Croatian Operational Research Review
CRORR **7**(2016), 81–96

Research project grouping and ranking by using adaptive Mahalanobis clustering

Željko Turkalj[1], **Damir Markulak**[2], **Slavica Singer**[1] **and Rudolf Scitovski**[3,*]

[1] *Faculty of Economics, Josip Juraj Strossmayer University of Osijek*
Trg Ljudevita Gaja 7, 31 000 Osijek, Croatia
E-mail: ⟨{*turkalj, singer*}*@efos.hr*⟩

[2] *Faculty of Civil Engineering, Josip Juraj Strossmayer University of Osijek*
Crkvena 21, 31 000 Osijek, Croatia
E-mail: ⟨*markulak@gfos.hr*⟩

[3] *Department of Mathematics, Josip Juraj Strossmayer University of Osijek*
Trg Ljudevita Gaja 6, 31 000 Osijek, Croatia
E-mail: ⟨*scitowsk@mathos.hr*⟩

Abstract. The paper discusses the problem of grouping and ranking of research projects submitted for a call. The projects are grouped into clusters based on the assessment obtained in the review procedure and by using the adaptive Mahalanobis clustering method as a special case of the Expectation Maximization algorithm. The cluster of projects assessed as best is specially analyzed and ranked. The paper outlines several possibilities for the use of data obtained in the review procedure, and the proposed method is illustrated with the example of internal research projects at the University of Osijek.

Key words: adaptive Mahalanobis clustering, multi-criteria decision making, evaluation, project clustering

Received: February 11, 2016; accepted: March 21, 2016; available online: March 31, 2016

DOI:10.17535/crorr.2016.0006

1. Introduction

Differences and similarities between certain phenomena are always an intriguing starting point not only for researchers, but also for decision makers, in which argumentation of the similarities and differences is important, e.g., for achieving as equitable allocation of limited resources (financial, human, material, etc.) as possible.

Why are the elements of some set more compact and separated better for some values of their features and how to group them better? For example, grouping a set of interested buyers of sports shoes with respect to age, education and purchasing power can be used to define the promotion policy of a manufacturer of sports shoes or grouping university students depending on the type of their previous education

*Corresponding author.

and the achieved GPA can be used to define the admissions policy of that university [4, 8, 14].

By including the criteria referring to limiting resources and expectations of an equitable distribution of such resources to mutually competitive activities, the answer to such question becomes exceptionally important from the application point of view in a situation where there is certain homogeneity of phenomena or activities that need to be assessed.

The issue of quality/excellence assessment of one's scientific achievements or research proposal is very topical and important not only for researchers, but also for the wider community the individual belongs to. There are many discussions referring thereto that have been published in scientific and professional papers and various publications (see e.g., [8, 12, 10, 2]), but also in the daily newspapers, which in fact is not surprising if one knows that the consequence of this process is the distribution of financial resources for the purpose of research, which are always limited. Hence, the debate is most often about whether the method of distribution of funds available for research corresponds to the actual scientific excellence of the respective research. And here we come to the basic problem - how someone's scientific performance or research proposal can be assessed in a clear, unambiguous, transparent and fair way?

Josip Juraj Strossmayer University of Osijek administration faced such situation when they decided to encourage research of young researchers through internal funding programs. This will be used as an example to illustrate the proposed method for project grouping and ranking.

A fairly large body of literature is dedicated to the assessment and ranking of research projects (see, for example, [3, 9, 11]) and ranking of departments, institutes and universities closely associated therewith (see, for example, [4, 8, 14]). Most approaches use different multi-criteria decision making methods, most frequently the well-known Analytic Hierarchy Process (AHP) [15]. In our paper, we have combined the AHP method and the adaptive Mahalanobis clustering (AMC) algorithm proposed in [13]. First, the set of projects that have passed the administrative verification was grouped into several clusters depending upon the features used. After that, ranking was conducted within the cluster of projects assessed as best by measuring the relative ranking "distance" from the *perfectly assessed project*, i.e., the project that has achieved the maximum grade possible.

The paper is organized as follows. The description and the structure of data that characterize the projects concerned are given in Section 2. This section also describes in more detail an example of internal competition for research projects at the University of Osijek. Section 3 outlines basic facts about cluster analysis and gives a short description of the AMC algorithm. Different approaches to the construction of the data set on the basis of which projects are grouped and ranked as well as appropriate examples are presented in Section 4.

2. Data

Suppose that N project proposal applications with full documentation were submitted in reply to a call for project proposals. Let us denote this set by $\mathcal{P}_N$. Projects will be assessed on the basis of features $f_1, \ldots, f_n$ describing the quality

of both the applicants and the project (the quality and relevance of the research proposal, the quality of the applicants, etc.) and the general impression F of the project.

By using the well-known AHP method (see, e.g., [3, 4, 15]), to each feature f_s we associate the weights $w_s > 0$, $s = 1, \ldots, n$ with the condition $\sum_{s=1}^{n} w_s = 1$.

Suppose further that for each project $p^i \in \mathcal{P}_N$ two independent, blind reviews $R^i_{\#1}$ and $R^i_{\#2}$ are obtained in which features $f_1, \ldots, f_n$ were assessed by grades $u^i_s, v^i_s \in [1, 5]$, $s = 1, \ldots, n$ and the general impression of the project by grades $U^i, V^i \in [1, 10]$. Grades (u^i_s, v^i_s, U^i, V^i) do not have to be integers. If for some project $p^{i_0} \in \mathcal{P}_N$ one of the grades referring to the general impression U^{i_0} or V^{i_0} is less than 6, such project is considered to be *negatively assessed* and it will not be considered for further evaluation.

For project $p^i \in \mathcal{P}_N$ assessed by grades $\big((u^i_s, v^i_s),\, s = 1, \ldots, n;\, U^i, V^i\big)$ we define the vector $\bar{f}^i$ of the GPA of features $\bar{f}^i_s = \frac{1}{2}(u^i_s + v^i_s)$, $s = 1, \ldots, n$ and the GPA of the general impression $\overline{F}^i = \frac{1}{2}(U^i + V^i)$.

In this way, for every project $p^i \in \mathcal{P}_N$ we have the following data:

$$\begin{aligned} &\bar{f}^i_1, \ldots, \bar{f}^i_n \text{ – the GPA of features based upon reviews } R^i_{\#1} \text{ and } R^i_{\#2}, \\ &\overline{F}^i \qquad\qquad \text{ – the GPA of the general impression of reviewers } R^i_{\#1} \text{ and } R^i_{\#2}, \end{aligned} \tag{1}$$

taking into account corresponding weights of features $w_1, \ldots, w_n > 0$.

Example 1. *Josip Juraj Strossmayer University of Osijek administration decided to encourage research of young researchers by internal funding programs and created a unique fund for that particular purpose. The main goal of this concept is to help young researchers, who have yet to acquire their own scientific recognition, in the implementation of their ideas as this is a difficult time if we take into account the reduced scope of financing scientific research on the national level and the related lower likelihood of approval of funding. Thus, the second call for internal scientific research project proposals was opened in the 2014-2015 academic year. In order to make the process more transparent, detailed conditions of the call as well as the assessment criteria were clearly defined on the website of the University*[‡]*. Two areas of scientific research were identified, i.e., the STEM fields and the fields of arts, humanities and social sciences. The maximum possible score for research, use of funds, etc. were defined for each field. Following standard administrative checks, all project proposals are supposed to undergo a peer-review process with two independent reviewers one of whom is from the specific field the project proposal refers to, and the other covers a broader project proposal research area. Both reviewers had to fill out an appropriate peer-review form which was the basis for establishing project proposal assessment criteria.*

However, for each of these fields, there is an open question of equity of the limited financial resources available to the University for this purpose, which is based on an equal comparison of all project proposals taking into account the same features.

[‡]`http://news.unios.hr/research/projects/open-calls/research-projects/`

In this and the following examples we will work out the problem of grouping and ranking research projects in the fields of arts, humanities and social sciences. $N = 61$ *project applications with full documentation were submitted (that have passed the administrative verification). Each project was reviewed by at least two independent, blinded reviewers:* $R^i_{\#1}$ *was selected from the field the respective project topic belongs to, and* $R^i_{\#2}$ *was selected from another related field. Reviewers* $R^i_{\#1}$ *and* $R^i_{\#2}$ *had similar but adapted forms in which they assessed* $n = 6$ *project features (see Table 1).*

Features	Reviewer $R^i_{\#1}$	Reviewer $R^i_{\#2}$	Weights w_s
f_1	The quality and relevance of the research proposal	The quality and relevance of the research proposal	0.30
f_2	The quality of applicants	The quality of applicants	0.20
f_3	Research feasibility study	Dissemination and utilization of research results 0.15	
f_4	Financial plan	Financial plan	0.10
f_5	Institutional support	Institutional support	0.10
f_6	Inclusion of students	Inclusion of students	0.15

Table 1: *The elements assessed by reviewers from Example 1 with corresponding weights*

Each reviewer also rated the general impression of the proposed project. Where the respective assessments of reviewers $R^i_{\#1}$ *and* $R^i_{\#2}$ *differed substantially, additional reviews were requested. If for some project the grade referring to the general impression was less than* 6*, such project was considered to be* negatively assessed *and it was not considered for further evaluation.*

The set of all positively assessed project proposals with corresponding data of the form (1) will be denoted by $\mathcal{P}$. The set $\mathcal{P}$ needed to be grouped according to their quality and a decision should be made on which projects shall be financed.

3. Data clustering

Clustering or grouping a data set $\mathcal{A} = \{a^i \in \mathbb{R}^n \colon i = 1, \ldots, m\}$ with n features in several compact and well-separated clusters has practical importance in a wide variety of applications, such as biology, medicine, physics, economy, environmental science, energy management, business, social sciences, etc. (see e.g. [1, 14, 17, 18, 19, 22]). A general problem is as follows: the set $\mathcal{A}$ should be partitioned into $1 \leq k \leq m$ nonempty disjoint subsets $\pi_1, \ldots, \pi_k$, such that

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \qquad \pi_r \cap \pi_s = \emptyset, \quad r \neq s, \qquad |\pi_j| \geq 1, \quad j = 1, \ldots, k. \tag{2}$$

Subsets $\pi_1, \ldots, \pi_k$ are called *clusters in* $\mathbb{R}^n$ and the set of all clusters is called a partition, which will be denoted by $\Pi = \{\pi_1, \ldots, \pi_k\}$. The collection of all such partitions will be denoted by $\mathcal{C}(\mathcal{A}, k)$.

If components a^i_s, $s = 1, \ldots, n$ of the data point a^i lie in intervals $[\alpha_i, \beta_i]$ which are not of equal range, i.e., if numbers $\beta_1 - \alpha_1, \ldots, \beta_n - \alpha_n$, are mutually significantly different, they should first be normalized [13]. This can be achieved by transforming

the set $\mathcal{A}$ into the set $\mathcal{B} = \{b^i = T(a^i)\colon a^i \in \mathcal{A}\} \subset [0,1]^n$ by using the mapping $T\colon [\alpha,\beta] \to [0,1]^n$, where

$$T(x) = D(x-\alpha), \qquad D = \operatorname{diag}\left(\tfrac{1}{\beta_1-\alpha_1}, \ldots, \tfrac{1}{\beta_n-\alpha_n}\right). \tag{3}$$

After clustering the set $\mathcal{B}$, the obtained results will be transformed again into $[\alpha,\beta]$ by the inverse mapping $T^{-1}\colon [0,1]^n \to [\alpha,\beta]$, $T^{-1}(x) = D^{-1}x + \alpha$.

If we introduce some distance-like function (see e.g. [1]) $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $\mathbb{R}_+ := [0,+\infty\rangle$, then to each cluster $\pi_j \in \Pi$ we can associate its center c_j defined by

$$c_j = \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{a\in\pi_j} d(x,a). \tag{4}$$

After that, a globally optimal k-partition $\Pi^\star \in \mathcal{C}(\mathcal{A},k)$ can be defined as a solution of the following global optimization problem

$$\Pi^\star = \operatorname*{argmin}_{\Pi\in\mathcal{C}(\mathcal{A},k)} \mathcal{F}(\Pi), \qquad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} d(c_j,a), \tag{5}$$

where $\mathcal{F}\colon \mathcal{C}(\mathcal{A},k) \to \mathbb{R}_+$ is the objective function (see e.g. [13, 16]).

3.1. Adaptive Mahalanobis clustering

Given the structure of the data set in this paper, the set $\mathcal{A}$ will be grouped into ellipsoidal clusters. An efficient algorithm for searching for a locally optimal partition with ellipsoidal clusters is the *Adaptive Mahalanobis k-means* (see [13]), which can be carried out as a special case of the well-known Expectation Maximization algorithm (see [24]), but its efficiency is significantly greater than the standard Expectation Maximization algorithm. The adaptive Mahalanobis k-means algorithm can be described by two steps which are iteratively repeated:

Step A: Based on the set of mutually different assignment points $c_1, \ldots, c_k \in \mathbb{R}^n$, the set $\mathcal{A}$ should be divided into k disjoint clusters $\pi_1, \ldots, \pi_k$ by using the minimum distance principle

$$\pi_j = \{a \in \mathcal{A}\colon d_M^j(c_j,a;S_j) \le d_M^s(c_s,a;S_s),\ \forall s \in J\}, \quad j \in J,$$

where

$$d_M^j(x,y;S_j) = \sqrt[n]{\det S_j}\,(x-y)^T S_j^{-1}(x-y), \tag{6}$$

is the adaptive Mahalanobis distance-like function, and

$$S_j = \tfrac{1}{|\pi_j|} \sum_{a^i\in\pi_j} (c_j - a^i)(c_j - a^i)^T, \tag{7}$$

is a *covariance matrix* (see e.g. [1], [13, 20, 21]);

Step B: For each cluster of the partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$, one can define the corresponding cluster centers

$$c_j = \frac{1}{|\pi_j|} \sum_{a^i \in \pi_j} a^i. \tag{8}$$

Remark 1. *According to [20] the covariance matrix S_j is positive definite if and only if the set of vectors*

$$\varphi_s = (a_s^1 - c_j^s, \ldots, a_s^{|\pi_j|} - c_j^s)^T \in \mathbb{R}^{|\pi_j|}, \quad s = 1, \ldots, n,$$

is linearly independent. The matrix S_j can become singular in some cases mentioned in [20]. That problem can than be solved by taking $S = I$ (identity matrix) or by introducing the small perturbation of the component of all data points in π_j. For more details see [20].

Searching for a globally optimal partition $\Pi^\star$ is a complex global optimization problem for the solution of which there is generally no effective method. An efficient incremental partitioning algorithm is proposed in the paper written by [13], which is able to find either a globally optimal partition or a locally optimal partition of the set $\mathcal{A} \subset \mathbb{R}^n$ close to the global one. By knowing an optimal r-partition ($r \geq 1$), the algorithm searches for the following additional cluster by using the well-known `DIRECT` algorithm for global optimization [6, 5, 7], and after that by using the adaptive Mahalanobis k-means algorithm it determines the optimal $(r+1)$-partition.

This algorithm successively gives optimal partitions (consisting of elliptical shape clusters that are as
compact and relatively strongly separated as possible) for $k = 2, \ldots, k_{max}$, where k_{max} is the maximum number of clusters that makes sense to be calculated. Therefore, this algorithm is also very suitable for searching for a partition with the most appropriate number of clusters by using some known indexes (see Section 3.2).

3.2. Choosing of a partition with the most appropriate number of clusters

In some cases, the number of clusters k is determined by the nature of the problem itself and therefore it is known in advance. If the number of clusters is not known in advance, then it is natural to search for an optimal partition which consists of clusters that are as compact and relatively strongly separated as possible. This can be done by using some of the well-known validity indexes (see e.g. [13, 23]). In our paper, we will use the Calinski-Harabasz (CH) index and the Davies-Bouldin (DB) index. More compact and better separated clusters in an optimal partition will result in a greater CH index and a smaller DB index, respectively.

4. Project clustering and ranking

The given set of positively assessed projects $\mathcal{P}$ should be grouped into $k \geq 1$ as compact and well-separated clusters as possible. The very nature of the data implies the need for searching for ellipsoidal clusters by applying the AMC algorithm described in Section 3.1.

4.1. Project clustering and ranking based upon the assessed project features

For each project $p^i \in \mathcal{P}$, first the vector $\bar{f}^i = (\bar{f}^i_1, \ldots, \bar{f}^i_n)^T \in [1,5]^n$ of the GPA and the vector $a^i = (w_1\bar{f}^i_1, \ldots, w_n\bar{f}^i_n)^T \in \mathbb{R}^n$ of the weighted GPA (WGPA) of features (1) as well as the set

$$\mathcal{A} = \{a^i = (w_1\bar{f}^i_1, \ldots, w_n\bar{f}^i_n)^T \in \mathbb{R}^n \colon i = 1, \ldots, m\} \subset [\alpha, \beta], \tag{9}$$

$$\alpha = (w_1, \ldots, w_n)^T, \quad \beta = 5(w_1, \ldots, w_n)^T, \tag{10}$$

should be defined (see Fig. 1a).

Since there is a bijection between the set $\mathcal{P}$ of all projects and the set $\mathcal{A}$, in order to group projects into groups by their quality, we will find an optimal partition of the set $\mathcal{A}$ (see Section 3.1) with the most appropriate number of clusters (see Section 3.2).

In order to ensure the same influence of such weighted grades, the data points should first be normalized by using the mapping $T\colon [\alpha, \beta] \to [0,1]^n$ given by (3), where

$$T(x) = Dx - \tfrac{1}{4}e, \quad D = \tfrac{1}{4}\operatorname{diag}\left(\tfrac{1}{w_1}, \ldots, \tfrac{1}{w_n}\right), \quad e = (1, \ldots, 1)^T \in \mathbb{R}^n. \tag{11}$$

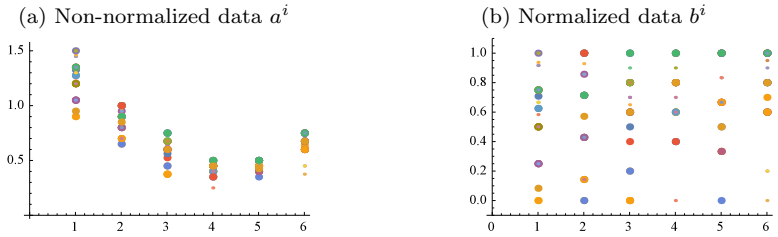

Figure 1: *Non-normalized data and normalized data*

This yields a normalized set $\mathcal{B} = \{b^i = T(a^i) \colon a^i \in \mathcal{A}\} \subset [0,1]^n$ (see Fig. 1b). Applying the AMC algorithm described in Section 3.1, by using validity indexes mentioned in Section 3.2 we obtain an optimal partition $\Pi^\star = \{\pi^\star_1, \ldots, \pi^\star_k\}$ with the most appropriate number of clusters $\pi^\star_1, \ldots, \pi^\star_k$ with centers $\zeta^\star_1, \ldots, \zeta^\star_k$. After clustering the set $\mathcal{B}$, the obtained results can be transformed again into $[\alpha, \beta]$ by the inverse mapping $T^{-1}\colon [0,1]^n \to [\alpha, \beta]$.

Project ranking will be carried out based upon measuring the "weighted Euclidean distance" to the perfectly assessed project $p^\star$, i.e., the project which the vector $f^\star = (5, \ldots, 5)^T \in \mathbb{R}^n$ is associated to. In this way, we will achieve a fine ranking structure in which all GPAs achieved as well as their weights will be taken into account. In this regard, we introduce the following definition.

Definition 1. *Let $\bar{f}^i = (\bar{f}^i_1, \ldots, \bar{f}^i_n)^T$ be a vector of the GPA of features of the project p^i and let $f^\star = (5, \ldots, 5)^T$ be a vector associated to the perfectly assessed project $p^\star$. The* quality measure of the project $p^i \in \mathcal{P}$ *is the weighted Euclidean*

distance $d(p^i, p^\star)$ *defined by*

$$d^2(p^i, p^\star) = \|\bar{f}^i - f^\star\|_w^2 = \sum_{s=1}^{n} w_s(\bar{f}_s^i - 5)^2. \tag{12}$$

Remark 2. *If the quality measure of the project* p^r *is less than the quality measure of the project* p^s*, i.e., if* $d(p^r, p^\star) < d(p^s, p^\star)$*, then the project* p^r *is "closer" to the perfectly assessed project* $p^\star$ *and it will be ranked higher than the project* p^s*.*

Furthermore, note that the quality measure of the project given by (12) *can be expressed by using normalized data. Specifically, by using* (11) *we obtain*

$$b^i = T(a^i) = D(a^i - \alpha) = \tfrac{1}{4}(\bar{f}^i - b^\star),$$

where $b^\star = (1, \ldots, 1)^T \in \mathbb{R}^n$ *(a normalized representant of the perfectly assessed project). Hence*

$$d(p^i, b^\star) := \|b^i - b^\star\|_w = \tfrac{1}{4}\|\bar{f}^i - f^\star\|_w = \tfrac{1}{4}d(p^i, p^\star). \tag{13}$$

This means that the same quality measure of the project can also be obtained such that we measure the weighted Euclidean distances of normalized data b^i *to the vector* $b^\star$*, what will be used below.*

Specially, the ranking of clusters within a partition can also be performed by measuring the weighted Euclidean distances of their centers to the vector of the perfectly assessed project. In this regard, the *quality measure* of the cluster $\pi_j^\star$ with the centers $\zeta_j^\star$ is defined as

$$d(\pi_j^\star, b^\star) := \|\zeta_j^\star - b^\star\|_w. \tag{14}$$

Example 2. *The set of all positively assessed projects* $\mathcal{P}$ *from Example 1 contains* $m = 47$ *projects. These projects should be grouped on the basis of the WGPA,* $w_1\bar{f}_1^i, \ldots, w_6\bar{f}_6^i$ *of these projects obtained based upon reviews by independent, blinded reviewers* $R_{\#1}^i$ *and* $R_{\#2}^i$ *and weights of features* $w_1, \ldots, w_n$ *that can be seen in Table 1.*

After defining the set $\mathcal{A} = \{a^i = (w_1\bar{f}_1^i, \ldots, w_6\bar{f}_6^i)^T \in \mathbb{R}^6 \colon i = 1, \ldots, m\}$, on the corresponding normalized set $\mathcal{B} = \{b^i = T(a^i) \colon a^i \in \mathcal{A}\} \subset [0,1]^6$ the AMC algorithm is carried out as described in Section 3.1. By using indexes specified in Section 3.2 it was shown that an optimal partition with the most appropriate number of clusters has four clusters. The obtained results were transformed again into $[\alpha, \beta]$. Characteristics of the optimal cluster partition (the number of projects per cluster $|\pi_j^\star|$, the standard deviation of the cluster $\sigma_j^\star$, the cluster center $c_j^\star = T^{-1}(\zeta_j^\star)$ and the quality measure $d(\pi_j^\star, b^\star)$ of the cluster) are given in Table 2.

Note that the quality measure $d(\pi_1^\star, b^\star)$ of the cluster of projects assessed as best is significantly lower than the quality measures of other clusters, whereby there is an insignificant difference between standard deviations by clusters (see Table 2). This means that the cluster of projects $\pi_1^\star$ assessed as best is significantly separated from other clusters.

Cluster	$\lvert\pi_j^\star\rvert$	$\sigma_j^\star$	$c_j^\star = T^{-1}(\zeta_j^\star)$						$d(\pi_j^\star, b^\star)$
$\pi_1^\star$	17	0.69	(4.527,	4.801,	4.713,	4.779,	4.956,	4.846)	0.077
$\pi_2^\star$	11	0.98	(4.432,	4.159,	4.182,	3.977,	4.364,	4.591)	0.178
$\pi_3^\star$	10	1.26	(3.908,	4.725,	3.95,	4.25,	4.65,	3.75)	0.229
$\pi_4^\star$	9	1.08	(3.519,	3.722,	3.5,	4.167,	4.528,	4.5)	0.301

Table 2: *Properties of clusters of the optimal partition*

The cluster $\pi_1^\star$ which consists of projects assessed as best contains 17 projects ranked according to the achieved quality measures $d(p^i, b^\star)$ as defined by (13) (see Table 3). GPA of all features of these projects, average weighted grades (AWG) of all features $\hat{f}^i = \sum_{s=1}^{n} w_s \bar{f}_s^i$, the GPA of the general impression $\overline{F}^i$ and the quality measures $d(p^i, b^\star)$ of these projects are also given in Table 3.

p^i	$\bar{f}_1^i$	$\bar{f}_2^i$	$\bar{f}_3^i$	$\bar{f}_4^i$	$\bar{f}_5^i$	$\bar{f}_6^i$	$\hat{f}^i$	F^i	$d(p^i, b^\star)$	Rank
p^{35}	5	5	5	5	5	5	5	10	0	1
p^{40}	5	5	5	5	5	5	5	9.5	0	2
p^6	4.833	4.75	5	5	5	4.875	4.881	9.75	0.038	3
p^{36}	5	5	4.5	4.5	5	5	4.875	9.5	0.062	4
p^{10}	4.5	4.75	5	4.75	5	5	4.775	10	0.076	5
p^1	4.417	5	5	5	5	5	4.825	9.25	0.080	6
p^{33}	4.5	5	4.5	5	5	5	4.775	10	0.084	7
p^7	4.5	5	5	5	5	4.5	4.775	9.5	0.084	8
p^8	5	4.5	5	4.5	5	4.5	4.775	9.5	0.084	9
p^{26}	4.5	4.5	5	5	5	5	4.750	9.5	0.088	10
p^{45}	4.5	4.5	5	5	5	5	4.750	8.5	0.088	11
p^{28}	4.875	4.875	4.125	4.5	5	5	4.756	9.875	0.096	12
p^9	4.5	4.75	4	4.5	5	5	4.600	9	0.128	13
p^{30}	4.333	5	4.75	4.5	5	4	4.563	8.75	0.141	14
p^3	4	5	4.5	4.5	5	5	4.575	9	0.151	15
p^{24}	4	4.5	4.25	4.5	4.75	4.5	4.337	8.5	0.177	16
p^{19}	3.5	4.5	4.5	5	4.5	5	4.325	9	0.222	17

Table 3: *Cluster of projects $\pi_1^\star$ assessed as best*

Theoretically, it may happen that for some project $p^r \in \pi_1$ and for some project $p^s \in \pi_2$ holds $d(p^r, b^\star) > d(p^s, b^\star)$, but application of ellipsoidal clusters reduces such possibility significantly.

Note that the ranking of projects in Table 3 does not follow the ranking of these projects by the AWG of all features (see column $\hat{f}^i$ in the table). For example, project p^1 has a higher average score than project p^{10}, but it is still ranked lower. The reason for that lies in its relatively low grade given to feature $\bar{f}_1^1$ of project p^1, which is much more important than other features ($w_1 = 0.30$).

Thus, the proposed method accepts better a fine structure of project feature ratings than the ordinary ranking obtained on the basis of the AWG of all features.

Since the well-known AHP method is in the background of ranking projects according to the AWG of all features (see e.g. [3, 9, 14]), in Table 3 it is possible to recognize advantages of the method we propose in relation to the AHP method.

4.2. Project clustering and ranking based upon assessed features and the general impression of the projects

Similarly to the previous section, for every project $p^i \in \mathcal{P}$ define $(n+1)$-dimensional vector $(\bar{f}_1^i, \ldots, \bar{f}_n^i, \overline{F}^i)^T \in [1,5]^n \times [1,10]$. The first n components of this vector are the GPAs of features, and the last component represents a GPA of the general impression of the project. A set of data $\mathcal{A}$ will be defined by means of these grades such that the grade of the general impression of the project $\overline{F}^i$ has the same impact as all weighted features $w_1\bar{f}_1^i, \ldots, w_n\bar{f}_n^i$ together. This means that we have the following set of data

$$\mathcal{A} = \{a^i = (w_1\bar{f}_1^i, \ldots, w_n\bar{f}_n^i, \overline{F}^i) \in \mathbb{R}^{n+1} \colon i = 1, \ldots, m\} \subset [\alpha, \beta], \tag{15}$$

$$\alpha = (w_1, \ldots, w_n, 1)^T, \quad \beta = (5w_1, \ldots, 5w_n, 10)^T. \tag{16}$$

Please note that the ratio of the impact of the general impression of the project and the weighted features $w_1\bar{f}_1^i, \ldots, w_n\bar{f}_n^i$ could also be defined in a different way.

Since there is a bijection between the set $\mathcal{P}$ of projects and the set $\mathcal{A}$, in order to group projects into groups by quality, we will find an optimal partition of the set $\mathcal{A}$ (see Section 3.1) with the most appropriate number of clusters (see Section 3.2).

In order to ensure the same impact of grades weighted in this way, the data points should first be normalized by using the mapping $T \colon [\alpha, \beta] \to [0,1]^{n+1}$ given by (3), where

$$T(x) = D(x - \alpha), \quad D = \operatorname{diag}\left(\tfrac{1}{4w_1}, \ldots, \tfrac{1}{4w_n}, \tfrac{1}{9}\right), \ \alpha = (w_1, \ldots, w_n, 1)^T \in \mathbb{R}^n. \tag{17}$$

Thus, this yields the set $\mathcal{B} = \{b^i = T(a^i) \colon a^i \in \mathcal{A}\} \subset [0,1]^{n+1}$ on which we applied the AMC algorithm described in Section 3.1 by using indexes given in Section 3.2 and obtained the optimal partition $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ with the most appropriate number of clusters $\pi_1^\star, \ldots, \pi_k^\star$ with centers $\zeta_1^\star, \ldots, \zeta_k^\star$. After clustering the set $\mathcal{B}$, the obtained results can be transformed again into $[\alpha, \beta]$ by the inverse mapping $T^{-1} \colon [0,1]^n \to [\alpha, \beta]$.

In accordance with Remark 2, project ranking can be carried out on the basis of the quality measures of the projects defined by (13).

Example 3. *The set $\mathcal{P}$ of $m = 47$ projects from Example 1 will be grouped equally, on the basis of the WGPA of 6 project features $w_1\bar{f}_1^i, \ldots, w_6\bar{f}_6^i$ and on the basis of the GPA of the general impression $\overline{F}^i$, which were obtained based upon reviews by independent, blinded reviewers $R^i_{\#1}$ and $R^i_{\#2}$.*

After defining the set $\mathcal{A} = \{a^i = (w_1\bar{f}_1^i, \ldots, w_6\bar{f}_6^i, \overline{F}^i) \in \mathbb{R}^7 \colon i = 1, \ldots, m\}$, on the corresponding normalized set $\mathcal{B} = \{b^i = T(a^i) \colon a^i \in \mathcal{A}\} \subset [0,1]^7$ the AMC algorithm is carried out as described in Section 3.1. By using indexes specified

in Section 3.2, it was shown that an optimal partition with the most appropriate number of clusters has four clusters. The obtained results were transformed into $[\alpha, \beta]$. Characteristics of the optimal cluster partition (the number of projects per cluster $|\pi_j^\star|$, the standard deviation of the cluster $\sigma_j^\star$, the cluster center $c_j^\star = T^{-1}(\zeta_j^\star)$ and the quality measure $d(\pi_j^\star, b^\star)$ of the cluster defined by (14)) are given in Table 4.

Cluster	$\lvert\pi_j^\star\rvert$	$\sigma_j^\star$	$c_j^\star = T^{-1}(\zeta_j^\star)$							$d(\pi_j^\star, b^\star)$
$\pi_1^\star$	18	0.37	(4.512	4.785	4.674	4.736	4.958	4.797	9.326)	0.250
$\pi_2^\star$	8	0.51	(4.187	4.656	4.031	4.094	4.187	4.250	8.347)	0.599
$\pi_3^\star$	3	0.34	(3.778	4.917	4.833	4.500	4.833	2.833	9.000)	0.559
$\pi_4^\star$	18	0.58	(3.912	3.889	3.681	4.083	4.569	4.556	7.982)	0.755

Table 4: *Properties of the optimal partition*

Note that the cluster $\pi_1^\star$ which consists of projects assessed as best is ranked significantly higher than other clusters, whereby there is an insignificant difference between standard deviations by clusters (see Table 4). This means that the cluster of projects $\pi_1^\star$ assessed as best is significantly separated from other clusters.

p^i	$\bar{f}_1^i$	$\bar{f}_2^i$	$\bar{f}_3^i$	$\bar{f}_4^i$	$\bar{f}_5^i$	$\bar{f}_6^i$	$\hat{f}^i$	$\overline{F}^i$	$\|b^i - b^\star\|$	Rank
p^{35}	5	5	5	5	5	5	5	10	0	1
p^6	4.833	4.75	5	5	5	4.875	4.881	9.75	0.047	2
p^{40}	5	5	5	5	5	5	5	9.5	0.055	3
p^{10}	4.5	4.75	5	4.75	5	5	4.775	10	0.076	4
p^{36}	5	5	4.5	4.5	5	5	4.875	9.5	0.0829	5
p^{33}	4.5	5	4.5	5	5	5	4.775	10	0.0834	6
p^{28}	4.875	4.875	4.125	4.5	5	5	4.756	9.875	0.096	7
p^8	5	4.5	5	4.5	5	4.5	4.775	9.5	0.1000	8
p^7	4.5	5	5	5	5	4.5	4.775	9.5	0.1001	9
p^{26}	4.5	4.5	5	5	5	5	4.750	9.5	0.104	10
p^1	4.417	5	5	5	5	5	4.825	9.25	0.115	11
p^9	4.5	4.75	4	4.5	5	5	4.600	9	0.169	12
p^3	4	5	4.5	4.5	5	5	4.575	9	0.187	13
p^{45}	4.5	4.5	5	5	5	5	4.750	8.5	0.188	14
p^{30}	4.333	5	4.75	4.5	5	4	4.563	8.75	0.197	15
p^{37}	4	4.5	4.25	4.5	4.75	4.5	4.275	8.5	0.240	16
p^{24}	4	4.5	4.25	4.5	4.75	4.5	4.337	8.5	0.243	17
p^{19}	3.5	4.5	4.5	5	4.5	5	4.325	9	0.248	18

Table 5: *Cluster of projects $\pi_1^\star$ assessed as best*

The cluster of projects $\pi_1^\star$ assessed as best in this case contains 18 projects. It is interesting to notice that these are all projects selected as best in the previous section (Example 2), but their order is modified under the influence of grades referring to the general impression of projects. For the very same reason, project p^{37} became part of the cluster projects assessed as best (data referring to this project can be seen in Table 5).

The ranking of projects in the cluster of projects $\pi_1^\star$ assessed as best is determined on the basis of the quality measures $d(p^i, b^\star)$ of the projects defined by (13) in accordance with Remark 2 and shown in Table 5. Since in this approach the impact of the GPA of the general impression of the project and the WGPA of all features is assumed to be equal, the quality measures of projects is primarily determined by the grades $\overline{F}^i$, but there are also fine corrections. For example, project p^6 was rated higher than project p^{40} due to a higher general impression grade, but project p^{19} is ranked lower than projects $p^{45}, p^{30}, p^{37}, p^{24}$, although it has a higher general impression grade. The reason for that lies in a relatively low grade given to feature f^1 of these projects, which is much more important than other features ($w_1 = 0.30$).

4.3. A possibility of simplification

A set of data (15) can have a great number of components, from which serious numerical problems in the implementation of data clustering may arise. Namely, in this case, there is a high possibility of the singularity of the covariance matrix (see [20]). That is why it makes sense to observe the following simplification: instead of the vector $(\bar{f}_1^i, \ldots, \bar{f}_n^i, \overline{F}^i)^T$ we could observe the vector $(\hat{f}^i, F^i)^T$, where $\hat{f}^i = \sum_{s=1}^{n} w_s \bar{f}_s^i$ are the AWG of all features of the i-th project. In this way, the problem would be reduced to the problem of grouping data with two features considered equally, i.e., we consider the set

$$\mathcal{A} = \{a^i = (\hat{f}^i, \overline{F}^i) \in \mathbb{R}^2 \colon \hat{f}^i = \sum_{s=1}^{n} w_s f_s^i, \quad i = 1, \ldots, m\}. \tag{18}$$

In this approach we should be aware of the fact that we have lost a fine structure of grades, but obtained a simpler set of data that can also be displayed graphically (see Fig. 2a).

Due to a disproportionate range of numbers $\hat{f}^i$ and $\overline{F}^i$, when grouping the set $\mathcal{A}$, the general impression would be preferred to the AWG of all features. In order to eliminate this discrepancy, in accordance with Section 3, the set of data $\mathcal{A}$ should be first transformed into the set $\mathcal{B} = \{b^i = T(a^i) \colon a^i \in \mathcal{A}\} \subset [0,1]^2$ in the unit square $[0,1]^2$. In this case, $T \colon [1,5] \times [1,10] \to [0,1]^2$,

$$T(x) = D(x - \alpha), \qquad D = \operatorname{diag}\left(\tfrac{1}{4}, \tfrac{1}{9}\right), \quad \alpha = (1,1)^T. \tag{19}$$

Thus we ensure a balanced simultaneous impact of the AWG of all features and the GPA of the general impression of the project.

Applying the AMC algorithm (see Section 3.1), by using indexes specified in Section 3.2 we obtain the optimal partition $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ with the most appropriate number of clusters $\pi_1^\star, \ldots, \pi_k^\star$ with centers $\zeta_1^\star, \ldots, \zeta_k^\star$. After clustering the set $\mathcal{B}$, the obtained results can be transformed again into $[1,5] \times [1,10]$ by the inverse mapping T^{-1}.

In accordance with Remark 2, in this case project ranking can also be carried out by comparing the quality measures $d(p^i, b^\star)$ of respective projects.

Example 4. *The set $\mathcal{P}$ of $m = 47$ projects from Example 1 will be grouped on the basis of the AWG of all project features $\hat{f}^i$ and the GPA of the general impression $\overline{F}^i$ obtained on the basis of reviews by independent, blinded reviewers $R^i_{\#1}$ and $R^i_{\#2}$.*

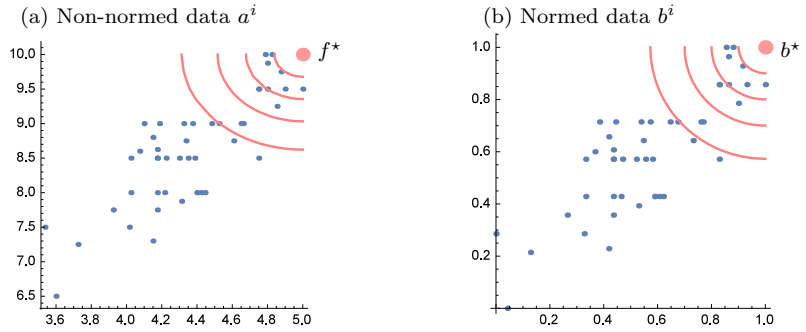

Figure 2: *Distances of non-normed data a^i to the vector $f^\star$ (ellipses) and distances of normed data b^i to the vector $b^\star$ (circles)*

After defining the corresponding set of data $\mathcal{A} = \{a^i = (\hat{f}^i, F^i) \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$, first the corresponding normalized set $\mathcal{B} = \{b^i = T(a^i) \colon a^i \in \mathcal{A}\}$ is defined by means of mapping (19). The set $\mathcal{A}$ is shown in Fig. 2a, and the corresponding set of normalized data is given in $\mathcal{B}$ in Fig. 2b. The AMC algorithm is applied to the set $\mathcal{B}$, as described in Section 3.1 (see also Fig. 3b). After clustering the set $\mathcal{B}$, the obtained results will be transformed again into $[1, 5] \times [1, 10]$.

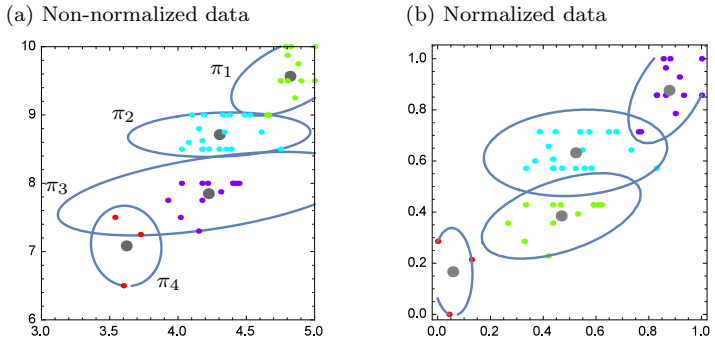

Figure 3: *Clusters of non-normalized and normalized data*

It was shown that the optimal partition with the most appropriate number of clusters has four clusters, too. Characteristics of the optimal cluster partition (the number of data per cluster $|\pi^\star_j|$, the standard deviation of the cluster $\sigma^\star_j$, the cluster center $c^\star_j = T^{-1}(\zeta^\star_j)$ and the quality measure $d(\pi^\star_j, b^\star)$ of the cluster) are given in Table 6.

The cluster of projects $\pi^\star_1$ assessed as best in this case contains 13 projects (see Table 7 and Fig. 3). It is interesting to notice that these are all projects selected as best in the previous sections, but projects $p^{19}, p^{24}, p^{30}, p^{37}$ and p^{45} are missing.

Cluster	$\lvert\pi_j\rvert$	$\sigma_j^\star$	$c_j^\star = T^{-1}(\zeta_j^\star)$		$d(\pi_j^\star, b^\star)$
π_1	13	0.21	(4.82	9.57)	0.066
π_2	19	0.62	(4.30	8.71)	0.226
π_3	12	0.82	(4.22	7.85)	0.308
π_4	3	1.27	(3.62	7.08)	0.473

Table 6: *Properties of clusters of the optimal partition*

Ranking clusters within the partition was carried out on the basis of the quality measures of the clusters $d(\pi_j^\star, b^\star)$, as defined in (14). The order of projects in the cluster of projects $\pi_1^\star$ assessed as best is determined based upon the quality measures $d(p^i, b^\star)$ of the projects defined by (13).

p^i	$\hat{f}^i$	F^i	$d(p^i, b^\star)$	Rank
p^{35}	5	10	0.	1
p^6	4.877	9.75	0.0209	2
p^{33}	4.825	10	0.0239	3
p^{40}	5	9.5	0.0248	4
p^{28}	4.8	9.875	0.02808	5
p^{36}	4.9	9.5	0.02837	6
p^{10}	4.7875	10	0.0291	7
p^3	4.65	9	0.0690	8
p^7	4.8	9.5	0.0369	9
p^1	4.8542	9.25	0.0422	10
p^{26}	4.75	9.5	0.0423	11
p^8	4.75	9.5	0.0423	12
p^9	4.6625	9	0.0678	13

Table 7: *Cluster of projects assessed as best*

Please note that this ranking is not the same any more as it was in previous sections, and it is shown in Table 7. Note also that the cluster of projects $\pi_1^\star$ assessed as best is ranked significantly higher than other clusters, whereby there is an insignificant difference between standard deviations by clusters (see Table 6). This means that the cluster of projects $\pi_1^\star$ assessed as best is significantly separated from other clusters.

A balanced simultaneous impact of the AWG of all features and the GPA of the general impression of the project determined the project ranking list.

5. Conclusions

The problem of a fair, equitable and transparent selection of research projects to be financed from a fund is important for both the institution that allocates financial resources and researchers, i.e., potential users. To tackle this problem, numerous approaches can be found in the literature, which are most often based on the AHP method. The combination of the AHP method and the AMC algorithm proved to

be a very reasonable approach to solve this problem because the proposed method for the formation of the cluster of projects assessed as best optimally connects all features of the data set, i.e., grades obtained in the review procedure. Note that the well-known Expectation Maximization Algorithm lies in the background of the method [22, 24].

The quality measure of projects in the cluster of projects assessed as best is defined such that it takes into account the weighted structure of grades obtained in the review procedure.

Based upon this grouping and ranking of positively assessed projects from Example 1, the obtained results were presented to University of Osijek constituent units and a list of projects to be financed was published on the University of Osijek website. It was observed that the reactions of applicants in the call for project proposals to this transparent and clear assessment process are generally very positive. In this way, we have maximally avoided possible objections and dissatisfaction of applicants whose project proposals were not selected for funding.

Acknowledgement

This work was supported by the Ministry of Science, Education and Sports, Republic of Croatia, through research grant 235-2352818-1034. The authors would like to thank the administration of Josip Juraj Strossmayer University of Osijek for preparing the data for this paper.

References

[1] Bezdek, J. C., Keller, J., Krisnapuram, R. and Pal, N. R. (2005). Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer.

[2] Cocchi, D., Cavaliere, G., Freo, M., Giannerini, S., Mazzocchi, M., Trivisano, C. and Viroli, C. (2014). A support for classifying scientific papers in a university department. Procedia Economics and Finance, 17, 47–54.

[3] Collan, M., Fedrizzi, M. and Luukka, P. (2013). A multi-expert system for ranking patents: An approach based on fuzzy pay-off distributions and a topsisahp framework. Expert Systems with Applications, 40, 4749–4759.

[4] Daraio, C., Bonaccorsi, A. and Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. Journal of Operational Research, 244, 918–930.

[5] Finkel, D. E. (2003). DIRECT Optimization Algorithm User Guide. Center for Research in Scientific Computation. North Carolina State University.

[6] Grbić, R., Nyarko, E. K. and Scitovski, R. (2013). A modification of the DIRECT method for Lipschitz global optimization for a symmetric function. Journal of Global Optimization, 57, 1193–1212.

[7] Jones, D. R., Perttunen, C. D. and Stuckman, B. E. (1993). Lipschitzian optimization without the Lipschitz constant. Journal of Optimization Theory and Applications, 79, 157–181.

[8] Kadziski, M. and Sowiski, R. (2015). Parametric evaluation of research units with respect to reference profiles. Decision Support Systems, 72, 33–43.

[9] Mandic, D., Jovanovic, P. and Bugarinovic, M. (2014). Two-phase model for multicriteria project ranking: Serbian railways case study. Transport Policy, 36, 88–104.
[10] Manouselis, N. and Verbert, K. (2013). Layered evaluation of multi-criteria collaborative filtering for scientific paper recommendation. Procedia Computer Science, 18, 1189–1197.
[11] Mardani, A., Jusoh, A. and Zavadskas, E. K. (2015). Fuzzy multiple criteria decision-making techniques and applications two decades review from 1994 to 2014. Expert Systems with Applications, 42, 4126–4148.
[12] Mlek, J., Hudekov, V. and Matjka, M. (2014). System of evaluation of research institutions in the Czech Republic. Procedia Computer Science, 33, 315–320.
[13] Morales-Esteban, A., Martynez-Alvarez, F., Scitovski, S. and Scitovski, R. (2014). A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic zoning. Computers & Geosciences, 73, 132–141.
[14] Rad, A., Naderi, B. and Soltani, M. (2011). Clustering and ranking university majors using data mining and ahp algorithms: A case study in Iran. Expert Systems with Applications, 38, 755–763.
[15] Saaty, T.L., 1980. The Analytic Hierarchy Process. Mc-Graw Hill.
[16] Sabo, K. and Scitovski, R. (2015). An approach to cluster separability in a partition. Information Sciences, 305, 208–218.
[17] Sabo, K., Scitovski, R., Vazler, I. and Zekić-Sušac, M. (2011). Mathematical models of natural gas consumption. Energy Conversion and Management 52, 1721–1727.
[18] dos Santos, T. R. and Zrate, L. E. (2015). Categorical data clustering: What similarity measure to recommend? Expert Systems with Applications, 42, 1247–1260.
[19] Scitovski, R. and Sabo, K. (2014). Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters. Knowledge-Based Systems 57, 1-7.
[20] Scitovski, S. and Šarlija, N. (2014). Cluster analysis in retail segmentation for credit scoring. Croatian Operational Research Review, 5, 235–245.
[21] Spath, H. (1983). Cluster-Formation und Analyse. R. Oldenburg Verlag, Munchen.
[22] Theodoridis, S. and Koutroumbas, K. (2009). Pattern Recognition. Academic Press, Burlington. 4th edition.
[23] Vendramin, L., Campello, R. J. G. B., Hruschka, E. R. (2009). On the comparison of relative clustering validity criteria, in: Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 May 2, 2009, Sparks, Nevada, USA, SIAM. pp. 733–744.
[24] Younis, K. S. (1999). Weighted Mahalanobis distance for hyper-ellipsoidal clustering. Ph.D. thesis. Air Force Institute of Technology, Ohio.