsk city	**25**	Barysau city	**50**	Mazyr city
3	Asipovicki district	**26**	Maladziecna city	**51**	Brahinski district
4	Babrujski district	**27**	Barysauski district	**52**	Buda-Kasaleuski district
5	Bialynicki district	**28**	Biarezinski district	**53**	Cacerski district
6	Bychauski district	**29**	Cervienski district	**54**	Chojnicki district
7	Causki district	**30**	Dziarzynski district	**55**	Dobruski district
8	Cerykauski district	**31**	Kapylski district	**56**	Homielski district
9	Chocimski district	**32**	Klecki district	**57**	Jelski district
10	Drybinski district	**33**	Krupski district	**58**	Kalinkavicki district
11	Harecki district	**34**	Lahojski district	**59**	Karmianski district
12	Hluski district	**35**	Lubanski district	**60**	Kastrycnicki district
13	Kasciukovicki district	**36**	Maladziecanski district	**61**	Lelcycki district
14	Kirauski district	**37**	Miadzielski district	**62**	Lojeuski district
15	Klicauski district	**38**	Minski district	**63**	Mazyrski district
16	Klimavicki district	**39**	Niasviski district	**64**	Naraulanski district
17	Krasnapolski district	**40**	Puchavicki district	**65**	Pietrykauski district
18	Kruhlanski district	**41**	Salihorski district	**66**	Rahaceuski district
19	Krycauski district	**42**	Slucki district	**67**	Recycki district
20	Mahileuski district	**43**	Smalavicki district	**68**	Svietlahorski district
21	Mscislauski district	**44**	Staradaroski district	**69**	Viatkouski district
22	Sklouski district	**45**	Staubcouski district	**70**	Zlobinski district
23	Slauharadzki district	**46**	Uzdzienski district	**71**	Zyktavicki district
		47	Valozynski district		
		48	Vilejski district		

Set of variables signifying a level of economic development of every particular region is the following:

- average salary
- retail turnover
- rate of unprofitable enterprises
- growth rate of industrial production
- sales profitability
- investment in accommodation facilities

We understand that the selected set of indexes is quite insufficient to obtain proper results in a particular sphere but as far as many parameters are unavailable in Belarusian official publications this set seems to be complete. Unfortunately, many indicators of the economic activities such as region's total output or investment in capital asset are unavailable in regional statistics but these would be very useful in our analysis.

As our 71 provinces are located within borders of Homiel, Mahileu and Minsk regions, in the following table we represent general descriptive statistics for all six regions of Belarus in order to give a brief description of the variables used in our paper. Mean data is provided in Table 2 for the observed period from 2000 to 2004 for six variables mentioned earlier.

Table 2: Descriptive statistics for Belarusian regions (mean values for 2000-2004 years according to [8]).

Region	Average salary (mln. Roubles to 1999)	Retail turnover (bln. R. per 10 000 people to 1999)	Unprofitable eterprises (%)	Growth rate of industrial production (chain, %)	Sales profitability (%)	Investment in accommodation facilities (ths.sq.m. per 10 000 people.)
Brest region	38.9	1.6	23.4[1]	108.5	10.2[1]	3.4
Homiel region	43.8	1.7	20.2[1]	109.3	21.1[1]	2.7
Hrodna region	39.8	1.7	20.1[1]	105.0	16.7[1]	3.3
Mahileu region	39.6	1.5	20.0[1]	104.6	4.1[1]	2.7
Minsk region	44.1	1.4	29.4[1]	108.1	14.7[1]	3.9
Vitabsk region	40.6	1.6	24.3[1]	106.1	15.1[1]	2.6
Total	41.3	1.6	24.5[1]	106.9	15.3[1]	3.1

As it was stated above the research work is based on SPSS and R tools for cluster analysis. We take under consideration seven different methods for classifying objects available in SPSS and six possible metrics for each of these methods thus the analysis involves 42 particular classification outcomes for 71 regions for five years (Squared Euclidean, Minkowski and a Customized distances are not included into analysis).

An analysis uses data taken from the official annual publications of the Ministry of Statistics of Belarus and also data obtained from the official websites of regional executive committees of districts that are under observation.

[1] Some of the mean values cannot be calculated due to unavailability of several statistical parameters. Values for the year 2004 are provided instead [8].

2 Brief description of clustering procedures

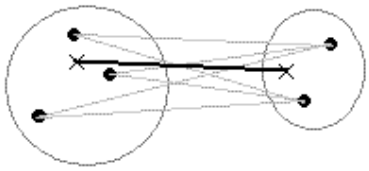
Cluster analysis is a collection of statistical methods, which identifies groups of samples that behave similarly or show similar characteristics. In common parlance it is also called look-a-like groups. The simplest mechanism is to partition the samples using measurements that capture similarity or distance between samples. In this way, clusters and groups are interchangeable words [6]. A cluster is a collection of objects "similar" to each other and are "dissimilar" to the objects belonging to other clusters. The similarity criterion is a between-objects distance: two or more objects belong to the same cluster if they are "close" according to a given distance [1]. All of the methods used in our work represent hierarchical clustering. Hierarchical cluster analysis is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted [1]. When there are N cases, this involves N-1 clustering steps. SPSS provides seven methods for hierarchical cluster analysis:

- Between-groups linkage
- Within-groups linkage
- Furthest neighbor
- Nearest neighbor
- Centroid clustering
- Median clustering
- Ward's method

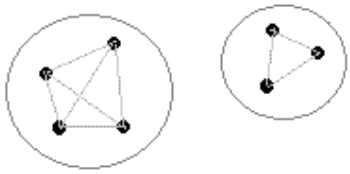
2.1 Methods used in the research

BETWEEN-GROUPS LINKAGE: The dissimilarity between clusters is calculated using cluster average values; of course there are many ways of calculating an average. The recommended one is UPGMA - Unweighted Pair-Groups Method with Arithmetic Mean [3]. The distance between any two clusters $\mathcal{A}$ and $\mathcal{B}$ is taken to be the average of all distances between pairs of objects x in $\mathcal{A}$ and y in $\mathcal{B}$. SPSS also provides two other methods based on averages, Centroid and Median.

$$D_{i,j} = \frac{1}{card\ A \cdot card\ B} \sum_{x \in A} \sum_{y \in B} d(x,y) \qquad (1)$$

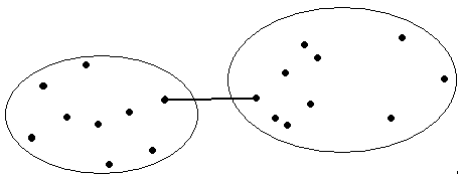


WITHIN GROUPS CLUSTERING: This is similar to UPGMA except clusters are fused so that within cluster variance is minimized. The distance is defined as the average of the distances between all pairs of cases in the cluster that would result if they were combined. This tends to produce tighter clusters than the UPGMA method [3].



NEAREST NEIGHBOR *(Minimum or Single linkage):* The dissimilarity between 2 clusters is the minimum dissimilarity between members of the two clusters. This method produces long chains which form loose, straggly clusters. This method has been widely used in numerical taxonomy [3]. The distance between two clusters $\mathcal{A}$ and $\mathcal{B}$ is the minimum distance between elements of each cluster (also called single linkage clustering):

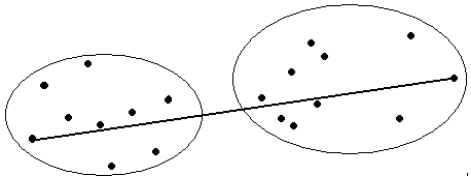
$$\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \,\} \qquad (2)$$



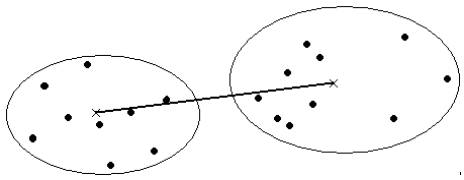
FURTHEST NEIGHBOR *(Maximum or Complete linkage):* The dissimilarity between 2 groups is equal to the greatest dissimilarity between a member of cluster *i* and a member of cluster *j*. This method tends to produce very tight clusters of similar cases [3].

The distance between two clusters $\mathcal{A}$ and $\mathcal{B}$ is the maximum distance between elements of each cluster (also called single linkage clustering):

$$\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \,\} \qquad (3)$$



CENTROID CLUSTERING (Unweighted Pair-Groups Method Centroid, UPGMC): uses the group centroid as the average. The average distance is calculated from the distance between each point in a cluster and all other points in another cluster. The two clusters with the lowest average distance are joined together to form the new cluster. The centroid is defined as the centre of a cloud of points. A problem with the centroid method is that some switching and reversal may take place, for example as the agglomeration proceeds some cases may need to be switched from their original clusters [3].



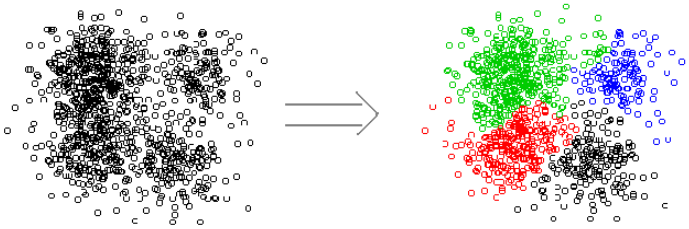
MEDIAN CLUSTERING: Clusters are weighted equally regardless of group size when computing centroids of two clusters being combined. This method also uses Euclidean distance as the proximity measure [4].

WARD'S METHOD (also known as *Minimum Variance)*: calculates the sum of squared Euclidean distances from each case in a cluster to the mean of all variables. The cluster to be merged is the one which will increase the sum the least. This is an ANOVA-type approach and preferred by some researchers for this reason [4]. The method searches for objects that can be grouped together while minimizing the increase in error sum of squares. Error sum of squares is computed as [5]:

$$ESS = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \qquad (4)$$

where:

x_i – the score of the i-th individual
n – number of objects in a cluster



Ward's method creates clusters of near equal size, having hyper spherical shapes.

2.1 Methods used in the research

There are several ways to work out the distance between two points in multi-dimensional space. The purpose of such measures is to give a numerical value to the amount of dissimilarity between two vectors. Measures used in this work are:

EUCLIDEAN DISTANCE: The straight line distance between two points

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \,. \qquad (5)$$

SQUARED EUCLIDEAN DISTANCE:

$$d = \sum_{i=1}^{n} (x_i - y_i)^2 \,. \qquad (6)$$

COSINE DISTANCE (non Euclidean distances for interval data): Cosine of vectors of variables. This is a pattern similarity measure. The cosine of the angle between 2 vectors is identical to their correlation coefficient [3].

$$Similarity(x, y) = \frac{\sum(x, y)}{(\sum x^2)(\sum y^2)} \,. \qquad (7)$$

PEARSON CORRELATION: is a measure of the correlation of two variables *X* and *Y* measured on the same object. It is a measure of the tendency of the variables to increase or decrease together. The result obtained is equivalent to dividing the covariance between the two variables by the product of their standard deviations [7]:

$$r = \frac{\sum z_x z_y}{n - 1} \qquad (8)$$

where:

z_x, z_y – standard scores of the two measures

(n-1) – number of degrees of freedom

The correlation coefficient adds a sign to show the direction of the relationship. The formula for the Pearson coefficient conforms to this definition, and applies when the relationship is linear [7].

CHEBYCHEV DISTANCE is a distance measure. In this formula absolute maximum difference between variable score is used.

$$d = \max_i \left| x_i - y_i \right| \qquad (9)$$

BLOCK DISTANCE (*Manhattan distance*): A new metric in which the distance between two points is the sum of the (absolute) differences of their coordinates. Manhattan distance is named so because it is the shortest distance a car would drive in a city laid out in square blocks. A route which follows the regular grid of roads is used [9].

$$d(x, y) = \sum \left| x - y \right| \qquad (10)$$

In order to select the most appropriate case and to obtain the best results we suggest a new approach in solving this issue which is discussed in detail further in our work.

Since we deal with six variables that have different nature and scaling, it is important that each metric contribute equally to the total distance. To remove this dependency on the range spanned by each metric, it is important to first *standardize* the values. This means that each metric, when compared over the full set of objects will have a mean of 0.0 and a variance (or standard deviation) of 1.0

2 Clustering procedure

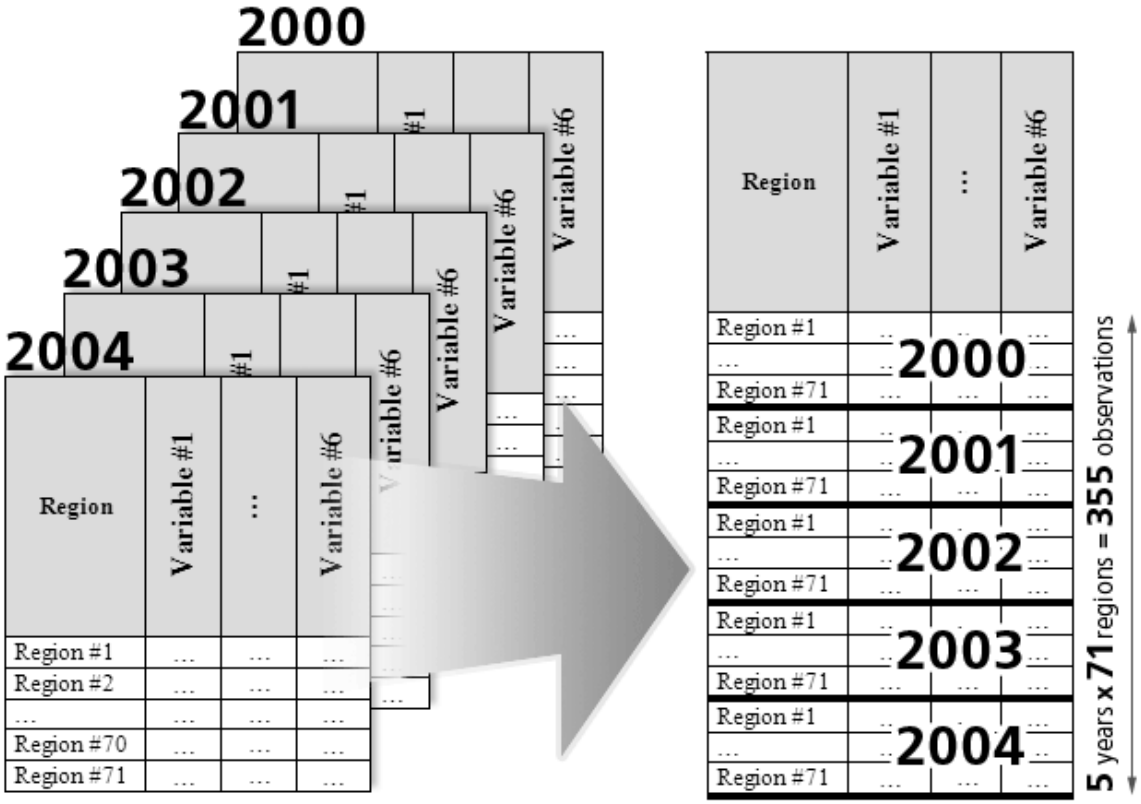
Classifying procedure of 71 regions due to economic development for a period of 2000-2004 is defined as follows and in fact is similar to that one described in [2].

Stage 1. Classification of regions in a multidimensional space according to their indexes using all seven methods for hierarchal clustering included in SPSS and using six different measures. Thus we examine 42 different ways of classification and 42 different results they lead to.

It is essential to add that the process of clustering in our case implies classification of objects not for every given year in particular but making a data pool for all five years of observation in order to get a proper understanding of processes and tendencies taking place in these regions and to track down the motion of these objects within clusters from year to year due to changes in their social and economic activities. The procedure of

creation of such a pool is illustrated on Picture 1. Thus, instead of dealing with five separate annual cluster classifications we combine the data into a single table. As we observe data for 71 Belarusian provinces and six variables for each of them then the resulting data table has a dimension of 355 rows (5 years for 71 regions) and six columns. Taxonomy of regions for every particular year is rather an additional tool for verifying the results of an obtained classification in case we consider any divisive issues of that classification.

Picture 1: Procedure of creation of a five year data pool from annual data.



To define an exact number of clusters needed to classify objects into groups we use a graphical analysis of distances (measured in percents of a maximal distance between grouped clusters in each of 42 classifications) growth for clusters being grouped. As a rule, the process of cluster merging is reasonable until the distance between clusters being grouped does not exceed 5-10 percent of its maximal value (which is attained at the last step of a hierarchical clustering procedure) but a final decision depends on individual tasks of a researcher ([2]). Such an analysis helps define an average suitable number of clusters that we can use to obtain our further classifications.

To formally define the best clustering method and measure we use the following approach. A classification is considered to be the best if this classification provides the largest dissimilarity **between** clusters while producing the highest similarity **within** clusters' objects. It means that we expect to obtain clusters with low within-cluster variance and high between-cluster distance which can be expressed in the following way (see Appendix A for more detail)

$$CDE = \sum_{k=1}^{m} \sum_{i=1}^{n} \left| \overline{x_i^k} - \overline{x}^k \right| - \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{l=1}^{p_i} \left| x_{il}^k - \overline{x_i^k} \right| \tag{11}$$

where *CDE* – Cluster Dissimilarity Estimate,

m – number of variables describing each object,

n – number of clusters in a particular classification,

p_i – number of objects in the *i*-th cluster,

$\overline{x_i^k}$ – mean values of the *i*-th cluster upon the *k*-th variable,

$\overline{x}^k$ – overall mean value upon *k*-th variable,

$x_{i\,l}^{\ k}$ – values of the *l*-th object in the *i*-th cluster upon the *k*-th variable.

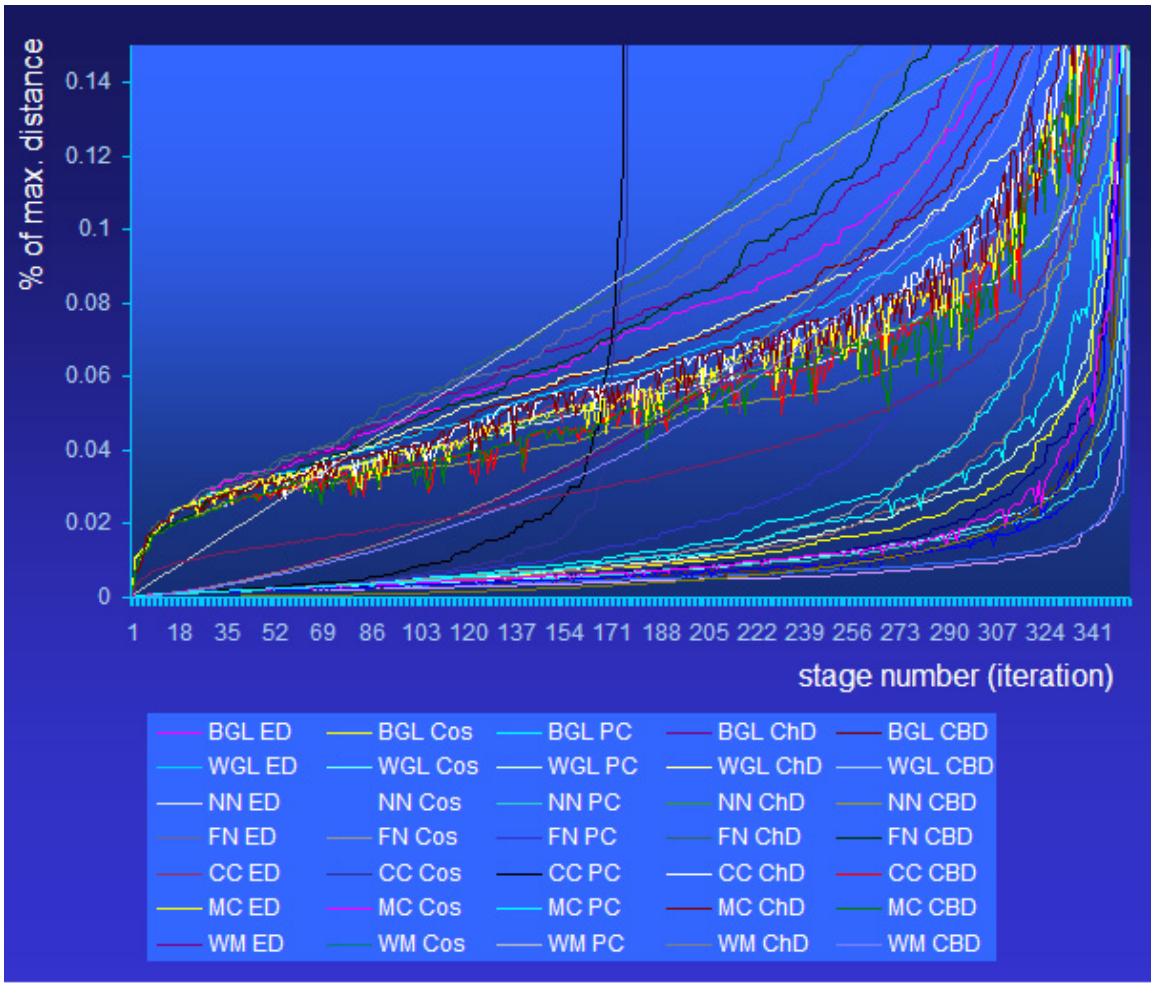
Stage 2. Once a desired classification is verified, we use its results to find out if some of the clusters have similar characteristics. Having a relatively large dataset and keeping in mind that we chose some 'average' number of clusters on a previous step (not the best one for this particular method) it is likely that some clusters have similar features. So, it is possible to minimize number of clusters and thus to simplify further analysis. For instance, it can happen that clusters number five, eight and seventeen contain objects that have far superior values over all the variables comparing to other clusters, while clusters two and ten have distinctively low indexes. Thus, we can combine clusters five, eight and seventeen into one cluster group, as well as clusters two and ten into another. Comparative and ranking tables can be used as additional tools for a better visualization of this procedure.

Stage 3. Having completed the process of cluster reduction and having a comparatively small number of cluster groups, we are ready to analyze dynamic changes in objects' state for a period of five years. Analyzing motion objects' motion from year to year, we try to find similar tendencies in their behavior for the whole period of observations and thus to define several classes of objects with distinctive tendencies.

3 Estimated results

We start with, we must determine a number of clusters which will be used in our further calculations. To obtain this, a hierarchical clustering procedure is applied to classify 355 cases using 42 method-measure combinations of our interest. Consequently, we get 42 sets of changes in distance growth between merged clusters received on each of the 354 steps (iterations) of clustering (agglomeration tables). Since all those distances greatly vary from one classification to another, we need to transform each set to a percentage scale taking a distance at the last 354th step as 100%. Keeping in mind restrictions and criteria concerning this procedure which we stated earlier we graphically illustrate the changes in distance growth between merged clusters (in percents of its maximal distance) for all 42 classifications.

Diagram 1: Distance growth between clusters being grouped (% of max. values).



where:

BGL – Between-groups linkage
WGL – Within-groups linkage
NN – Nearest neighbor
FN – Furthest neighbor
CC – Centriod clustering
MC – Median clustering
WM – Ward's clustering

ED – Euclidean distance
Cos – Cosine distance
PC – Pearson correlation
ChD – Chebychev distance
CBD – City block distance

The next step will be to find out how average distance altered on each step and to determine the step where its value exceeds a 10% barrier. We must admit that due to a heavy dispersal in distances it was not easy to determine an appropriate quantity of clusters that would "satisfy" most clustering algorithms. Nevertheless, after having analyzed the changes in the distances' growth in each of the 42 different classifications (both graphically and analytically) we come to a conclusion that a number of clusters equal to 18 satisfies the criteria mentioned above (in our case average distance between grouped clusters does exceed 11%). However, we should keep in mind that this cluster quantity is idealized in a certain sense and hence is not wholly suitable for some of the methods. That is, in case we needed to find a number of clusters for each method in particular there is no guarantee that this number would be equal 18. We can even assume their values will unlikely coincide. But since we want to put all 42 clustering algorithms into equal conditions we must follow theoretically accepted procedure (though modified) in order to check the quality of classifications produced by different methods and measures.

Further we obtain results of 42 classifications each of which includes 18 clusters. Appendix B contains results of final classifications and also summarized information regarding the number of objects in each cluster of those classifications. After that, following the procedure described in Appendix A of this paper, we are now ready to formally determine a best classification. Table 3 contains final calculation to evaluate the quality of 42 different classifications as combinations of seven clustering methods and six measures available in SPSS and R for hierarchical cluster analysis.

Table 3: Resulting table for defining a formally best classification.

Classification	Nominal values		Standardized values		CDE nominal	CDE adjusted (final)
	Sum of inter-clusters distances (IINTER)	Sum of intra-clusters distances (INTRA)	Sum of inter-clusters distances (INTER)	Sum of intra-clusters distances (INTRA)		
BGL SED	54.6	1396.6	0.1	-0.9	-1342.0	1.0
BGL ED	58.2	1450.1	0.4	0.4	-1391.9	0.0
BGL Cos	50.8	1403.2	-0.2	-0.8	-1352.4	0.6
BGL PC	45.9	1419.1	-0.6	-0.4	-1373.2	-0.2
BGL ChD	49.9	1432.1	-0.3	-0.1	-1382.2	-0.2
BGL CBD	54.6	1418.3	0.1	-0.4	-1363.7	0.5
WGL SED	49.4	1412.1	-0.3	-0.6	-1362.6	0.2
WGL ED	54.3	1389.2	0.1	-1.1	-1334.8	1.2
WGL Cos	46.3	1389.1	-0.6	-1.1	-1342.8	0.6
WGL PC	41.9	1403.1	-0.9	-0.8	-1361.2	-0.1
WGL ChD	55.1	1393.9	0.1	-1.0	-1338.8	1.1
WGL CBD	48.5	1400.8	-0.4	-0.8	-1352.3	0.4
NN SED	74.8	1490.9	1.7	1.3	-1416.1	0.3
NN ED	74.8	1490.9	1.7	1.3	-1416.1	0.3
NN Cos	65.6	1503.6	0.9	1.6	-1438.0	-0.7
NN PC	70.9	1495.7	1.4	1.5	-1424.8	-0.1
NN ChD	69.4	1495.7	1.2	1.5	-1426.3	-0.2
NN CBD	72.8	1492.6	1.5	1.4	-1419.7	0.1
FN SED	47.1	1417.1	-0.5	-0.4	-1370.0	-0.1
FN ED	47.1	1417.1	-0.5	-0.4	-1370.0	-0.1
FN Cos	41.4	1384.9	-1.0	-1.2	-1343.6	0.3
FN PC	40.1	1396.1	-1.1	-0.9	-1356.0	-0.1
FN ChD	41.8	1439.6	-0.9	0.1	-1397.8	-1.0
FN CBD	44.1	1406.5	-0.7	-0.7	-1362.4	0.0
CC SED	66.0	1486.1	1.0	1.2	-1420.1	-0.2
CC ED	71.6	1494.9	1.4	1.4	-1423.3	0.0
CC Cos	37.0	1407.1	-1.3	-0.7	-1370.1	-0.6
CC PC	37.2	1397.6	-1.3	-0.9	-1360.4	-0.4
CC ChD	74.0	1493.6	1.6	1.4	-1419.6	0.2
CC CBD	67.1	1494.3	1.1	1.4	-1427.2	-0.4
MC SED	64.3	1446.6	0.8	0.3	-1382.4	0.6
MC ED	59.9	1450.8	0.5	0.4	-1390.9	0.1
MC Cos	53.8	1453.0	0.0	0.4	-1399.2	-0.4
MC PC	47.7	1436.1	-0.5	0.0	-1388.4	-0.5
MC ChD	69.2	1496.7	1.2	1.5	-1427.6	-0.3
MC CBD	70.7	1494.4	1.3	1.4	-1423.7	-0.1
WM SED	39.0	1407.8	-1.1	-0.7	-1368.9	-0.5
WM ED	38.4	1399.7	-1.2	-0.9	-1361.3	-0.3
WM Cos	38.6	1404.7	-1.2	-0.7	-1366.2	-0.4
WM PC	33.9	1407.7	-1.5	-0.7	-1373.8	-0.9
WM ChD	38.3	1403.7	-1.2	-0.8	-1365.3	-0.4
WM CBD	42.6	1374.1	-0.9	-1.5	-1331.5	0.6

As it is shown in Appendix A of this paper, the above table contains two types of estimation. Following the formula (11), we calculate nominal values for both inter- and intra-distances as well as the corresponding CDE values. In addition to this, further we standardize the inter- and intra-cluster dissimilarity distances separately to equalize influence of those values on the final CDE values. Hence, standardized values of the resulting between- and within cluster distances provide adjusted CDE estimates which give us information regarding quality of the observed 42 classifications.

Thus, formally speaking we can conclude that several classifications resulted in higher CDE values in comparison with other estimates. Among these classifications we can point out the following: Within-Groups Linkage with Euclidean Measure, Within-Groups Linkage with Chebychev Distance and Between-Groups Linkage with Squared Euclidean Distance (with CDE values respectively 1.2, 1.1 and 1.0). The lowest CDEs belong to classifications built according to Furthest Neighbor with Chebychev Distance, Ward's Clustering with Pearson Correlation and Nearest Neighbor with Pearson Correlation (-1.0, -0.9 and -0.7).

We can also state that in this particular task the two clustering methods (Within-Groups Linkage and Between-Groups Linkage) showed relatively better results comparing with other algorithms. It is also interesting that Ward's Clustering, one of the most well known and widely used methods in cluster analysis, in this application demonstrated quite poor results. As we can see from Table 3 Ward's method showed good potential in generating homogenous clusters (what directly follows from its definition) but at the same time could not provide sufficient dissimilarity among those clusters and thus resulted in low overall CDE estimates.

Speaking about metrics, we have to admit that Euclidean and Squared Euclidean Distance provided classifications with comparatively high between- and low within-cluster dissimilarities distances, that is, in our task these two metrics appear to be more preferable than the others. On the other hand, Pearson Correlation in all seven methods of our analysis led to the lowest CDE final values and thus benefits of its application in this paper is rather dubious.

It is also important to notice that during the process of calculations using both R and SPSS software some differences between these programs where revealed concerning classification results and some of the techniques used within them[2].

4 Conclusions

Estimated results in a whole demonstrate that some algorithms for evaluating the quality of taxonomies built with different clustering methods and metrics can be applied to identify formally best classifications. A new approach suggested in this paper for evaluating procedure is based on the well known and widely accepted theoretical assumption that a high quality taxonomy will produce homogenous and highly distinctive clusters. Its simplicity and logicality make this algorithm useful while simultaneously dealing with various classifications and a task for finding an acceptable classification arises. It can be helpful in different areas of cluster analysis's application as in most cases a researcher makes choices in favor of this or that method and measure in his particular case at the beginning of his research and thus certain mistakes or contretemps may occur throughout his analysis or when drawing final conclusions. Our algorithm allows to conduct clustering process in several ways and to make judgment of accuracy and applicability of any clustering method and measure in one's analysis.

Concerning data clustering itself we can say that cluster analysis like any other method has its own cons and restrictions. In particular, a structure and a number of clusters are determined by the chosen criteria of classification. That is, the results of any partition vary greatly from one method to another, from a selected measure and a chosen number

[2] When conducting a cluster analysis in R or SPSS, there are two points which need to pay attention to: 1) a different way of standardizing data in SPSS and R; 2) R and SPSS have different way to deal with the missing data.

SPSS standardization method (Z scores) is defined as subtracting variable's mean from each value and dividing the result by the variable's standard deviation**.** In R (function *"Agnes"*) standardization procedure implies subtracting the variable's mean and dividing the expression by the variable's mean absolute deviation.

As to the missing data, in a case there is missing data in SPSS, this case will be omitted and will not be included into any cluster. If there are any missing values in R (noted as "NA") the "dist" function can still compute the distance matrix in (translating the "NA" into numeric values), so we can still get an integrated distance matrix and obtain a classification with all the cases, including those with missing values.

of clusters as well. In the process of numerous partitions and aggregations of data one may lose distinctive features of some of the objects due to substituting its characteristics with the aggregate cluster values. It is also important to notice that a clustering procedure is very sensitive to the missing values of variables and that adding or changing any pieces of data involves this procedure to be held again from the very beginning. Moreover, missing values are interpreted differently in SPSS and R packages what lead to greatly varying results so one should be very careful when deciding which program to use to classify objects. But in general clusterization allows us to break a general sample of objects into unique groups and to reveal their distinctive features. It is obvious that such a procedure should always be taken over researcher's control and its results should always be checked with an additional quantitive and qualitative analysis.

References

1. A Tutorial on Clustering Algorithms. (http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/)

2. B. Boots, S. Drobyshevsky, O. Kochetkova, G. Malginov, V. Petrov, G. Fedorov, Al. Hecht, A. Shekhovtsov, A. Yudin. Typology of Russian Regions. *The Institute for the Economy in Transition*, 2002.

3. Cluster Analysis, *Manchester Metropolitan University*, (http://149.170.199.144/multivar/ca.htm).

4. Cluster Analysis, *Statistics Solution Inc*., (http://www.statisticssolutions.com/Cluster-Analysis.htm).

5. J. Ward, Jr. Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Volume 58, Issue 301, 1963, pp. 236-244.

6. N. Sambamoorthi. Hierarchical Cluster Analysis, *CRMportals Inc*., (http://www.crmportals.com/hierarchical_cluster_analysis.pdf).

7. Pearson product-moment correlation coefficient, *Wikipedia*, (http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient).

8. Regions of Republic of Belarus. *The Ministry of Statistics and Analysis of the Republic of Belarus*, 2005 (published in Russian).

9. Taxicab geometry, *Wikipedia*, (http://en.wikipedia.org/wiki/Taxicab_geometry).

Additional Reading

10. Borland, J, Hirschberg, J, Lye, J. Data Reduction of Discrete Responses: An Application of Cluster Analysis. *Department of Economics - Working Papers Series*, 1998.

11. D. Babicki, U. Valetka, A. Ivanouski, S. Salas. Barysau and Region : the Development Strategy. *Civil initiative "Clean Barysau"*, 2004 (published in Belarusian).

12. Economy and Society of Belarus: Disproportions and Development Perspectives. National Report of Human Development 2004-2005. *UNDP Report*, 2005 (published in Russian).

13. Investment Ranking of the Russian Regions 1999-2000. *Expert*, 2000, #41 (published in Russian).

14. K. Franz. A Study of the Competitiveness of Regions based on a Cluster Analysis: The Example of East Germany. *ERSA conference papers*, 2003.

15. M. Chernysh. Experience of Using Cluster Analysis. *Sociology: 4M*, #12, 2000 (published in Russian).

16. M. Steiner, V. Boikov, B. Frolov, F. Hiss. Regional Differentiation in the Russian Federation - a Cluster-based Typification. *ERSA conference papers*, 1998.

17. M. Valevski, A. Chubrik. Households in Belarus in 1995-2000 // Belarusian Economy: from Market to Plan. Volume 1. *Center for Social and Economic Research,* 2002 (published in Russian).

18. S. Aivazian. Integral Indicators of the Living Standards: Building and Using in Management and Interregional Comparison. *Central Economics and Mathematics Institute of Russian Academy of Sciences*, 2000 (published in Russian).

19. Y. Kozhich. Grounds of the Administrative and Territorial Division of the Republic of Belarus. [Thesis], 2005 (published in Russian).

Appendix A

Our particular task is to find a best classification for 71 Belarus regions within the period of five years. Observing 71 regions during the period of five years will give us 355 cases. Using seven methods for hierarchical clustering included in SPSS software (Between-Group Linkage, Within-Group Linkage, Nearest Neighbor, Further Neighbor, Centroid Custering, Median Clustering, Ward's Clustering) and six different measures (Euclidean Distance, Cosine, Pearson Correlation, Chebychev Distance, City Block Distance), we examine 42 different ways of classification.

To compare results of classifications and thus to formally trace the best classification in the previous work we used the following approach suggested in [4]:

$$H = -\sum_{i=1}^{n} \frac{N_i}{N} \log_2 \frac{N_i}{N} = \log_2 N - \frac{1}{N} \sum_{i=1}^{n} N_i \log_2 N_i \qquad (12)$$

The main idea was to get classification which is considered to be the best if all of its objects are divided into groups more or less evenly. That means that all or most clusters should be filled with objects.

We suggest a new approach in determining the quality of a classification. This new criterion of finding a best classification is based on the theoretical assumptions that a good clustering will produce high quality clusters if:

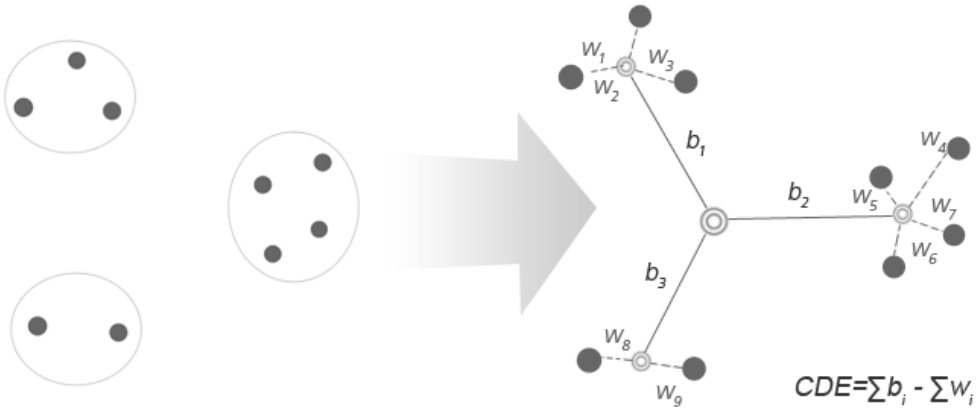
- the inter-cluster similarity is low, that is all the found clusters differ from each other as much as possible.

- the intra-cluster class similarity is high, which means that within-cluster objects' likeness is encouraged to be maximized.

Thus, a good classification is that one which provides a large dissimilarity among clusters while producing the highest objects' similarity within clusters. Using this and other information from previous stages of our analysis (such as an average number of clusters suitable for 42 different classifications), we define the inter-cluster dissimilarity as the sum of absolute differences between a mean value of each clusters and the overall mean, and intra-cluster similarity as the sum of absolute distances between each object in a cluster and cluster's mean (Picture 2). We also define the Cluster Dissimilarity Estimate (CDE) as the difference between inter- and intra-cluster dissimilarities values.

Since the good classification requires maximal distance between clusters and minimal variance within clusters we need our Cluster Dissimilarity Estimate to be maximized in order to get a formally best classification.

Picture 2: Graphical illustration of CDE values calculations (in case of 3 clusters).



where

b_1-b_3 – between-cluster dissimilarity distances,

w_1-w_9– within-cluster similarity distances.

In general we will define Cluster Dissimilarity Estimate in the following way:

$$CDE = \sum_{k=1}^{m} \sum_{i=1}^{n} \left| \overline{x_i^k} - \overline{x}^k \right| - \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{l=1}^{p_i} \left| x_{il}^k - \overline{x_i^k} \right|$$

where *CDE* – Cluster Dissimilarity Estimate,

m – number of variables describing each object,

n – number of clusters in a particular classification,

p_i – number of objects in the *i*-th cluster,

$\overline{x_i^k}$ – mean values of the *i*-th cluster upon the *k*-th variable,

$\overline{x}^k$ – overall mean value upon *k*-th variable,

x_{il}^k – values of the *l*-th object in the *i*-th cluster upon the *k*-th variable.

The best classification will correspond to the maximum of the CDE value.

It is important to notice that due to different nature of the variables we need to deal with their standardized values rather then with their absolute values. Also, in this particular work all of the six indexes are considered of the same importance though it can be quite easily changed using some adjustment coefficients.

To illustrate the main idea and some of the results of our new approach consider the following example. Suppose, we have 9 observations in two-dimensional space, that is, every observation is defined by two variables, x_i and y_i. We want to compute the CDE values given the following information:

- the methods and measures we want to compare are (chosen arbitrary): Between-Groups Linkage with Euclidean Distance (BGL ED), Between-Groups Linkage with Cosine Distance (BGL Cos), Within-Groups Linkage with Cosine Distance (WGL Cos), Within-Groups Linkage with Pearson Correlation (WGL PC) and Ward's Method with Pearson Correlation (WM PC);

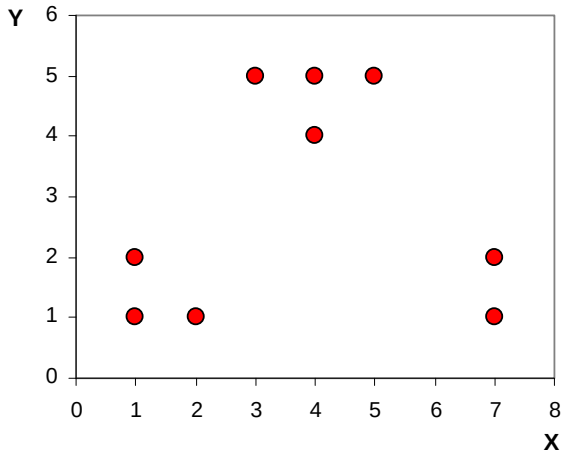
- the number of clusters varies from two to six (in order to see how CDE values change if given different number of clusters).

The data and the graphical illustration for this example are given below.

Table 4: The input data and standardized values for CDE calculations example.

Observation Number	X values	Y values	X values (standardized)	Y values (standardized)
Observation 1	1	1	-1.22	-1.03
Observation 2	1	2	-1.22	-0.48
Observation 3	2	1	-0.78	-1.03
Observation 4	3	5	-0.34	1.15
Observation 5	4	4	0.10	0.61
Observation 6	4	5	0.10	1.15
Observation 7	5	5	0.54	1.15
Observation 8	7	1	1.41	-1.03
Observation 9	7	2	1.41	-0.48

Picture 3: Graphical illustration of the input data (absolute values).



As it is stated above we want to see how CDE values change if different number of clusters is taken under our consideration. As an example, the results of the partition into three and six clusters for all five methods are given below.

Table 5: Results of classification of nine objects into three and six clusters for five methods.

Three clusters					Six clusters				
BGL ED	**BGL Cos**	**WGL Cos**	**WGL PC**	**WM PC**	**BGL ED**	**BGL Cos**	**WGL Cos**	**WGL PC**	**WM PC**
1	1	1	1	1	1	1	1	1	1
1	2	1	2	2	2	2	2	2	2
1	3	2	3	3	1	1	1	3	3
2	2	1	2	2	3	3	3	4	4
2	1	1	2	1	4	4	4	5	5
2	1	1	2	2	5	4	4	6	6
2	1	1	3	1	5	5	4	6	6
3	3	3	3	3	6	6	5	3	3
3	3	3	3	3	6	6	6	3	3

After having calculated the inter- and intra-cluster similarities according to (11), we compute the CDE values for all five methods and for different partitions (from 2 to 6 clusters). The following table contains the summarized information on CDE values and some intermediate calculations.

Table 6: Results for CDE values for five methods and five partitions (2-6 clusters).

Number of clusters	Method used	Nominal values		Standardized values		CDE nominal	CDE adjusted
		Sum of inter-clusters distances	Sum of intra-clusters distances	Sum of inter-clusters distances	Sum of intra-clusters distances		
2 clusters	**BGL ED**	2.9	9.5	0.7	-0.7	-6.6	1.4
	BGL Cos	2.9	9.5	0.7	-0.7	-6.6	1.4
	WGL Cos	2.9	9.5	0.7	-0.7	-6.6	1.4
	WGL PC	2.3	11.9	-1.1	1.1	-9.6	-2.2
	WM PC	2.3	11.9	-1.1	1.1	-9.6	-2.2
3 clusters	**BGL ED**	5.2	3.6	1.2	-1.7	1.6	2.9
	BGL Cos	3.2	11.4	-0.9	0.7	-8.2	-1.6
	WGL Cos	4.7	8.8	0.7	-0.1	-4.1	0.8
	WGL PC	4.2	10	0.1	0.3	-5.8	-0.2
	WM PC	3.1	11.9	-1.1	0.8	-8.8	-1.9
4 clusters	**BGL ED**	6.8	3.2	0.7	-0.7	3.6	1.4
	BGL Cos	6.8	3.2	0.7	-0.7	3.6	1.4
	WGL Cos	6.8	3.2	0.7	-0.7	3.6	1.4
	WGL PC	4.8	9.1	-1.1	1.1	-4.3	-2.2
	WM PC	4.8	9.1	-1.1	1.1	-4.3	-2.2
5 clusters	**BGL ED**	8.6	2.3	0.7	-0.7	6.3	1.4
	BGL Cos	8.6	2.3	0.7	-0.7	6.3	1.4
	WGL Cos	8.6	2.3	0.7	-0.7	6.3	1.4
	WGL PC	6.1	8.6	-1.1	1.1	-2.5	-2.2
	WM PC	6.1	8.6	-1.1	1.1	-2.5	-2.2
6 clusters	**BGL ED**	9.6	1.4	-0.2	-0.8	8.2	0.6
	BGL Cos	10.1	1.5	0.5	-0.8	8.6	1.3
	WGL Cos	10.8	1.8	1.5	-0.6	9	2.1
	WGL PC	9.1	4.1	-0.9	1.1	5	-2
	WM PC	9.1	4.1	-0.9	1.1	5	-2

The above table illustrates (results under “Nominal values”) a significant variation of the final CDE values from one classification to another due to the changes in our inter- and intra-cluster distances which are caused, first of all, by the selected number of clusters in classifications. To eliminate this problem, we standardize the inter- and intra-cluster distances separately in order to achieve approximately equal influence of those values on the final CDE estimates. Therefore, Table 6 also contains standardized values of the resulting between- and within cluster distances, as well as adjusted CDE estimates calculated from them. Standardization is done separately for every of the partitions (from 2 to 6 clusters) presented in this example as our aim does not implies testing which of these partitions delivers a more “reasonable” taxonomy but rather answering the question which of the methods yields better results in dividing objects into a definite number of groups.

Thus, it will be incorrect to compare resulting CDE values for classifications involving different numbers of clusters. Conclusions which classification is formally considered better should be drawn only within analogical partition, that is, we can compare CDE values only for classifications that are built using the same number of clusters. Otherwise, the obtained values can cause confusion.