Journal of Classification 28 (2011)
DOI: 10.1007/s00357-011-9084-x

Clustering of Distributions: A Case of Patent Citations

Nataša Kejžar

University of Ljubljana, Slovenia

Simona Korenjak-Černe

University of Ljubljana, Slovenia

Vladimir Batagelj

University of Ljubljana, Slovenia

Abstract: Often the data units are described with discrete distributions (work described with citation distribution over time, population pyramid described as age-sex distribution etc.). When the set of such units is very large, appropriate clustering methods can reveal the typical patterns hidden in the data.

In this paper we present an adapted leaders method combined with a compatible adapted agglomerative hierarchical method that are based on relative error measure between a unit and the corresponding cluster representative–leader. The proposed approach is illustrated on citation distributions derived from the data set of US patents from 1980 to 1999. These new methods were developed because clustering of units, described with distributions, with classical *k*-means method reveals patterns with single high peaks which correspond to a single year. These patterns prevail over other distribution shapes also present in the data. Compared with centers in *k*-means method, clusters' representatives obtained with the proposed new methods better detect typical distribution shapes of units. The obtained main cluster types for different sets of units show three main patterns: patents with early or late peak of importance to the community, and patents where the importance is slowly increasing throughout the time period.

Keywords: Clustering; Distribution; Leaders method; *k*-means method; Agglomerative hierarchical clustering method; Temporal citation distribution; Citation network; Relative error measure; Patents.

The authors would like to thank the anonymous referees for many valuable comments and suggestions how to improve this paper. This work was partially supported by the Slovenian Research Agency, Project J1-6062-0101.

Authors' Addresses: Nataša Kejžar, University of Ljubljana, Faculty of Medicine, Institute of Biostatistics and Medical Informatics, IBMI, Vrazov trg 2, 1000 Ljubljana, Slovenia, e-mail: natasa.kejzar@mf.uni-lj.si; Simona Korenjak-Černe, University of Ljubljana, Faculty of Economics, Department of Statistics, e-mail: simona.cerne@ef.uni-lj.si; Vladimir Batagelj, University of Ljubljana, Faculty of Mathematics and Physics, Department of Mathematics, e-mail: vladimir.batagelj@fmf.uni-lj.si.

Published online

1. Introduction

A motivation for this paper stems from the question of how to determine main patterns or shapes of the works' (scientific papers, books, patents etc.) citation distributions.

A citation network consists of works (vertices) and their citations (arcs). For each work we know also the information about its first appearance (publication year). Since a newly published work can cite only works that existed (were published) before it, the citation networks are (almost) acyclic. From the time of the first appearance of each work and the citation relation it is easy to construct for each work the distribution of its citations through time–*temporal citation distribution*. The work was published somewhere in the past and from that time on, it was open to citations. The shape of the temporal citation distribution carries information on the importance or "relevance" of the work at subsequent points in time. If the number of citations is increasing in time, the work is becoming more and more interesting, whereas if the number of citations decreases, it has very likely already reached its peak of importance. As such, the shape of the temporal citation distribution is one of the indicators of the work's importance, although the importance of works can be measured also in other ways (IMU Report 2008). An interesting approach for describing the shape of the temporal citation distribution could be functional data analysis (see Ramsey 2005).

It is always a challenge how to include this type of information into the methodology for selection of the appropriate quality measure of such works. Some approaches how to include time in the evaluation of cited works can be found for example in bibliometrics (Garfield 1998a; Research report by Universities UK 2007 or Sidiropoulos, Katsaros, and Manolopoulos 2006 and Katsaros, Sidiropoulos, and Manolopoulos 2007). We have to be aware of some essential differences that ought to be considered when we evaluate bibliographic works (Garfield 1985; 1998b).

The aim of this article is to find and describe an appropriate clustering method that reveals main patterns or shapes of the works' citation distributions. Since the set of works is usually large, the first selection of clustering approach is frequently the combination of standard *k*-means (leaders) method and Ward's clustering method. Because of the nature of the squared Euclidean distance on which these two methods are based, the final clustering gives mostly cluster representatives with single high peak. To obtain more informative results we are proposing to use new error measures on which the methods are based and derive the corresponding adapted leaders method and agglomerative hierarchical clustering method.

In the next section we present the US patents data set that is used throughout the paper as an example of the data where the new, adapted methods can be used. Section 3 presents the heart of our work. Several new error measures between unit and cluster representative (leader) are proposed and an illustration of how to use them in generalized leaders method is presented. When using the combination of a leaders method (to reduce the size of the data) and a hierarchical method (to determine the final number and composition of clusters), the use of compatible methods is preferred. Therefore we also adapt the agglomerative hierarchical clustering method according to the proposed new error measures.

Further, the results obtained with the new methods on US patents data set are presented. To validate the obtained clusters, we inspect whether they can be characterized by values of the other patent variables that were not included in the clustering process.

2. Data—US Patents

In this work we use the US patents' data set (see Hall, Jaffe and Tratjenberg 2001). It is a very large (almost 3.8 million patents and more than 16.5 million citations), high quality and easily accessible temporal data set about a real citation network. The data consists of information about patents granted between January 1963 and December 1999. Patent citations are available only from 1975 on. We limited the data in this article on patents and their citations from 1980 till 1999. The in-degree distribution of patents network is close to scale-free distribution (Newman 2005).

Each patent in the data set is described with some additional variables. The patent's *grant year* serves as a time point we use to obtain the empirical temporal citation distributions. For each patent X the time distribution of citations to it—*temporal citation frequency distribution* F_{X}—is created. $F_{\text{X}} = [f_1, f_2, \ldots, f_m]$, where f_i is the number of citations to patent X in the i-th year. To make the citation distributions comparable we decided to convert their descriptions into the relative frequency distributions (for $f \neq 0$)

$$X = [f_1/f, f_2/f, \ldots, f_m/f],$$

where $f = \sum f_i$. Because of the length of the term *relative temporal citation distribution* (a special case of (discrete) distribution) for the unit's description we will use its abbreviation RTCD.

Other variables that we use as explanatory variables are (for a detailed description see Hall et al. 2001):

- *number of citations*.
- *technological category* (1–6): (Hall et al. 2001) simplification grouping of USPTO classification (1 Chemical, 2 Drugs&Medical, 3 Computer& Communication, 4 Electrical&Electronics, 5 Mechanical, 6 Others).

- *assignee type* (1–6): types of assignees are defined according to non-government institutions (2 US, 3 non-US), individuals (4 US, 5 non-US) or government organizations (6); 1 stays for unassigned.
- *generality*: $g_i = 1 - \sum_j^{n_i} s_{ij}^2$, where s_{ij} denotes the percentage of citations *received* by patent i that belong to patent class j out of n_i patent classes. A high generality measure means that a patent got cited by patents of a wide range of fields. It is presumed that such a patent has a wider impact and influences innovations in many different fields.
- *originality*: similarly to generality, $o_i = 1 - \sum_j^{n_i} t_{ij}^2$, where t_{ij} denotes the percentage of citations *made*. If a patent cites patents from a narrow field the measure will be low, whereas the citation of patents from various fields gives the high value of originality.
- *percentage of self-citations*: for each patent with an assignee code the number of citations that it made to (previous) patents that have the same assignee code is counted, and is divided by the number of citations (with the assignee code) that it made. This is the "upper bound" (Hall et al. 2001) for the self-citations percentage.

There were 55,537 patents granted in 1980 and cited at least once in 20 years. RTCDs were determined for the interval of 20 years from 1980 to 1999. Because the procedure uses the relative descriptions, an often cited patent cited only in one year has the same representation as a patent cited only once in that year. We found that patents with very few citations influence the shapes of the clusters too much so it was natural to limit the data set to those patents with at least some prescribed number of citations (8–17,492 patents, 15–6,123 patents, or 20–3,245 patents).

The second data set consists of 22,514 patents granted from 1980–84 and cited at least 15 times in a 15-year time period. Here, we were interested also in the grant year—does the grant year of a patent influence its citation pattern?

Therefore two different data sets were considered:

- patents from *one grant year* only (1980) and observed in a specified time period (20 years from 1980 till 1999);
- patents from *different grant years* (1980–1984), but observed for the same amount of time (same *time window*–15 years from patent's grant year).

3. Methods

The patent RTCD's data set is too large for hierarchical clustering, therefore a non-hierarchical clustering method is used. The drawback of the non-hierarchical methods is that we have to know or guess the appropriate number of clusters. To avoid this problem two methods are combined:

non-hierarchical method for patents, described with RTCDs, and a compatible (based on the same criterion function) hierarchical method for leaders of the obtained clusters of patents. Therefore the clustering of patents was performed in two steps:

(1) clustering a large data set of *patents* with the *leaders method* to produce a smaller set of patent cluster representatives;
(2) clustering *patent cluster representatives* using a *hierarchical method* to reveal the relations among them and to determine the most suitable number of clusters.

3.1 Leaders Method

The most popular non-hierarchical clustering method is k-means method (Forgy 1965; MacQueen 1967; Anderberg 1973). It is a special version of a more general leaders clustering method (Vinod 1969; Hartigan 1975; Diday 1979). To describe the leaders clustering method the following notation will be used: X—*unit*; $\mathbf{U}$—a finite *set of units*; C—a *cluster*, $C \subseteq \mathbf{U}$ and $C \neq \emptyset$; $\mathbf{C}$—*clustering*, $\mathbf{C} = \{C_i\}$; Φ—a set of *feasible clusterings*; P—a *criterion function*, $P : \Phi \to \mathbb{R}_0^+$, the nonnegative reals. With these notions we can express the *clustering problem* (Φ, P) as follows:

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

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

A clustering $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ is a *partition* iff $\bigcup_{C\in\mathbf{C}} = \mathbf{U}$ and $C_i \cap C_j = \emptyset$, for $i \neq j$. In our case we shall consider $\Phi = \{\mathbf{C} : \mathbf{C}$ is a partition in k clusters$\}$. Each partition determines a clustering function $c : \mathbf{U} \to \mathbb{N}$

$$c(\mathrm{X}) = i \ \Leftrightarrow\ \mathrm{X} \in C_i.$$

From a set of leaders Λ, for each cluster C a leader Y is selected. The leaders method is based on the *generalized criterion function*

$$W(\mathbf{C}, \mathbf{L}) = \sum_{\mathrm{X}\in\mathbf{U}} \Delta(\mathrm{X}, \mathrm{Y}_{c(\mathrm{X})}),$$

where $\mathbf{L} = \{Y_1, Y_2, \ldots, Y_k\} \subset \Lambda$ and the *unit's error* $\Delta(\mathrm{X}, \mathrm{Y})$ is the error produced by replacing a unit X with the leader Y. Note that in general the leader Y and unit X can be from different spaces. For example, we could describe clusters by regression lines—therefore $\mathrm{Y} \in \mathbb{R}^2$.

Between the generalized criterion function W and the clustering criterion function P the following relation is required

$$P(\mathbf{C}) = \min_{\mathbf{L}\subset\Lambda} W(\mathbf{C}, \mathbf{L}).$$

As is well known from the literature (Brucker 1978) the problem to get the exact solution of clustering problem is in most cases computationally hard. In practice, local optimization procedures are used since they give quite satisfactory results.

For large data sets the most popular is the leaders method which can be described by the following pseudo-code (Salton 1989; Hartigan 1975):

```
**C**–initial clustering
**repeat**
    determine new leaders $\mathbf{L} = \{\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_k\}$ :
        $\mathrm{Y}_i = \mathrm{argmin}_{\mathrm{Y}} \sum_{\mathrm{X} \in C_i} \Delta(\mathrm{X}, \mathrm{Y}), \qquad i = 1, \ldots, k$
    determine new clustering $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ :
        $c(\mathrm{X}) = \mathrm{argmin}_{i \in 1..k}\, \Delta(\mathrm{X}, \mathrm{Y}_i)$
**until** leaders stabilize
```

The leaders method produces a locally optimal solutions of the corresponding clustering problem.

In our case a unit $\mathrm{X} = [x_1, x_2, \ldots, x_m] \in \mathbb{R}^m$ is described by values of m variables, and also a leader $\mathrm{Y} = [y_1, y_2, \ldots, y_m] \in \Lambda$, is presented as m-dimensional vector from $\mathbb{R}^m$. We assume the "orthonormal model" (the total error is the sum of the partial errors) in which the unit's error $\Delta(\mathrm{X}, \mathrm{Y})$ can be expressed component-wise

$$\Delta(\mathrm{X}, \mathrm{Y}) = \sum_{t=1}^{m} \delta(x_t, y_t), \tag{1}$$

where $\delta(x, y)$ is a *component error measure*. It is required that $\delta(x, y) \geq 0$ and $\delta(x, x) = 0$.

The new leader of a cluster is calculated from all units currently in the cluster:

$$\begin{aligned} \mathrm{Y}_i^\star = \mathrm{argmin}_{\mathrm{Y}} \sum_{\mathrm{X} \in C_i} \Delta(\mathrm{X}, \mathrm{Y}) &= \mathrm{argmin}_{\mathrm{Y}} \sum_{\mathrm{X} \in C_i} \sum_{t=1}^{m} \delta(x_t, y_t) \\ &= \left[\mathrm{argmin}_{y_t} \sum_{\mathrm{X} \in C_i} \delta(x_t, y_t) \right]_{t=1}^{m} . \end{aligned}$$

Based on this equality we can further limit ourselves on only one component

$$y_{it}^\star = \mathrm{argmin}_{y_t} \sum_{\mathrm{X} \in C_i} \delta(x_t, y_t),$$

and therefore omit the subscript t. The leader's component of the cluster C can be written in simplified notation as

$$y^\star = \text{argmin}_{y \in \mathbb{R}} \sum_{\mathrm{X} \in C} \delta(x, y), \tag{2}$$

For example, for the classical error measure $\delta(x, y) = (x - y)^2$ the optimal solution is

$$y^\star = \frac{1}{|C|} \sum_{\mathrm{X} \in C} x.$$

Therefore, the well-known *k*-means method is a special version of the leaders method, where units are represented with numerical vectors, the error of a unit is the squared Euclidean distance and the representative (leader) of the cluster is its center of gravity.

The criterion function P can be also written as a sum of *cluster errors* $p(C)$, which in our case can be expressed component-wise as

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C) \qquad \text{and} \qquad p(C) = \sum_{t=1}^{m} p_t(C),$$

where

$$p_t(C) = \sum_{\mathrm{X} \in C} \delta(x_t, y_t^\star). \tag{3}$$

3.2 Example: Patents Granted in Year 1980, Clustered with Classical K-means Method

The classical *k*-means method was applied to the data sets of the most cited patents with at least 8 citations (17,492 patents), with at least 15 citations (6,123 patents) and with at least 20 citations (3,245 patents). In Figure 1 we present two examples of clusters obtained with classical *k*-means method. The gray lines on each graph represent RTCDs of patents. The cluster leader (the center of gravity of the cluster) is marked with a thicker black line. In both panels we can see cluster representatives with single high peak.

In most of the obtained clusters only one year seems to characterize all patents inside the cluster. The main reason for this is the nature of the squared Euclidean distance used for clustering RTCDs that have all the values in the interval $[0, 1]$—its value is mainly determined by few largest values. Citation distributions of some patents can have the highest value in the same year but are otherwise quite different in shape. However, with the squared Euclidean distance they would be classified to the same cluster. This motivated our investigation of the new error measures.

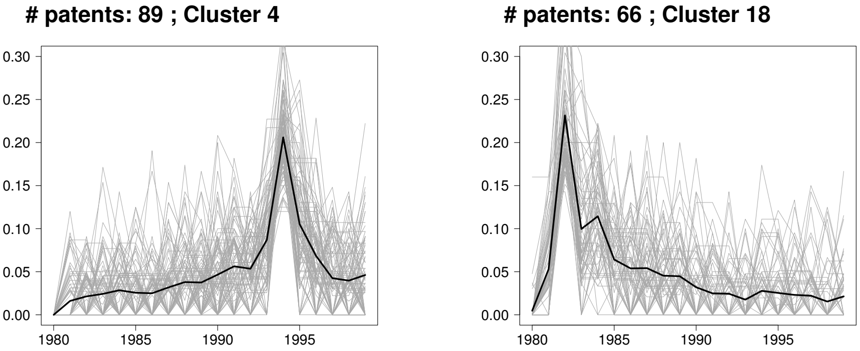

Figure 1. Clusters of patents with patents' temporal distributions and cluster leaders.

To obtain more informative clusters an appropriate relative measure has to be used. In Table 1 nine different component error measures δ are presented together with corresponding leaders $y^\star$–optimal solutions of the problem (2). For error measures δ_3 to δ_9 we assume that $x, y \in \mathbb{R}_0^+$. The first two of them are well-known distances and therefore symmetrical. The error measure δ_1 is the squared Euclidean distance which is used in the classical *k*-means method, and the second one, δ_2, is the city-block or Manhattan distance. The Euclidean and Manhattan distance are both special cases of the more general Minkowski distance (Gower and Legendre 1986). To our knowledge, the component error measures δ_3 to δ_9 represent the novel options to clustering procedures. Measures from δ_3 to δ_6 are asymmetrical.

In error measures special attention should be given to possible divisions with zero. If the numerator is zero, then also the value of the fraction is set to zero, otherwise the value is infinite. In our computer program (clustddist–R package; Kejžar, Korenjak-Černe and Batagelj 2009) a very large value of 10^6 is used instead of infinity.

In this article we selected for the relative component error measure the function δ_4 from Table 1. The adapted leaders method and adapted hierarchical clustering method were developed. Some of the results obtained using the proposed methods on the US patent data set are presented in the following sections. A comparison of methods based on error measures δ_1 to δ_9 were presented at the IFCS 2009 conference (Kejžar et al. 2009).

3.3 Clustering Method with Relative Error Measure

We are proposing a clustering approach based on the component error measure δ_4 from Table 1 for which we derived the corresponding versions of the leaders method and a compatible hierarchical clustering method. As

Table 1. The component error measures with corresponding optimal solutions.

	$\delta(x,y)$	$y^\star$
1	$(x-y)^2$	$y^\star = \frac{1}{\lvert C\rvert}\sum_{\mathrm{X}\in C} x = A_t(C)$
2	$\lvert x-y\rvert$	$y^\star = \mathrm{med}\{x : \mathrm{X}\in C\}$
3	$(\frac{x-y}{y})^2$	$y^\star = \frac{\sum_{\mathrm{X}\in C} x^2}{\sum_{\mathrm{X}\in C} x} = \frac{S_t(C)}{A_t(C)}$
4	$\frac{(x-y)^2}{y}$	$y^\star = \sqrt{\frac{1}{\lvert C\rvert}\sum_{\mathrm{X}\in C} x^2} = \sqrt{S_t(C)}$
5	$(\frac{x-y}{x})^2$	$y^\star = \frac{\sum_{\mathrm{X}\in C}\frac{1}{x}}{\sum_{\mathrm{X}\in C}\frac{1}{x^2}}$
6	$\frac{(x-y)^2}{x}$	$y^\star = \frac{\lvert C\rvert}{\sum_{\mathrm{X}\in C}\frac{1}{x}} = H_t(C)$
7	$\frac{(x-y)^2}{xy}$	$y^\star = \sqrt{\frac{\sum_{\mathrm{X}\in C} x}{\sum_{\mathrm{X}\in C}\frac{1}{x}}} = \sqrt{A_t(C)H_t(C)}$
8	$(\frac{x-y}{x+y})^2$	solve : $y\sum_{\mathrm{X}\in C}\frac{x}{(x+y)^3} = \sum_{\mathrm{X}\in C}\frac{x^2}{(x+y)^3}$
9	$\frac{(x-y)^2}{x+y}$	solve : $\lvert C\rvert = \sum_{\mathrm{X}\in C}(\frac{2x}{x+y})^2$

we showed in the previous section we can limit our derivation only to one component.

The optimal leader of the cluster for component error measure δ is determined by minimizing the sum of all component errors between the units and the leader (see Eq. (2)), therefore

$$\frac{\partial \sum_{\mathrm{X}\in C}\delta(x,y)}{\partial y} = 0.$$

For the component error measure δ_4 we get:

$$\frac{\partial}{\partial y}\left(\sum_{\mathrm{X}\in C}\frac{(x-y)^2}{y}\right) = \sum_{\mathrm{X}\in C}\left(-\frac{(x-y)^2}{y^2} - \frac{2(x-y)}{y}\right) = \sum_{\mathrm{X}\in C}\left(1-\frac{x^2}{y^2}\right) = 0.$$

Solving this equation for the leader y we get

$$y^\star = \sqrt{\frac{1}{|C|}\sum_{\mathrm{X}\in C} x^2} = \sqrt{S_t(C)}.$$

Optimal leaders for all proposed error measures from Table 1 are given in the third column of that table. They are derived in a similar way as for δ_4. Note that for distributions the corresponding leader is not for all δ a distribution itself—the sum of all components is not always 1.

3.4 Hierarchical Method Adapted on the Relative Error Measure

When using the combination of leaders and hierarchical methods as one procedure, it is desirable to preserve the same relation among clusters throughout the procedures—they should be based on the same criterion function $P(\mathbf{C})$, therefore also on the same component error measure $\delta(x, y)$. For example, the *k*-means method and Ward's hierarchical clustering method are both based on the squared Euclidean distance δ_1.

As *k*-means is a special case of (non-hierarchical) leaders method, also Ward's (1963) method belongs to the family of more general agglomerative hierarchical clustering methods. In the agglomerative hierarchical clustering method (Anderberg 1973), two nearest clusters are merged in each step. A procedure produces a series of partitions $(\mathbf{C}_i)$ of the units according to the dissimilarity measure $D(C_p, C_q)$ between clusters. The procedure can be described by the following pseudo-code:

```
$\mathbf{C}_0 = \{\{X\};\ X \in \mathbf{U}\}$ ; $h_0 = 0$
for $k := 1$ to $|\mathbf{U}| - 1$ do
    $(u, v) = \mathrm{argmin}_{(p,q):\ C_p, C_q \in \mathbf{C}_{k-1}} D(C_p, C_q)$
    $C_{u,v} = C_u \cup C_v$
    $\mathbf{C}_k = \mathbf{C}_{k-1} \setminus \{C_u, C_v\} \cup \{C_{u,v}\}$
    $h_k = D(C_u, C_v)$
    update $D(C_{u,v}, C_s), \quad C_s \in \mathbf{C}_k$
endfor.
```

Here h_k is the height of merging of clusters C_u and C_v.

Let us assume the component-wise form of $D(C_p, C_q) = \sum_{t=1}^{m} D_t(C_p, C_q)$. For example, for the component error measure $\delta_1(x, y) = (x - y)^2$ we have $y_C^\star = \frac{1}{|C|} \sum_{X \in C} x = A(C)$ (see Table 1). For disjoint clusters C_u and C_v fused into cluster $C_z = C_u \cup C_v$, where u, v and z are the leaders of corresponding clusters, it holds

$$z = \frac{|C_u| u + |C_v| v}{|C_u| + |C_v|}.$$

It is well known (for example Späth 1977) that for δ_1 we get

$$D(C_u, C_v) = \frac{|C_u| \cdot |C_v|}{|C_u| + |C_v|} d(u, v),$$

the *Ward's dissimilarity* between clusters C_u and C_v, where $d(u, v)$ is the squared Euclidean distance between the centers u and v of the clusters C_u and C_v.

Most of agglomerative hierarchical clustering methods can be seen as greedy procedures based on the model (Batagelj 1988)

$$p(C_z) = p(C_u \cup C_v) = p(C_u) + p(C_v) + D(C_u, C_v),$$

because for it

$$P(\mathbf{C}_k) = P(\mathbf{C}_{k-1}) + D(C_u, C_v)$$

holds. An intuitive explanation of the model would consider the division of the variance (inertia): total Var (of C_z) $=$ within Var (C_u and C_v)$+$between Var(D).

Since also the cluster's error $p(C)$ can be expressed component-wise (see Eq. (3)) we can without loss of generality limit ourselves to only one component and omit the subscript t from further derivations. For component error measure $\delta_4(x,y) = \frac{(x-y)^2}{y}$ we derived the following update formulas for hierarchical clustering. From Table 1 we know:

$$y_C^\star = \sqrt{\frac{1}{|C|}\sum_{\mathrm{X}\in C} x^2} \qquad \text{or} \qquad |C|\, {y_C^\star}^2 = \sum_{\mathrm{X}\in C} x^2.$$

Therefore, considering the last equation, for disjoint clusters C_u and C_v fused into cluster $C_z = C_u \cup C_v$ we have

$$|C_z|z^2 = \sum_{x\in C_z} x^2 = \sum_{x\in C_u} x^2 + \sum_{x\in C_v} x^2 = |C_u|u^2 + |C_v|v^2.$$

This gives (the component of) the leader of the fused cluster:

$$z = \sqrt{\frac{|C_u|u^2 + |C_v|v^2}{|C_u| + |C_v|}}. \tag{4}$$

Using the greedy approach model this yields (for simplicity we use $\delta(x,y) = \delta_4(x,y)$)

$$\begin{aligned} D(C_u, C_v) &= p(C_u \cup C_v) - p(C_u) - p(C_v) \\ &= \sum_{x\in C_u} (\delta(x,z) - \delta(x,u)) + \sum_{x\in C_v} (\delta(x,z) - \delta(x,v)). \end{aligned}$$

Let us evaluate the term

$$\delta(x,z) - \delta(x,u) = \frac{(x-z)^2}{z} - \frac{(x-u)^2}{u} = \frac{ux^2 + uz^2 - zx^2 - zu^2}{uz}.$$

Since $\sum_{x \in C_u} x^2 = |C_u| u^2$, summing the terms over C_u we get

$$\sum_{x \in C_u} (\delta(x,z) - \delta(x,u)) = \frac{u|C_u|u^2 + |C_u|uz^2 - z|C_u|u^2 - |C_u|zu^2}{uz}$$
$$= \frac{|C_u|(u-z)^2}{z} = |C_u|\delta(u,z)$$

and finally:

$$D(C_u, C_v) = |C_u|\delta(u,z) + |C_v|\delta(v,z). \tag{5}$$

Further computation of dissimilarity between clusters gives us

$$\begin{aligned} D(C_u, C_v) &= |C_u|\frac{(u-z)^2}{z} + |C_v|\frac{(v-z)^2}{z} \\ &= \frac{1}{z}(|C_u|(u^2 - 2uz + z^2) + |C_v|(v^2 - 2vz + z^2)) \\ &= \frac{1}{z}(2(|C_u| + |C_v|)z^2 - 2z(|C_u|u + |C_v|v)) \end{aligned}$$

and finally $D(C_u, C_v) = 2(|C_z|z - |C_u|u - |C_v|v)$.

Since $D(C_u, C_v) \geq 0$ it follows from the equation above:

$$z \geq \frac{|C_u|u + |C_v|v}{|C_u| + |C_v|} \qquad \text{(weighted mean of } u \text{ and } v\text{)}.$$

Also (for singletons): $D(\{u\}, \{v\}) = 2(\sqrt{2(u^2 + v^2)} - u - v)$.

Similarly the dissimilarity $D(C_u, C_v)$ can be obtained for other error measures from Table 1. They are presented in Table 2, where $s_u = \sum_{X \in C_u} x$ and $h_u = \sum_{X \in C_u} \frac{1}{x}$. The derivations not published in this paper can be obtained from: `http://www.educa.fmf.uni-lj.si/datana/pub/papers/ClDDisSup.pdf`.

The monotonicity of the proposed relative error measures has not been proven yet. However, in the obtained clusterings of US patents based on δ_4 inversions never occurred.

4. Clustering Results for US Patents Data

4.1 Patents Granted in Year 1980 with at Least 20 Citations

In the first step, clustering of patents is performed with the adapted leaders method that uses error measure δ_4 from Table 1. The results of clustering of patents with at least 20 citations in 40 clusters are presented in

Table 2. The component error measures with corresponding dissimilarities $D(C_u, C_v)$ and new leaders z.

	$\delta(x,y)$	z	$D(C_u, C_v)$
1	$(x-y)^2$	$\frac{\lvert C_u\rvert u+\lvert C_v\rvert v}{\lvert C_u\rvert+\lvert C_v\rvert}$	$\frac{\lvert C_u\rvert\cdot\lvert C_v\rvert}{\lvert C_u\rvert+\lvert C_v\rvert}\delta(u,v)$
3	$(\frac{x-y}{y})^2$	$\frac{us_u+vs_v}{s_u+s_v}$	$\frac{s_u}{u}\delta(u,z)+\frac{s_v}{v}\delta(v,z)$
4	$\frac{(x-y)^2}{y}$	$\sqrt{\frac{\lvert C_u\rvert u^2+\lvert C_v\rvert v^2}{\lvert C_u\rvert+\lvert C_v\rvert}}$	$\lvert C_u\rvert\delta(u,z)+\lvert C_v\rvert\delta(v,z)$
5	$(\frac{x-y}{x})^2$	$\frac{h_u+h_v}{\frac{h_u}{u}+\frac{h_v}{v}}$	$uh_u\delta(u,z)+vh_v\delta(v,z)$
6	$\frac{(x-y)^2}{x}$	$\frac{\lvert C_u\rvert+\lvert C_v\rvert}{h_u+h_v}$	$\lvert C_u\rvert\delta(u,z)+\lvert C_v\rvert\delta(v,z)$
7	$\frac{(x-y)^2}{xy}$	$\sqrt{\frac{s_u+s_v}{\frac{s_u}{u^2}+\frac{s_v}{v^2}}}$	$\frac{s_u}{u}\delta(u,z)+\frac{s_v}{v}\delta(v,z)$

Figure 2 and Figure 3. As in Figure 1, also in this one, the gray lines on each graph represent the RTCDs of single patent and the cluster leader is marked with a thicker black line.

The cluster leaders in this case are more informative, since more than one year is detected to be important for some of the patents' clusters. Although clustering methods do not take the time (temporal ordering) into account, most of the resulting top leaders have a "continuous" shape.

To determine the main patterns of the clusters of patents, obtained by the leaders method, they were further clustered using the compatible hierarchical clustering procedure (second step of clustering). As can be seen from Eq. (4) and Eq. (5), the leaders and dissimilarity measures in hierarchical clustering weigh clusters according to their sizes, therefore the last 10 clusters have negligible influence on the outcome. This allows us to exclude 10 clusters that have less than 10 patents from further analysis.

The dendrogram over the remaining 30 clusters is presented in Figure 4. On the right side of the picture the patent clusters are clustered into 7 clusters, which are represented by graphs of units–leaders and the cluster leader. Light gray lines correspond to the clusters' leaders and black lines that represent main patterns correspond to the leaders of the joint clusters. The quantitative descriptions of these 7 clusters' leaders are given in Table 3. Note that from this table it can be seen that leaders are not necessarily relative distributions (the sum of the components is not necessarily equal to 1).

The optimal cluster leaders are based on a selected relative measure of cluster errors. Although the proposed clustering methods do not consider the time (ordering), the additional hierarchical clustering of the patent clusters pointed out three main patterns that are based on the time (see Figure 5):

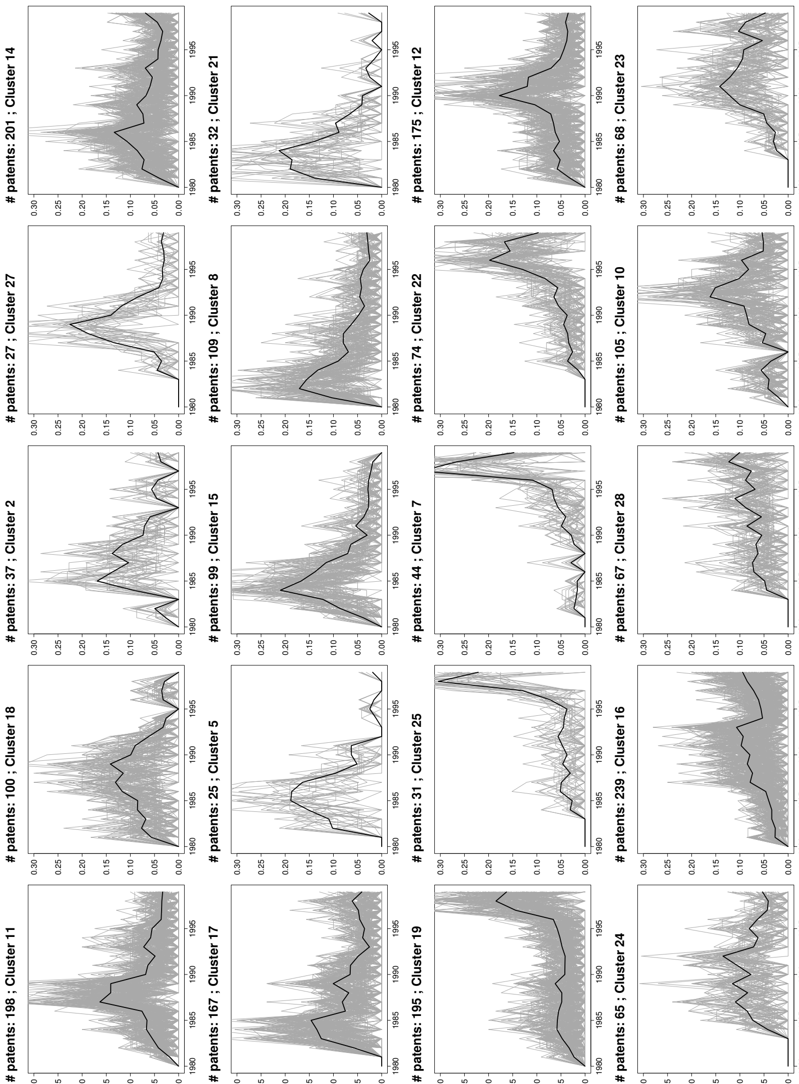

Figure 2. Citations 20+, leaders for δ_4 - 1st part.

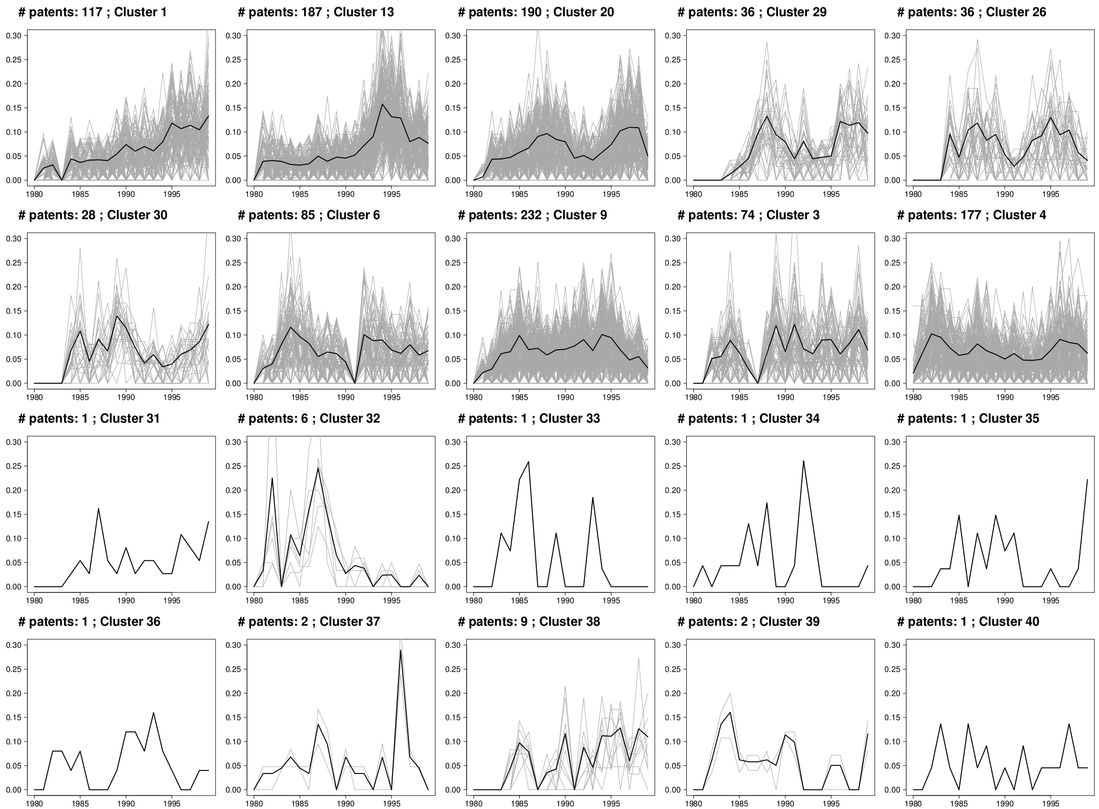

Figure 3. Citations 20+, leaders for δ_4 - 2nd part.

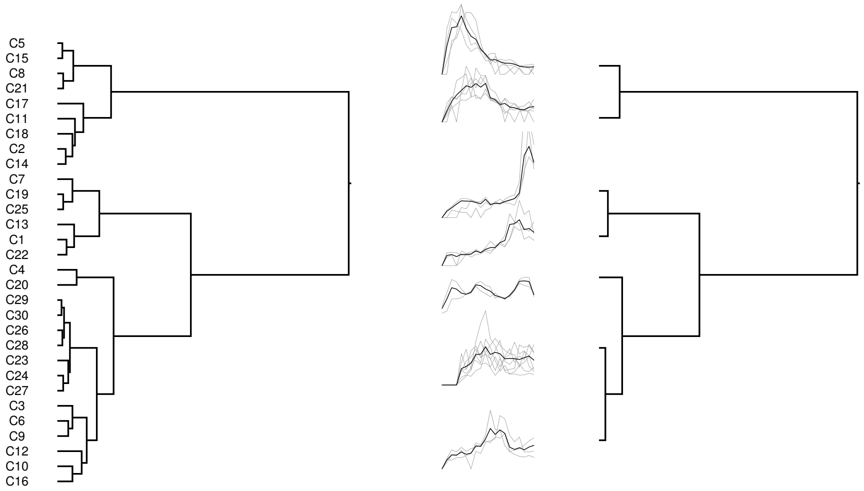

Figure 4. Relations among 30 clusters of patents. The right hand side of the figure is derived from the left.

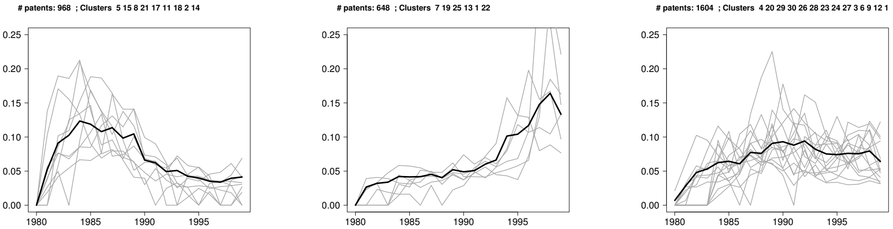

Figure 5. Three main cluster patterns.

(1) the set of patents that reached their importance in the first and second quarter of the observed period and were losing their position towards the end (the first two hierarchical clusters in the right part of Figure 4);
(2) the set of patents whose importance was increasing towards the end of the observed period (the third and fourth hierarchical clusters); and
(3) the set of patents with slightly increasing (or stable) importance that were cited over the whole observed period of 20 years.

We further inspected the seven hierarchical clusters from Figure 4 in order to see if there are any other meaningful differences that have to be taken into account. We took the variables we described in Section 2 and checked if there are any statistically significant differences in the distribution of a variable over the whole data set and the data from a cluster. With this we can additionally evaluate the clusterings that we obtained.

Table 3. The seven leaders from the dendrogram in Figure 4 in numbers. The leaders (columns) in the table follow the dendrogram top down.

	1	2	3	4	5	6	7
# of patents	265	703	270	378	367	327	910
1980	0.00	0.00	0.00	0.00	0.0146	0.00	0.00
1981	0.085	0.0325	0.0195	0.0308	0.0435	0.00	0.0272
1982	0.142	0.0625	0.0297	0.0337	0.0777	0.00	0.0425
1983	0.144	0.0817	0.041	0.0275	0.0734	0.00	0.0418
1984	0.177	0.0956	0.051	0.0342	0.0613	0.0496	0.051
1985	0.140	0.11	0.0501	0.0341	0.0578	0.0564	0.0442
1986	0.115	0.105	0.0497	0.0352	0.0641	0.0666	0.0477
1987	0.106	0.116	0.0459	0.0448	0.0865	0.093	0.0699
1988	0.0763	0.105	0.0422	0.0391	0.0839	0.0933	0.0682
1989	0.0601	0.117	0.0557	0.0493	0.0737	0.116	0.091
1990	0.043	0.073	0.0396	0.0547	0.0672	0.0928	0.123
1991	0.0445	0.0675	0.0436	0.0553	0.054	0.0991	0.104
1992	0.0379	0.053	0.0431	0.0698	0.0497	0.0948	0.118
1993	0.0337	0.0561	0.0488	0.0763	0.0445	0.0794	0.107
1994	0.0334	0.0453	0.0533	0.125	0.0542	0.0806	0.0655
1995	0.0309	0.0428	0.0577	0.127	0.0711	0.0799	0.0589
1996	0.0242	0.0386	0.0751	0.139	0.0972	0.0774	0.064
1997	0.0221	0.0373	0.189	0.109	0.0984	0.0829	0.0577
1998	0.0216	0.0445	0.216	0.113	0.0965	0.0879	0.0662
1999	0.0226	0.0464	0.168	0.101	0.0564	0.0754	0.0713

We used Pearson's χ^2 test of independence for the factor variables category and assignee type and the two-sample Kolmogorov-Smirnov test (Bickel and Doksum 1977) for the interval variables (the number of citations, generality, originality and self-citations). The results of statistical tests for factor variable category can be seen in Table 4. Table for assignee type can be obtained from: `http://www.educa.fmf.uni-lj.si/datana/pub/papers/ClDDisSup.pdf`. The results of statistical tests for the interval variables can be seen in Table 5.

Table 4 shows statistical significance in some clusters that could roughly be summarized with the following: patents of cluster 1 and 2 (that achieve

Table 4. p-values of statistical tests of possible factor explanatory variable category and its relative frequencies for hierarchical clusters and the whole data set.

	category	probabilities for categories (1–6)
cluster 1	0.000**	0.06, 0.59, 0.33, 0.02, 0.00, 0.00
cluster 2	0.001**	0.07, 0.64, 0.27, 0.01, 0.01, 0.01
cluster 3	0.386	0.15, 0.61, 0.23, 0.00, 0.00, 0.00
cluster 4	0.000**	0.19, 0.59, 0.17, 0.03, 0.02, 0.00
cluster 5	0.106	0.15, 0.59, 0.22, 0.02, 0.01, 0.01
cluster 6	0.000**	0.25, 0.56, 0.16, 0.02, 0.01, 0.01
cluster 7	0.000**	0.16, 0.59, 0.21, 0.01, 0.03, 0.00
overall		*0.14, 0.60, 0.23, 0.01, 0.00, 0.01*

Table 5. p-values of statistical tests of possible explanatory variables for hierarchical clusters.

	# citations	generality	originality	selfcitations
cluster 1	0.000**	0.000**	0.094	0.000**
cluster 2	0.382	0.073	0.998	0.001**
cluster 3	0.000**	0.174	0.692	0.001**
cluster 4	0.016*	0.803	0.905	0.064
cluster 5	0.882	0.015*	0.346	0.739
cluster 6	0.000**	0.219	0.981	0.000**
cluster 7	0.616	0.276	0.308	0.081

early peak in citations) are short in the category Chemical and larger in Comp&Comm. Patents of cluster 6 over-represent category Chemical. Patents of cluster 4 and 6 under-represent category Comp&Comm. Last three categories are represented in small percentages which are slightly higher in cluster 4, 6 and partly 7. However, one has to keep in mind that only patents cited at least 20 times are considered in the analysis. The pattern of category percentages could change if a lower citation threshold was taken, i.e. early citation of software patents or patent thickets (intellectual property strategies) of pharmaceutical patents.

Similarly, assignee type appears to be statistically significant for some clusters: clusters 4 and 6 represent patents with more institutional non-US non-government assignees. Clusters 1 and 2 indicate more individual and non-governmental patents which could indicate their stemming from the contemporary usability problems.

Table 5 shows that some of results for the interval variables are statistically significant. However, only the plotting of the *empirical cumulative distribution functions* (ecdf) for the whole data set with ecdf of the specific cluster shows the side at which a significant dislocation occurs. Figure 6 shows that the patents in cluster 1 (also cluster 2 and 6—not shown) exhibit a percentage of self-citations which is significantly larger than that for the whole data set. This could also explain and is consistent with the large number of citations in the first quarter of the time interval. Cluster 3 however exhibits a significantly lower percentage of self-citations—therefore a larger amount of time was necessary to get "acknowledged", and the peak of the distribution is in the last part of the time interval.

Two more variables, that show some statistical significance with respect to clustering are also tightly connected with citations: number of citations and generality. Interestingly, originality does not seem to have influence on main patterns.

The number of citations is significantly larger for clusters 3 and 4 (which is increasing its importance at the end of the time interval), whereas it is lower for clusters 1 (that achieved its peak importance in the first half of the time period) and 6. Generality proves to be statistically significant in cluster 1 (patents in this cluster are far more narrow than the overall data—this somehow connects to the large percentage of self-citations observed in the cluster), and in cluster 5, which is on the contrary—relatively more general (widely dispersed) than the whole data set. We can conjecture that cluster 5 consists of "evergreen" patents—patents that are general and interesting enough to yield a base and be cited by many other patents throughout the observed time period (see Figure 7).

4.2 Patents Granted in Years 1980–84, 15-year Time Period

In Figure 8 and Figure 9 we present the results of analysis using the adapted leaders method on the other data set of 22,514 patents granted from 1980–84 and cited at least 15 times in 15-year time period.

We include all the cluster leaders in further analysis where we use compatible hierarchical clustering of the clusters. The dendrogram is presented in Figure 10. On the right side of the picture the patent clusters are clustered into 7 clusters that are represented by a graph (gray lines correspond to the cluster leaders and black lines to the leaders of joint clusters) and numerically in Table 6.

Similarly to the previous section additional hierarchical clustering of the patent clusters pointed out the same three main patterns (see Figure 11): (1) the set of patents that reached their importance in the first part of the observed period and were losing their position towards the end (upper two

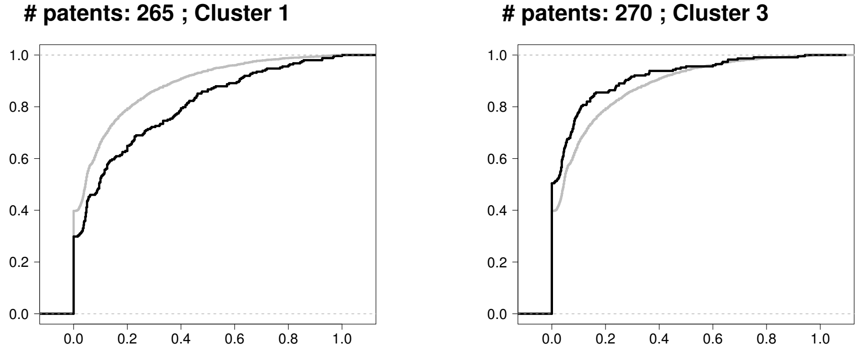

Figure 6. Ecdf for the percentage of self-citations. The gray line represents the ecdf of the whole data set, black line the ecdf of a specific hierarchical cluster.

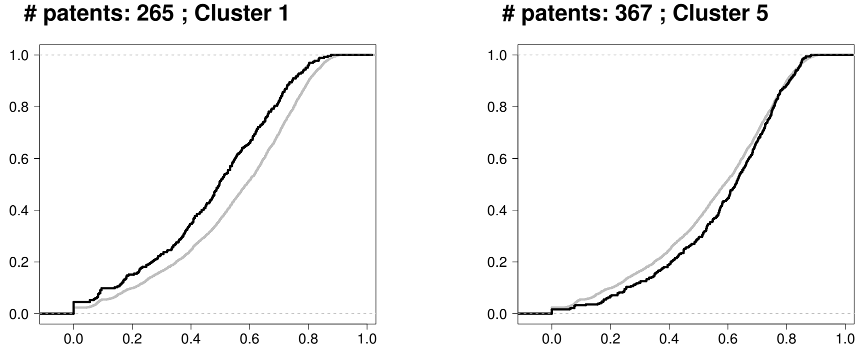

Figure 7. Ecdf for generality. Gray line represents the ecdf of the whole data set, black line the ecdf of a specific hierarchical cluster.

hierarchical clusters in the right part of Figure 10); (2) the set of patents with slightly increasing importance over all the observed period of 15 years; and (3) the set of patents whose importance was slightly increasing in the first two quarters of the observed period and slightly decreasing in the second two (the last four hierarchical clusters).

From further inspection of the seven hierarchical clusters (i.e. checking statistically significant differences in the distribution of other variables from the clusters—see the explanation from previous section) we were especially interested whether the grant year of patents in any way influences citation patterns. We used Pearson's χ^2 test of independence.

Grant year (see Table 7) is a statistically significant variable. The oldest patents are over-represented in the last four clusters, specially in cluster 4 and 5. With the youngest patents the pattern is reversed.

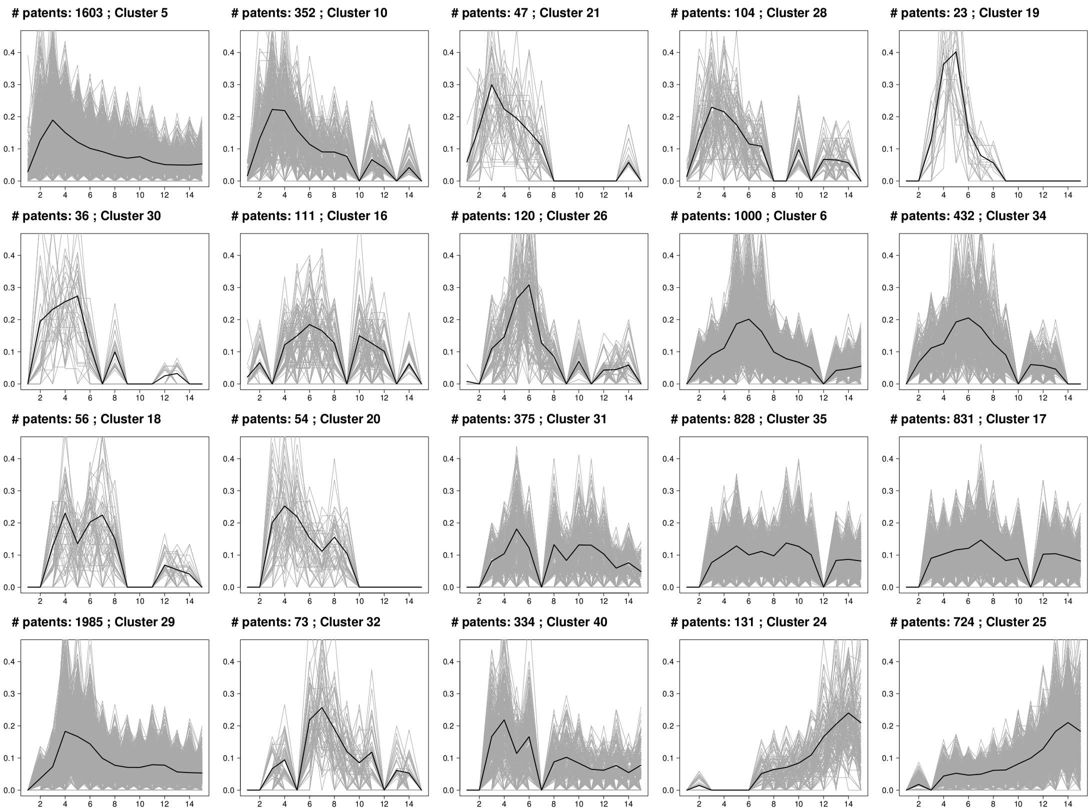

Figure 8. Citations 15+, leaders clustering based on δ_4 in 40 clusters, years 1980–84 - 1st part.

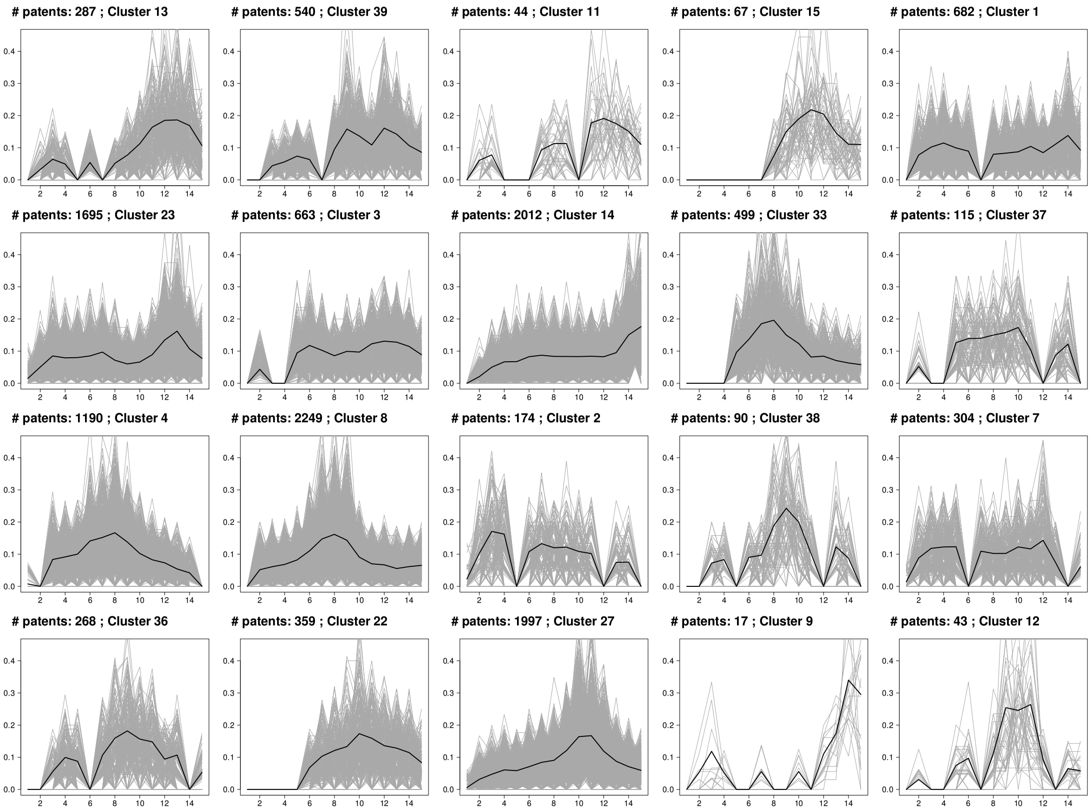

Figure 9. Citations 15+, leaders clustering based on δ_4 in 40 clusters, years 1980–84 - 2nd part.

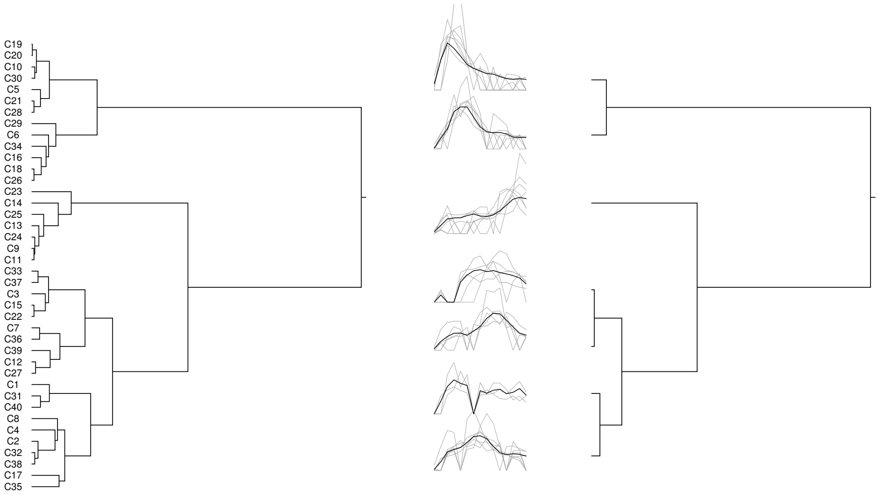

Figure 10. Relations among 40 clusters of patents.

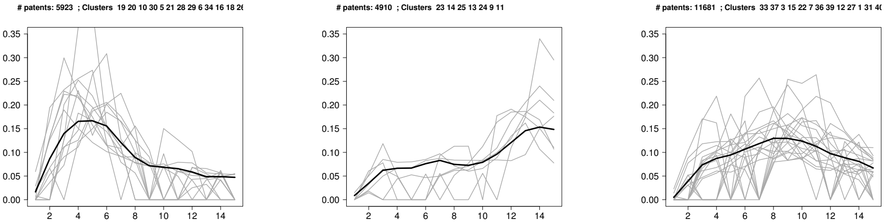

Figure 11. Three main cluster patterns.

We could conjecture that the grant year pattern emerges due to the increased tendency for citation in the later time periods (from the average of 5 citations per patent in 1975 to 10 in 1995). This happened partly because of the computerization of the patent data base in the mid 1980s (Hall et al. 2001).

Tables with detailed results of tests for other variables can be obtained from: `http://www.educa.fmf.uni-lj.si/datana/pub/papers/ClDDisSup.pdf`. As in the previous example, category and assignee type are statistically significant. Patents of the first two clusters are under-represented in category Chem, and over represented in category on Comp&Comm. The opposite is valid for the most of the other clusters. Non-governmental non-US institutional assignees are over-represented in cluster 3 (with patents with slightly increasing importance across all the observed period).

Table 6. The seven leaders from the dendrogram in Figure 10 in numbers. The leaders (columns) in the table follow the dendrogram top down.

	1	2	3	4	5	6	7
# of patents	2,219	3,705	4,909	1,703	3,152	1,391	5,435
1	0.0268	0.00395	0.00878	0.00	0.00642	0.00	0.00536
2	0.128	0.0462	0.0348	0.0303	0.0372	0.0542	0.0379
3	0.201	0.0845	0.0621	0.00	0.0578	0.116	0.0793
4	0.177	0.158	0.0664	0.00	0.0718	0.144	0.0895
5	0.144	0.179	0.0669	0.0853	0.0726	0.130	0.0978
6	0.108	0.178	0.0761	0.115	0.063	0.121	0.120
7	0.0923	0.135	0.083	0.133	0.0812	0.00	0.144
8	0.081	0.0937	0.0747	0.138	0.100	0.0984	0.148
9	0.0697	0.0729	0.0726	0.129	0.135	0.0881	0.135
10	0.0677	0.0689	0.079	0.134	0.157	0.100	0.103
11	0.057	0.0702	0.0962	0.126	0.154	0.105	0.076
12	0.0488	0.0637	0.121	0.120	0.127	0.0856	0.0681
13	0.045	0.0506	0.146	0.113	0.099	0.0917	0.0715
14	0.0477	0.0493	0.153	0.103	0.072	0.108	0.0692
15	0.0452	0.0485	0.148	0.0774	0.0641	0.0795	0.0615

Table 7. p-values of statistical tests for grant year and their relative frequencies for hierarchical clusters and the whole data set.

	grant year	probabilities for years (1980–1984)
cluster 1	0.001**	0.12, 0.17, 0.18, 0.23, 0.30
cluster 2	0.020*	0.13, 0.19, 0.18, 0.22, 0.28
cluster 3	0.000**	0.13, 0.15, 0.17, 0.22, 0.34
cluster 4	0.000**	0.18, 0.21, 0.19, 0.20, 0.21
cluster 5	0.000**	0.19, 0.21, 0.18, 0.20, 0.22
cluster 6	0.005**	0.16, 0.17, 0.20, 0.18, 0.30
cluster 7	0.000**	0.15, 0.19, 0.20, 0.21, 0.25
overall		*0.15, 0.18, 0.18, 0.21, 0.28*

Interval variables also show statistically significant dislocations which match the ones from the previous section. Clusters 1 and 2 (early peak of importance) show statistically significant higher percentage of self-citations, smaller generality and number of citations. On the other side, clusters 4, 5 and 7 (later peak) and 3 (still increasing) receive a significantly larger number of citations and have lower percentage of self-citations. The difference is observed in the variable originality. It is statistically significant for clusters 1 (lower) and 3 (higher than overall). A possible explanation is, that patents with early peak of importance gather information from a narrower pool of knowledge (these patents are more "instant"), whereas patents with slowly increasing number of citations do the opposite. They contribute to more fields of knowledge which gradually find them interesting.

5. Conclusions

We presented an adapted leaders method and a compatible agglomerative hierarchical clustering method based on relative error measure for clustering units described with distributions. This study was motivated by unexpected patterns obtained in analyses of temporal citation distributions of patents (e.g. US patents data) with classical clustering methods. Since the data set was too large to use only hierarchical clustering method, the combination of non-hierarchical and hierarchical clustering methods was used.

When clustering units with the well-known *k*-means method, the resulting patterns were typically of single high peak because of the nature of the squared Euclidean distance on which the method is based. Single high peak is determined by the highest value of the patents with little influence of the values in their other data points. To diminish the influence of large values on the error measure that the algorithm minimizes, several relative error measures between units in the cluster and the corresponding cluster leader are proposed in this article. Special attention should be given to cases when in the measures the division by zero is possible. Based on these error measures and their corresponding optimal leaders, adapted leaders and adapted agglomerative hierarchical clustering methods were developed. They proved to be efficient for large data sets.

To observe the relations among clusters and also to avoid the problem with the selection of an appropriate number of clusters, a larger number of clusters obtained by the non-hierarchical clustering method (e.g. an adapted leaders method) should be clustered further using a compatible (according to error measure) hierarchical clustering method. Therefore we also derived the dissimilarities between clusters in the adapted agglomerative hierarchical clustering methods compatible with the proposed relative error measures.

The procedure based on the adapted leaders method and adapted hierarchical clustering method with error measure δ_4 was applied on the data on US patents. The results were compared with the clustering results obtained with k-means method. The adapted version gave more relevant and informative results than the methods based on squared Euclidean distance. The expected forms of cluster shapes were obtained. The clusters show that temporal citation distributions are mainly unimodal, but there exist a non-ignorable number of patents that do not exhibit a noticeable peak in their citation distribution (e.g. hierarchical cluster 5 from the first data set).

Differences in citation patterns were observed for different grant years. This may be due to an increase in patents citation in the later time periods.

Other variables of patents were inspected in order to evaluate the obtained results of clusterings. They were not included in the clustering procedure, but nevertheless some of them turned out to characterize the clusters. We expect that further applications by researchers who work in scientometrics will provide better understanding of the proposed methods.

Although each year was considered as independent in the proposed approach, adapted clustering methods reveal three main patterns of patent citation distributions: (1) the set of patents that reached their importance in the second quarter of the observed period and are losing their position towards the end; these patents are narrower in field coverage, less frequently cited than overall and show large percentage of self-citations (2) the set of patents whose importance was increasing later in the observed period; these patents are more frequently cited than overall and show significantly lower percentage of self-citations, and (3) the set of patents with slightly increasing importance, that were cited over the whole observed period; they exhibit significantly larger field coverage and could possibly be seen as a base for newer patents.

From the obtained results we can say that the clustering procedures based on relative error measure gave more informative results compared with the results obtained by using classical clustering procedures. The clustering methods based on all proposed error measures were implemented in R package “clustddist”. In the future, we plan to make the package “clustddist” available at CRAN (The Comprehensive R Archive Network, see R Development Core Team 2008), compare the impacts of different measures and apply the methods on other data sets with units described as frequencies or distributions.

References

ANDERBERG, M.R. (1973), *Cluster Analysis for Applications*, New York: Academic Press.

BATAGELJ, V. (1988), “Generalized Ward and Related Clustering Problems”, in *Classification and Related Methods of Data Analysis*, ed. H.H. Bock, North-Holland: Amsterdam, pp. 67–74.

BICKEL, P.J., and DOKSUM, K.J. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, Oakland: Holden-Day, Inc.

BRUCKER, P. (1978), "On the Complexity of Clustering Problems", in *Lecture Notes in Economics and Mathematical Systems: Optimizing and Operational Research*, eds. R. Henn, B. Korte, and W. Oletti, Berlin: Springer, pp. 45–54.

CLUSTDDIST–R PACKAGE (2009), *Test Version of an R Package for Clustering of Distributions*, by N. Kejžar, V. Batagelj, and S. Korenjak-Černe, `https://r-forge.r-project.org/projects/clustddist/`.

DIDAY, E. et al. (1979), *Optimisation en classification automatique, Tomes 1., 2.*, Rocquencourt: INRIA.

FORGY, E.W. (1965), "Cluster Analysis of Multivariate Data: Efficiency Vs. Interpretability of Classifications", *Biometrics, 21*, 768–769.

GARFIELD, E. (1985), "Uses and Misuses of Citation Frequency", *Current Contents. Essays of an Information Scientist, 8*, 403–409.

GARFIELD, E. (1998a), "Long-Term Vs. Short-Term Journal Impact: Does It Matter?", *The Scientist, 12,* 3.

GARFIELD, E. (1998b), "The Impact Factor and Using It Correctly", *Der Unfallchirurg, 101(6)*, 413.

GOWER, J.C., and LEGENDRE, P. (1986), "Metric and Euclidean Properties of Dissimilarity Coefficients", *Journal of Classification, 3*, 5–48.

HALL, B.H., JAFFE, A.B., and TRATJENBERG, M. (2001), "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools", NBER Working Paper 8498, NBER, `http://papers.nber.org/papers/w8498.pdf`.

HARTIGAN, J.A. (1975), *Clustering Algorithms*, New York: Wiley-Interscience.

HIRSCH, J.E. (2005), "An Index to Quantify an Individual's Scientific Research Output", *Proceedings of the National Academy of Sciences of the United Stated of America, 102*, 16569–16572.

IMU REPORT (2008), "Citation Statistics. A Report from the International Mathematical Union (IMU) in Cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)", by R. Adler, J. Ewing, and P. Taylor, `http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf`.

KAUFMAN, L., and ROUSSEEUW, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley.

KATSAROS, D., SIDIROPOULOS, A., and MANOLOPOUS, Y. (2007), "Age Decaying H-Index for Social Network of Citations", *Proceedings of Workshop on Social Aspects of the Web*, Poznan, Poland, April 27.

KEJŽAR, N., KORENJAK-ČERNE, S., and BATAGELJ, V. (2009) "Clustering of Discrete Distributions: New R Package and Comparison of Its Methods", Abstract for the International Conference IFCS 2009 in Dresden, March 2009.

MACQUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations", *5th Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297.

NEWMAN, M.E.J. (2005), "Power Laws, Pareto Distributions and Zipf's Law", *Contemporary Physics, 46, 5*, 323–351.

RAMSEY, J., and SILVERMAN, B.W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer-Verlag.

R DEVELOPMENT CORE TEAM (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, `http://www.R-project.org`.

RESEARCH REPORT BY UNIVERSITIES UK (2007), "The Use of Bibliometrics to Measure Research Quality in UK Higher Educational Institutions", 40, October 2007, `http://www.universitiesuk.ac.uk/Publications/Pages/Publication-275.aspx`.

SALTON, G. (1989), *Authomatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Reading, Massachusetts: Addison-Wesley.

SIDIROPOULOS, A., KATSAROS, D., and MANOLOPOUS, Y. (2006), "Generalized H-index for Revealing Latent Facts in Social Networks of Citations", *Proceedings of the 4th ACM International Workshop on Link Analysis: Dynamics and Static of Large Networks (LinkKDD), (in conjunction with ACM KDD)*, ACM Press, pp. 45–52.

SPÄTH, H. (1977), *Cluster-Analyse-Algorithmen*, München: R. Oldenbourg.

VINOD, H. (1969), "Integer Programming and the Theory of Grouping", *Journal of American Statistical Association, 64*, 506–517.

WARD, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association, 58*, 236–244.