Cluster Analysis

Anuška Ferligoj

University of Ljubljana, Slovenia
NRU HSE, Moscow, Russia

Outline

1. Introduction
2. Clustering
3. Dissimilarities
4. Criterion function
5. Clustering problem
6. Relocation algorithm
7. Hierarchical algorithm
8. Monotonicity
9. Leader algorithm
10. Examples

Introduction

Grouping units into clusters so that those within a cluster are as similar to each other as possible, while units in different clusters as dissimilar as possible, is a very old problem. Although the clustering problem is intuitively simple and understandable, providing solution(s) remains a very current activity. The increasing number of recent papers on this topic, in both theoretical and applied statistical journals, is notable.

Further, the *Journal of Classification*, was established in 1984 and the *International Federation of Classification Societies* was formed in 1985. In 2007 a new journal appeared *Advances in Data Analysis and Classification*.

There are two main reasons for this lively interest:

- Prior to 1960, clustering problems were solved separately in different scientific fields with little concern for integrating specific solutions. Attempts to unify different problems and solutions first appeared in the sixties with *Sokal and Sneath* (1963) providing the first extensive statement. With this as a point of departure, the field of cluster analysis developed as a specific field within data analysis.
- The development of cluster analysis was influenced greatly by developments in *computing technology*. Computers allowed the application of more demanding computational procedures and the processing of large data sets. Theoretical results in computer science on *computational complexity* were important also.

Basic notions

Let us start with the formal setting of the clustering problem. The following notations will be used:

X – *unit*

$\mathcal{U}$ – finite *set of units*

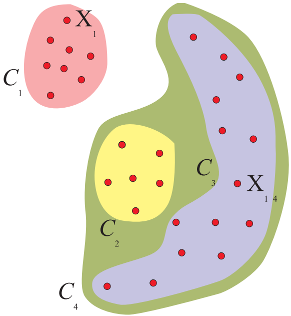
C – *cluster*, $\emptyset \subset C \subseteq \mathcal{U}$

$\mathbf{C}$ – *clustering*, $\mathbf{C} = \{C_i\}$

Φ – set of *feasible clusterings*

P – *criterion function*,

$$P : \Phi \to \mathbb{R}_0^+$$



Clustering

There are several types of clusterings, e.g., partition, hierarchy, pyramid, fuzzy clustering, clustering with overlapping clusters. The most frequently used clusterings are partition and hierarchy. A clustering $\mathbf{C} = \{C_1, C_2, \dots C_k\}$ is a *partition* of the set of units $\mathcal{U}$ if

$$\bigcup_i C_i = \mathcal{U}$$

$$i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

A clustering $\mathbf{H} = \{C_1, C_2, \dots C_k\}$ is a *hierarchy* if for each pair of clusters C_i and C_j from $\mathbf{H}$ it holds

$$C_i \cap C_j \in \{C_i, C_j, \emptyset\}$$

and it is a complete hierarchy if for each unit X it holds $\{X\} \in \mathbf{H}$, and $\mathcal{U} \in \mathbf{H}$.

Dissimilarity

For solving a clustering problem, the choice of an appropriate dissimilarty measure between two units is crucial. The issues to consider when selecting a (dis)similarity measure include its mathematical properties, its behavior when confronted with data, and the nature of the data.

A dissimilarity can be described by a mapping, *a measure of dissimilarity*, where a real number is assigned to each pair of units (x, y)

$$d : (x, y) \mapsto R$$

We usually assume the following conditions hold:

1. $d(x, y) \geq 0$ nonnegativity
2. $d(x, x) = 0$
3. $d(x, y) = d(y, x)$ symmetry

If, for a dissimilarity measure, the following two conditions also hold,

4. $d(x, y) = 0 \Longrightarrow x = y$
5. $\forall z : d(x, y) \leq d(x, z) + d(z, y)$ triangle inequality

the dissimilarity is called *distance*.

(Dis)similarity measures for numerical data

For the units X and Y decribed by m numerical variables

$$X = (x_1, x_2, \dots, x_m)$$

$$Y = (y_1, y_2, \dots, y_m)$$

the *euclidean distance* is defined in the following way:

$$d(X, Y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Often the *Manhattan distance* is used:

$$d(X, Y) = \sum_{i=1}^{m} |x_i - y_i|$$

Minkowsky distance

Both distances are special cases of the *Minkowsky distance*

$$d(X, Y) = (\sum_{i=1}^{m} |x_i - y_i|^r)^{\frac{1}{r}} \quad , \quad r > 0$$

For $r = 2$ it is euclidean distance, and for $r = 1$ it is Manhattan distance.

Minkowsky distance has the following property: for large values of r the larger distances between two units $|x_i - y_i|$ have larger impact on the distance. In the case of $r = \infty$, the Minkowsky distance is

$$d(X, Y) = \max_i |x_i - y_i|$$

It is called *Chebyshev distance*.

Correlation coefficient

Also *Pearson correlation coefficient* can be used for numerical data

$$r(X,Y)=\frac{\sum_{i=1}^{m}(x_i-\mu_X)(y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{m}(x_i-\mu_X)^2\sum_{i=1}^{m}(y_i-\mu_Y)^2}}$$

A property of the correlation coefficient is that it remains the same if one or the other unit is linearly transformed.
Let us assume that we add a constant to all values of a unit. The profiles of both units are the same. If the euclidean distance is calculated it can be rather large if that constant is large. But the correlation coefficient would be equal to 1.
Therefore, it is important to understand the research problem and to choose an appropriate (dis)similarity measure.

Standardization

If the variables are measured on different scales, we standardize numerical variables before calculating the distances between units. The most usual standardization is

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where x_{ij} is the value of the variable X_j for the unit i, μ_j is the arithmetic mean and σ_j is the standard deviation of the variable X_j.

Other standardizations

$$z_{ij} = \frac{x_{ij}}{\sigma_j}$$

$$z_{ij} = \frac{x_{ij}}{\max X_j}$$

$$z_{ij} = \frac{x_{ij}}{\mu_j}$$

$$z_{ij} = \frac{x_{ij}}{\max X_j - \min X_j}$$

$$z_{ij} = \frac{x_{ij} - \min X_j}{\max X_j - \min X_j}$$

Milligan in Cooper (1988) proposed several possible standardizations and compared them.

Dissimilarity measures for binary data

Many similarity measures have been defined for units described by binary variables. They are determined mostly by the frequencies of the contingency table for a pair of units for which the similarity is measured. The contingency table for the units X and Y where the values of all m variables are $+$ and $-$ is:

		Unit Y	
		$+$	$-$
Unit X	$+$	a	b
	$-$	c	d

The sum of all four frequences is equal to the number of variables ($a + b + c + d = m$).

Some matching similarity measures

1. Sokal-Michener similarity (1958)

$$\frac{a+d}{a+b+c+d}$$

2. Russell-Rao similarity (1940)

$$\frac{a}{a+b+c+d}$$

3. Jaccard similarity (1908)

$$\frac{a}{a+b+c}$$

All of these similarity measures are defined in the interval from 0 to 1. See Batagelj, Bren (1995): Comparing resemblance measures. J Classif 12 (1): 73-90.

Relationship between similarities and dissimilarities

The clustering algorithms in most of the cases assume dissimilarity measures. A similarity measure s can be tranformed into a dissimilarity measure d and vice versa. There are several transformations possible. The choice depends also on the interval of the measure that has to be transformed. If a similarity measure s is defined on the interval $[0, 1]$, the usual transformation to a dissimilarity mesure d is

$$d = 1 - s$$

There are several choices if the similarity measure is defined on the interval $[-1, 1]$.

Criterion function

Clustering criterion functions can be constructed *indirectly* as a function of a suitable (dis)similarity measure between pairs of units (e.g., Euclidean distance) or *directly*. In most cases the criterion function is defined indirectly.

In the case of partitions into k clusters the *Ward criterion function* is usually used

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{X \in C} d(X, t_C)$$

where t_C is the centroid of the cluster C and d the squared Euclidean distance.

Several other types of criterion functions were proposed in the literature.

Clustering problem

Cluster analysis (also classification, taxonomy) deals mainly with the following *clustering problem*:

> *Determine the clustering* $\mathbf{C}^\star \in \Phi$ *for which*

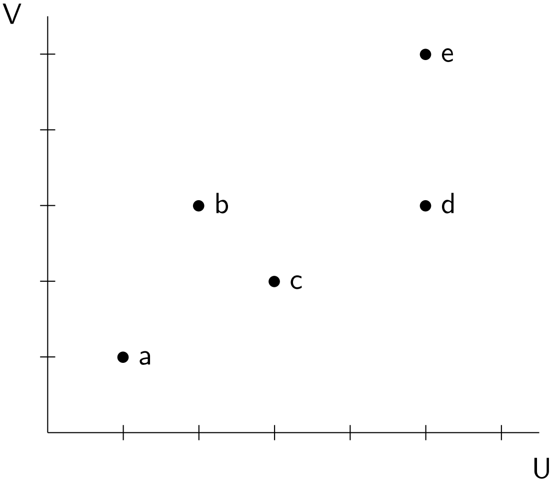
$$P(\mathbf{C}^\star) = \min_{\mathbf{C}\in\Phi} P(\mathbf{C})$$

Since the set of units $\mathcal{U}$ is finite, the set of feasible clusterings is also finite. Therefore, the set $\mathrm{Min}(\Phi, P)$ of all solutions of the problem (optimal clusterings) is not empty. (In theory) the set $\mathrm{Min}(\Phi, P)$ can be determined by the complete search.

An example

Consider the set of five units $U = \{a, b, c, d, e\}$ for which there are measurements in terms of two variables (U and V):

	a	b	c	d	e
U	1	2	3	5	5
V	1	3	2	3	5


V
e
b
d
c
a
U

The units are grouped into two clusters (a partition) using the following criterion function:

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \sum_{X \in C} d(X, t_C)$$

where $t_C = (\overline{u}_C, \overline{v}_C)$ is the center of the cluster C and the dissimilarity d is euclidean distance.
All possible partitions into two clusters, together with the calculated values of the criterion function, are shown in the next slide.

$\mathbf{C}$	C_1	C_2	t_1	t_2	$P(\mathbf{C})$
1	a	$bcde$	$(1.0, 1.0)$	$(3.75, 3.25)$	6.65
2	b	$acde$	$(2.0, 3.0)$	$(3.50, 2.75)$	8.18
3	c	$abde$	$(3.0, 2.0)$	$(3.25, 3.00)$	8.67
4	d	$abce$	$(5.0, 3.0)$	$(2.75, 2.75)$	7.24
5	e	$abcd$	$(5.0, 5.0)$	$(2.75, 2.25)$	5.94
6	ab	cde	$(1.5, 2.0)$	$(4.33, 3.33)$	6.66
7	ac	bde	$(2.0, 1.5)$	$(4.00, 3.67)$	7.21
8	ad	bce	$(3.0, 2.0)$	$(3.33, 3.33)$	9.58
9	ae	bcd	$(3.0, 3.0)$	$(3.33, 2.67)$	9.48
10	bc	ade	$(2.5, 2.5)$	$(3.67, 3.00)$	8.48
11	bd	ace	$(3.5, 3.0)$	$(3.00, 2.67)$	9.34
12	be	acd	$(3.5, 4.0)$	$(3.00, 2.00)$	8.08
13	cd	abe	$(4.0, 2.5)$	$(2.67, 3.00)$	8.58
14	ce	abd	$(4.0, 3.5)$	$(2.67, 2.33)$	9.11
15	de	abc	$(5.0, 4.0)$	$(2.00, 2.00)$	5.41

The lowest value of the criterion function is (for the last partition):

$$P(\mathbf{C}_{15}) = 5.41$$

The best clustering (partition) for this criterion function is therefore

$$\mathbf{C}^* = \{\{a, b, c\}, \{d, e\}\}$$

From the graphical display, this is the obvious solution. For this simple example we can search the set of all 15 possible clusterings. In general, however, if there are n units there are

$$2^{n-1} - 1$$

different partitions with 2 clusters. The number of partitions exponentially increases with the number of units.

Consequences

Although there are some polynomial types of clustering problems it seems that they are mainly NP-hard.
From these results it follows (it is believed) that no efficient (polynomial) exact algorithm exists for solving the clustering problem.
Therefore, the algorithms should be used which give "good" results, but not necessarily the best, in a reasonable time.
The most important types of such algorithms are:

- relocation algorithm
- hierarchical (agglomerative, divisive and adding) algorithm
- leaders, K-MEANS or dynamic clusters algorithm
- graph theory algorithms
- ...

Relocation algorithm

The algorithm assumes that the user can specify the number of clusters of the partition.
The scheme of the relocation algorithm is:

```
Determine the initial clustering $\mathbf{C}$;
while
    there exists $\mathbf{C}'$, obtained by moving a unit $X_i$ from cluster
    $C_p$ to cluster $C_q$ in the clustering $\mathbf{C}$ or by interchanging
    units $X_i$ and $X_j$ between two clusters;
    such that $P(\mathbf{C}') \leq P(\mathbf{C})$
repeat:
    substitute $\mathbf{C}'$ for $\mathbf{C}$ .
```

While different criterion functions can be used in this approach, the Ward criterion function is used most often.
It is a **local** optimization procedure. To obtain a good solution the algorithm has to be repeated several hundreds times.

Hierarchical agglomerative algorithm

Agglomerative hierarchical clustering algorithm assumes that all relevant information on the relationships between the n units from the set $\mathcal{U}$ is summarized by a dissimilarity matrix $D = [d_{ij}]$.

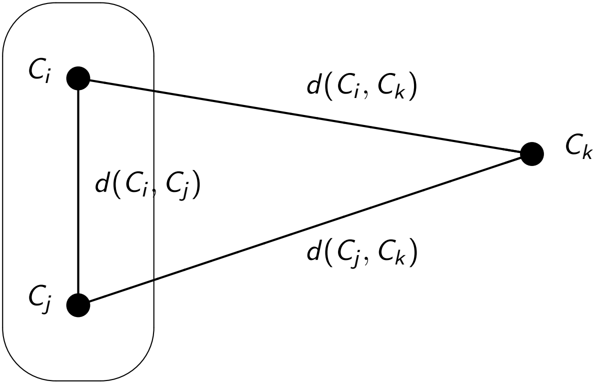
```
Each unit is a cluster: $C_i = \{X_i\}$ , $X_i \in \mathcal{U}$ , $i = 1, 2, \ldots, n$;
**repeat** while there exist at least two clusters:
    determine the nearest pair of clusters $C_p$ and $C_q$:
        $d(C_p, C_q) = \min_{u,v} d(C_u, C_v)$ ;
    fuse the clusters $C_p$ and $C_q$ to form a new cluster
        $C_r = C_p \cup C_q$;
    replace $C_p$ and $C_q$ by the cluster $C_r$;
    determine the dissimilarities between the cluster $C_r$
        and other clusters.
```

The resulting clustering (hierarchy) can be represented graphically by *dendrogram*. The fusion level is:

$$h(C_r) = d(C_p, C_q)$$

$$h(C_i) = 0$$

Dissimilarities between clusters



Methods

- The **Minimum method**, or single linkage, (Florek et al., 1951; Sneath, 1957):

$$d(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

- The **Maximum method**, or complete linkage, (McQuitty, 1960):

$$d(C_i \cup C_j, C_k) = \max(d(C_i, C_k), d(C_j, C_k))$$

- The**McQuitty method** (McQuitty, 1966; 1967):

$$d(C_i \cup C_j, C_k) = \frac{d(C_i, C_k) + d(C_j, C_k)}{2}$$

- The **Ward method** (Ward, 1963):

$$d(C_i \cup C_j, C_k) = \frac{(n_i + n_j)n_k}{(n_i + n_j + n_k)} d^2(t_{ij}, t_k)$$

t_{ij} denotes the centroid of the fused cluster $C_i \cup C_j$ and t_k the centroid of the cluster C_k. n_i denotes the number of units in C_i.

Lance in Williams' formula

Lance and Williams (1967) showed that it is possible to present most of the hierarchical agglomerative methods as special cases of the method in which the new dissimilarity measures can be determined by the following formula

$$d(C_i \cup C_j, C_k) =$$

$$= \alpha_1 d(C_i, C_k) + \alpha_2 d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|$$

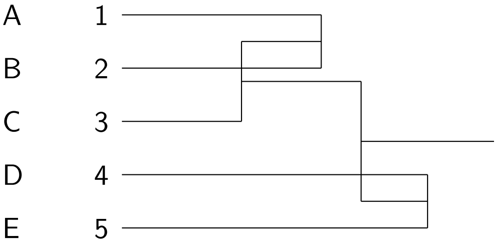
By an appropriate selection of the coefficients α_1, α_2, β and γ in the formula above most of the known methods can be obtained.

method	α_1	α_2	β	γ
minimum	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
maximum	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
McQuitty	$\frac{1}{2}$	$\frac{1}{2}$	0	0
average	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Gower	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{n_j + n_k}{n_i + n_j + n_k}$	$-\frac{n_k}{n_i + n_j + n_k}$	0

Monotonicity

On the basis of Lance and Williams' formula one can generate infinitely many hierarchical methods. A question is whether a certain set of four coefficients provides a method that makes sense. This question can be answered by creating criteria with which we can assess whether a certain method makes sense. One such a criterion is the *monotonicity* of the method. When two clusters C_i in C_j are fused into the new cluster $C_r = C_i \cup C_j$ it can happened that the dissimilarity measure (the fusion level) where the two clusters C_i in C_j are fused is smaller than the fusion levels of clusters C_i and C_j at the previous steps. The method in which such a phenomenon may occure builds 'unnatural' dendrogram with 'inversions'.

An example of nonmonotonic dendrogram



Definition of the monotonicity

Let h be the fusion level in the dendrogram defined in the following way:

$$X \in E \Longrightarrow h(\{X\}) = 0$$

$$C_r = C_i \cup C_j \Longrightarrow h(C_r) = d(C_i, C_j)$$

The dendrogram is monotonic if and only if for each cluster $C_r = C_i \cup C_j$ in the dendrogram holds:

$$h(C_r) \geq \max(h(C_i), h(C_j))$$

Batagelj (1981) proved that a hierarchical method based on Lance and Williams' formula ensures monotonic dendrograms if the following requirements hold:

$$\gamma \geq -\min(\alpha_1, \alpha_2)$$

$$\alpha_1 + \alpha_2 \geq 0$$

$$\alpha_1 + \alpha_2 + \beta \geq 1$$

The first and the second requirement holds for all mentioned methods, but the third one does not hold for Gower's metod.

Characteristics of agglomerative algorithm

- The ‘greediness’ of the agglomerative algorithm can be seen. The early fusion of clusters can preclude the later formation of more optimal clusters: Clusters fused early cannot be separated later even if the early fusion is incorrect. The negative effects of greediness are usually noticed at the higher levels of agglomeration (with smaller numbers of clusters).
- Agglomerative hierarchical algorithm is very popular as it is very simple and its solution can be presented nicely by a dendrogram. In general, it is very quick also for some hundreds of units and users do not need to have an explicit idea about the number of clusters hidden within the data.

Characteristics of some methods

- The minimum method is very effective for finding long, non-elliptic, clusters (with a ‘sausage’ shape). If there are overlapping clusters, the effect of using the minimum method is chaining, where, in each iteration, only one unit is added to a cluster.
- The maximum method searches for very cohesive clusters. The maximum method is best for spherical clusters.
- The Ward method is the most suitable for finding ellipsoidal clusters.

Leader algorithm

Among the *nonhierarchical algorithms*, the most popular is the leader algorithm (Hartigan, 1975), or K-MEANS (e.g., MacQueen, 1967) or the dynamic clusters algorithm (Diday, 1974). It assumes that users can determine the number of clusters of the partition they want to obtain.

The basic scheme of the leader algorithm is:

```
Determine the initial set of leaders $\mathcal{L} = \{l_i\}$;
repeat
    determine the clustering $\mathbf{C}$ in the way to classify
        each unit to the nearest leader;
    for each cluster $C_i \in \mathbf{C}$ compute its centroid $\overline{C_i}$.
        The centroid $\overline{C_i}$ determines the new leader $l_i$
        of the cluster $C_i$;
until the leaders do not change.
```

Some characteristics of the leader algorithm

- Very large sets of units can be efficiently clustered using the leader algorithm, while the standard agglomerative hierarchical algorithm has some limits on the number of units.
- The leader algorithm is a **local** optimization procedure. Different initial sets of leaders can provide different local optima and corresponding partitions. Consequently, several initial sets of leaders should be used to assess the set of obtained solutions to the clustering problem.

Steps of solving a clustering problem

1. Select the set of units.
2. Observe or measure appropriate variables according to the given problem.
3. Choose an appropriate dissimilarity according to the given problem and type of measured variables.
4. Choose an appropriate type of clustering (e.g., partition, hierarchy).
5. Select or create an appropriate criterion function (e.g., Ward criterion function).
6. Choose or devise an algorithm for the given clustering problem.
7. Determine the clustering(s) which optimize(s) the chosen criterion with the selected algorithm.
8. Apply various tests to detect whether the obtained solutions has some underlying structure or not. Use descriptive statistics to summarize the characteristics of each cluster.

Example 1: Small companies

Research problem

Which are typical groups of owners of small companies who have similar believes on what is important for the business success of their companies?

Data

The population consists of small companies employing at least 1 and up to 50 employees in all sectors of the economy except agriculture in Slovenia. A random sample was drown from the data-bank of the Chamber of Commerce of Slovenia and from the Crafts Chamber. Out of the 200 companies selected, 151 agreed to participate in the study. Each company was visited and a questionnaire was filed by personal interview. The survey was conducted in 1993.

Survey question:

Please indicate your opinion about the influence of the following factors on business success of your company on 5 point scale (1 - not important at all, 5 - very important)?

PROD-MET	– improvement of productive methods
MARK-MET	– improvement of marketing methods
PRODUCT	– improvement of products
RELATION	– good relations among employees
SKIL-EMP	– skilled employees
SKIL-MAN	– skilled managers
FAMILY	– support of the family
ECON-ASO	– support of economic associations
POL-CON	– political connections
LOC-AUT	– support of local authorities
STATE	– support of the state
COMPANY	– support of other companies

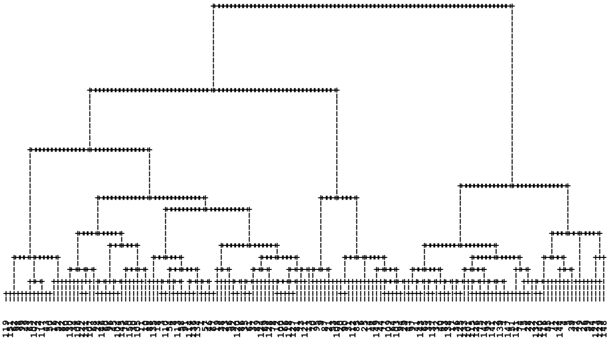
Results

The solution was obtained by the hierarchical aglomerative algorithm (Ward metod).

The variables were standardized and euclidean distance calculated.

The dendrogram on the next slide shows 5 clusters.

Dendrogram



Averages for 5 obtained clusters

	avr.	C 1		C 2		C 3		C 4		C 5	
PROD-MET	3.46	4.06	+	3.81		2.22	-	1.79	- -	4.33	+ +
MARK-MET	3.85	3.92		4.13	+	3.67	-	3.00	- -	4.53	+ +
PRODUCT	3.57	4.00		4.25	++	2.28	- -	1.96	- -	4.47	+ +
RELATION	3.82	3.65		4.38	+	4.22	+	2.79	- -	4.53	+ +
SKIL-EMP	4.04	4.02		4.53	+	4.11		3.00	- -	4.67	+
SKIL-MAN	4.03	3.92		4.38	+	3.83		3.54	- -	4.73	+ +
FAMILY	3.44	3.08	-	3.59		4.00	+	2.92	- -	4.73	+ +
ECON-ASO	2.78	2.77		3.84	++	3.50	+	1.75	- -	1.33	- -
POL-CON	2.07	2.08		3.28	++	1.89		1.25	-	1.00	- -
LOC-AUT	2.53	2.29		3.72	+	4.22	+ +	1.25	- -	1.00	- -
STATE	2.69	2.23		3.81	+	4.39	+ +	1.58	- -	1.93	-
COMPANY	2.19	2.34		3.13	++	2.00		1.38	-	1.07	- -
n. of units	151	62		32		18		24		15	

Clusters description

cluster	number of cases	cluster description
1	62	AVERAGE
2	32	YES - SAYERS
3	18	BAD GUYS
4	24	NO - SAYERS
5	15	GOOD GUYS

Example 2: Clustering of Slovenian sociologists

The agglomerative hierarchical procedure is applied to the Slovenian researchers publication performance.

The dataset was obtained from the **Current Research Information System** (SICRIS) which includes the information of all active researchers registered at the Slovenian Research Agency and at the co-operative **On-Line Bibliographic System & Services** (COBISS) which officially maintains database of all publications published by Slovenian reserachers.

In this study the units are researchers who were in September 2008 in SICRIS registered to work in the field of sociology in Slovenia in the years from 1996 to 2007. There are 89 sociologists studied.

Publication performance is measured by the number of publications by

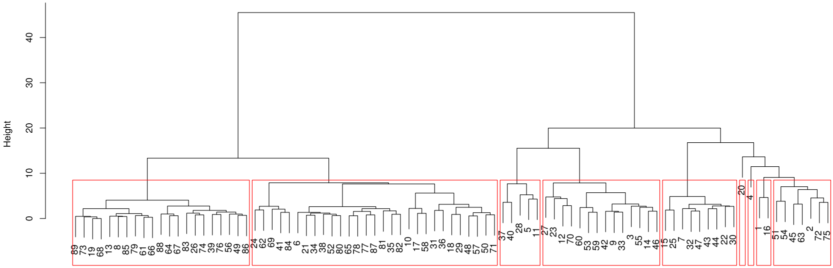
- type of publication: articles in the journals with an impact factor, other original scientific articles, chapters in scientific monographs, and scientific monographs
- language: English, Slovenian, or other languages.

The publication performance variables

- number of articles in journals with an impact factor,
- number of articles in English language scientific journals,
- number of chapters in English language scientific monographs,
- number of scientific English language monographs,
- number of articles in Slovene scientific journals,
- number of chapters in Slovene scientific monographs,
- number scientific Slovene monographs,
- number of articles in scientific journals in other languages,
- number of chapters in scientific monographs in other languages,
- number of scientific monographs in other languages.

Hierarchical clustering of sociologists according to their publication performance

Dissimilarity between researchers was measured by the euclidean distance. The Ward dendrogram was obtained for clustering of 89 sociologists considering the standardized variables.



Average publication performance of obtained clusters

		journals with IF	in English language			in Slovenian language			in other languages		
cluster	N		journals	chapters	books	journals	chapters	books	journals	chapters	books
1	21	0,43	0,48	0,71	0	1,71	0,62	0,67	0	0	0
2	29	0,41	1,21	2,14	0,1	3,52	3,1	1,86	0,14	0,07	0
3	5	0,2	1,4	3,2	0,2	3,8	3,2	4,4	2,8	2,8	0,2
4	14	0,71	0,93	4,43	1,43	3,5	2,86	3,36	0,43	0,71	0,07
5	9	0,56	0,89	1,11	0,44	10,89	5,78	2	0,11	0,11	0
6	1	0	3	9	2	3	11	12	0	2	3
7	1	14	9	4	3	3	5	0	2	1	0
8	2	1	2,5	8	1	7,5	12,5	9	0,5	0	0
9	7	2,86	7,14	4,14	0,14	9,43	4,86	2,71	0,57	0,14	0,29
together	89	0,82	1,57	2,51	0,4	4,39	3,21	2,29	0,36	0,35	0,08

Clustering into 9 clusters

- **Cluster 1**: researchers with the lowest performance.
- **Cluster 2**: sociologists with still below average performance.
- **Cluster 3**: they mostly publish Slovene monographs and publications in other languages.
- **Cluster 4**: they publish English chapters and monographs.
- **Cluster 5**: they mostly publish articles in Slovene journals.
- **Cluster 6**: (s)he mostly publishes chapters and scientific monographs in all languages and less articles in journals.
- **Cluster 7**: (s)he typically publishes articles in journals with an impact factor and monographs in English language.
- **Cluster 8**: they mostly publish English and Slovene chapters and monographs.
- **Cluster 9**: they typically publish articles in English and Slovene journals.

Main results

The clustering of the sociologists according to the publication performance variables only shows that they publish their research results in very specific ways. E.g., some of them publish mostly (only) in the Slovenian language, some of them just chapters in the scientific monographs, some of them typically in English journals.

The results clearly show that there is no typical common culture of the publishing performance in the field of sociology in Slovenia.

Benefits from the optimizational approach to clustering problem

The *optimizational approch* to clustering problem offers two possibilities to adapt to a concrete clustering problem: the definition of the *criterion function P* and the specification of the *set of feasible clusterings Φ* (see Ferligoj, A., Kronegger, L. 2009).
For example: *blockmodeling* is searching for a clustering according to the **relational data only** and the solution can be obtained by an appropriatelly defined *criterion function*.
If a clustering is searched for **attribute and relational data** *clustering with relational constraint* can be used. Here, an appropriately defined set of *feasible clusterings* according to the relational data has to be determined and an appropriately defined *criterion function* according to the attribute data.