Scientometrics (2014) 99:911–926
DOI 10.1007/s11192-014-1258-y

On the quality evaluation of scientific entities in Poland supported by consistency-driven pairwise comparisons method

Waldemar W. Koczkodaj · Konrad Kułakowski · Antoni Ligęza

Received: 15 August 2013 / Published online: 19 March 2014

Abstract Comparison, rating, and ranking of alternative solutions, in case of multicriteria evaluations, have been an eternal focus of operations research and optimization theory. There exist numerous approaches at practical solving the multicriteria ranking problem. The recent focus of interest in this domain was the event of parametric evaluation of research entities in Poland. The principal methodology was based on pairwise comparisons. For each single comparison, four criteria have been used. One of the controversial points of the assumed approach was that the weights of these criteria were arbitrary. The main focus of this study is to put forward a theoretically justified way of extracting weights from the opinions of domain experts. Theoretical bases for the whole procedure are based on a survey and its experimental results. Discussion and comparison of the two resulting sets of weights and the computed inconsistency indicator are discussed.

Keywords Pairwise comparisons · Inconsistency analysis · Expert opinion · Academic entity quality · Performance evaluation

Introduction and problem statement

The question of how to measure the performance of scientific entities is one of the most basic in the scientific community. The answer to this question is primarily related to: 'how

The research is supported by Euro Grant "Human Capital" and AGH University of Science and Technology, contract no.: 11.11.120.859

W. W. Koczkodaj (✉)
Laurentian University, Sudbury, ON, Canada
e-mail: wkoczkodaj@cs.laurentian.ca

K. Kułakowski · A. Ligęza
AGH University of Science and Technology, Kraków, Poland
e-mail: kkulak@agh.edu.pl

A. Ligęza
e-mail: ligeza@agh.edu.pl

should research funds be distributed among different research units?' (addressed in Wang et al. 2011; Geuna and Martin 2003), or 'what should be the policy of the state in the promotion of science?' (see Geuna et al. 1999) to name a few. Due to the many factors that can affect the final assessment, the problem of finding clear and widely acceptable performance indicators is not easy. Numerous legal environments and various scientific practices in different countries add to the problem complexity.

In academia, we are better in evaluating our students than ourselves. However, the Ministry of Science and Higher Education stipulates that evaluation of the academic performance of a scientific unit[1] is conducted in Poland on the basis of the algorithm presented in the government regulation[2] with four major criteria for all scientific entities (Table 1).

Each criterion is subdivided into many sub-criteria that depend on the type of scientific entity. The ranking process seems to be easy, yet it is not. One of the important problems is to determine the significance of criteria $c_1, \ldots, c_4$. Relating c_i to c_j is both subjective and difficult due to the intangible and abstract nature of the criterion itself. In the adopted algorithm ("Research entities evaluation: The official procedure" Section), the criteria importance must be expressed as the real numbers. One of the ways allowing the subjective judgments to be transformed into the numerical values is the pairwise comparisons (PC) method. Therefore, to improve the algorithm proposed by the Ministry of Science and Higher Education ("Research entities evaluation: The official procedure" Section), the authors propose to add one additional step. The step in which the weights for criteria $c_1, \ldots, c_4$ are explicitly estimated by experts. The experiment conducted by the authors ("An experimental survey procedure" Section), the survey among the scientists, provides a sample of how the weights of the criteria $c_1, \ldots, c_4$ might look like, when they were determined by the PC method.

Preliminaries of the PC method

As indicated in "Introduction and problem statement" Section, the evaluation of research units and induction of the final linear ordering is based on four different criteria. These predefined criteria are shown in Table 1. Precise interpretation of these criteria is provided by the Ministry of Science and Higher Education (2012); here the intuitive understanding of them is sufficient.

Note that in view of Table 1 quality evaluation of research units is not only a Multicriteria Decision Problem (or, more precisely, Multicriteria Ranking Problem), but all the criteria are in fact of *qualitative nature*. Hence, the first step in the procedure consists of defining the transformation of non-measurable characteristics into a single numbers. This is done for each criterion of each research unit. For example, calculation of the value of criterion c_1 consists in summing up points assigned to a list of publications of the last four years published by the research workers of the unit. The procedure is presented in detail in Ministry of Science and Higher Education (2012). Now, the problem can be approached by pairwise comparisons.

[1] Following the official regulation, the authors understand that the scientific entity (sometimes also referred to as the scientific unit) means: a research unit within the university such as faculty or department, an independent research institute (national and international), and research units within The Polish Academy of Science.

[2] http://www.bip.nauka.gov.pl/_gAllery/19/31/19319/poz._877

Table 1 Comparison criteria for scientific entities

Code	Criterion name
c_1	Scientific and/or creative achievements
c_2	Scientific potentiality
c_3	Tangible benefits of the scientific activity
c_4	Intangible benefits of the scientific activity

Let us briefly review the roots and ideas of the approach. It is believed that, in 1785, Condorcet was the first researcher who used pairwise comparisons for improving voting results in Condercet (1785). However, it was Fechner who described the *PC* method in in 1860 (reprinted in Fechner 1966), but he did it only from the psychometric perspective. Thurstone not only described the PC method in Thurstone (1994), but for the first time proposed a solution based on statistical analysis. In his seminal work Saaty (1977), Saaty introduced a hierarchy, which is instrumental for practical applications, and eigenvalue-based inconsistency.

Regretfully, the proposal of Saaty constitutes only a global inconsistency indicator and, as such, could not localize the most inconsistent elements of the matrix. The first ever localizing inconsistency definition was proposed in Koczkodaj (1993). Both inconsistencies were recently analyzed in Bozóki and Rapcsak (2008).

There are several different ways for deriving weights in the pairwise comparisons method Crawford (1987); Kułakowski (2013). For the purpose of this paper, the authors adopted probably the second most popular geometric means based method. The Monte Carlo study presented in Herman and Koczkodaj (1996) provided evidence that, for small inconsistencies, both the geometric means solution (used in this study) and the eigenvector solution (as proposed by Saaty in 1977) are similar enough from the statistical point of view. In fact, eigenvector and geometric means solutions are identical for fully consistent matrices and the geometric means is slightly better (approx. 7 out of 10 wins) than the principal eigenvector solution.

The procedure usually begins (after an appropriate feasibility study and data gathering, which are not addressed here) with a listing of all possible criteria. In our case, the four criteria mentioned in "Preliminaries of the PC method" Section are used.

Let us consider n alternative items to be compared. Let m_{ij} expresses a relative preference of entity s_i over $s_j, i, j = 1, \ldots, n$. The pairwise comparison matrix M is defined as

$$M = \begin{bmatrix} 1 & m_{12} & \cdots & m_{1n} \\ \frac{1}{m_{12}} & 1 & \cdots & m_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{m_{1n}} & \frac{1}{m_{2n}} & \cdots & 1 \end{bmatrix} \tag{1}$$

A pairwise comparisons matrix M describing the relationship between n given alternative items is called *reciprocal* if $m_{ij} = \frac{1}{m_{ji}}$ for every $i, j = 1, \ldots, n$ (then automatically $m_{ii} = 1$ for every $i = 1, \ldots, n$). Let we say that $M = [m_{ij}] \in R^{n \times n}$ is a pairwise comparisons (PC) matrix if $m_{ij} > 0$ for all $i, j = 1, ..., n$. A PC matrix M is called *consistent* (or transitive) if $m_{ij} \cdot m_{jk} = m_{ik}$ for every $i, j, k = 1, \ldots, n$. Note that while every consistent matrix is reciprocal, the converse is false in general. Consistent matrices correspond to the ideal situation in which there are the exact values $s_1, \ldots, s_n$ for the entity. The elements of matrix M defined as quotients $m_{ij} = s_i/s_j$ form a consistent matrix. The vector $\mathbf{s} = [s_1, \ldots, s_n]$ is unique up to a multiplicative constant.

Every pairwise comparisons question has been answered by all respondents and many different pairwise comparisons matrices could be produced (one matrix for each expert involved in the survey). In our study, we produced the survey summary for all the partial results $M_1, \ldots, M_q$ (every $M_r = [m_{ij}^{(r)}]$ corresponding to responses given by the i'th expert) by synthesizing them into one summary PC matrix $\widehat{M} = [\widehat{m}_{ij}]$ following the geometric mean synthesizing function proposed in Aczél and Saaty (1983); Saaty (2008) where all the experts are equally important.[3] Hence, the resulting ratios have the following form:

$$\widehat{m}_{ij} = \left(\prod_{r=1}^{q} m_{ij}^{(r)} \right)^{1/q} \tag{2}$$

According to the geometric mean method used in this study the final ranking vector s is calculated as:

$$s = [\sigma^{-1} S_1, \ldots, \sigma^{-1} S_n] \tag{3}$$

where

$$S_i = \left(\prod_{j=1}^{n} \widehat{m}_{ij} \right)^{1/n} \quad \text{and} \quad \sigma = \sum_{r=1}^{n} S_r \tag{4}$$

In the formulas above S_i represents the rank of the i-th alternative before normalization, and σ^{-1} is the normalization coefficient so that all $s_i = \sigma^{-1} S_i$ for $i = 1, \ldots, n$, sum up to one.

Let us see how the pairwise comparisons method works in practice by providing the following simple numeric example in which four widely recognized scientists apply for some award for their research results. For simplicity let us assume that the assessment procedure is based on an review of a few major scientific achievements of each. During the evaluation panel two experts provide their assessments forming two matrices M_1 and M_2.

$$M_1 = \begin{bmatrix} 1 & 0.4 & 1.1 & 0.91 \\ 1/0.4 & 1 & 1.28 & 1.1 \\ 1/1.1 & 1/1.28 & 1 & 0.5 \\ 1/0.91 & 1/1.1 & 1/0.5 & 1 \end{bmatrix} \quad M_2 = \begin{bmatrix} 1 & 0.5 & 1.3 & 0.85 \\ 1/0.5 & 1 & 1.28 & 1.2 \\ 1/1.3 & 1/1.28 & 1 & 0.5 \\ 1/0.85 & 1/1.2 & 1/0.5 & 1 \end{bmatrix} \tag{5}$$

To determine the PC matrix, each expert had to perform six comparisons (ones corresponding to the values above the matrix diagonal). For example, by indicating that the achievements of the first candidate are 0.4 of the achievements of the second candidate $m_{1,2}^{(1)} = 0.4$ the first expert indicated that $m_{2,1}^{(1)} = 1/0.4 = 2.5$. Thus, the values below the diagonal in $M^{(1)}$ and $M^{(2)}$ are generated in an automatic way. The matrices $M^{(1)}$ are used $M^{(2)}$ to compute the matrix of collective results $\widehat{M}$ according to the formula (Eq. 2).

$$\widehat{M} = \begin{bmatrix} 1 & 0.447 & 1.196 & 0.879 \\ 2.236 & 1 & 1.28 & 1.149 \\ 0.836 & 0.781 & 1 & 0.5 \\ 1.137 & 0.87 & 2 & 1 \end{bmatrix} \tag{6}$$

[3] Please compare with Theorem 4 in Saaty (2008), where the importance weight assigned to every judge is identical.

The final assessment $\mathbf{s}$ is formed as the normalized geometric means of rows of $\widehat{M}$ and is $\mathbf{s} = [0.201, 0.327, 0.184, 0.288]^T$. Hence, the winner is the second scientist with the rank 0.327, then respectively the scientist number four, one and three.

Data inconsistency and how to deal with it

Observe that matrix $M^\star$ given by (5) is not consistent. For example $m^\star_{1,2} \cdot m^\star_{2,3} \neq m^\star_{1,3}$ as $0.4 \cdot 1.1 \neq 0.5$. The question arises what can one do about that?

When an $n \times n$ matrix M is not consistent the consistency index needs to be computed to determine the degree of inconsistency. One of the popular inconsistency index Saaty (1977) is defined as follows:

$$Ic(M) = \frac{\lambda_{max} - n}{n - 1} \tag{7}$$

where λ_{max} is the principal eigenvalue of M. It is commonly assumed that the matrix M is sufficiently consistent if $Ic(M) \leq 0.1$ Saaty (1977). In such a case the results calculated using e.g. the geometric means method are considered to be reliable.

Another approach, perceived as more restrictive inconsistency index comes from Koczkodaj (1993). It is defined as:

$$\mathscr{K}(M) = \max_{i,j,k \in \{1,\ldots,n\}} \left\{ \min \left\{ \left| 1 - \frac{m_{ij}}{m_{ik} m_{kj}} \right|, \left| 1 - \frac{m_{ik} m_{kj}}{m_{ij}} \right| \right\} \right\} \tag{8}$$

where $i, j, k = 1, \ldots, n$ and $i \neq j \wedge j \neq k \wedge i \neq k$. For sufficiently consistent matrices, it should not be too high.

When a matrix M is inconsistent (especially when the inconsistency is high), we must compute a consistent $n \times n$ PC matrix C which differs from the matrix M 'as little as possible'. This is a relatively simple and natural way of dealing with the problem. Note that the approximation is really reduced to a problem of norm selection and the distance minimization. For the Euclidean norm, the vector of geometric means (equal to the principal eigenvector for the transitive matrix) is the one which generates it.

Many approximation solutions have been proposed in the past starting with Jensen (1984). More recently, Bozóki et al. (2010), and others Anholcer et al. (2010); Grzybowski (2012) proposed a practical optimization. No study has ever provided an analytic proof of the substantial superiority of any method for approximation over another. Strong statistical evidence (based on 1,000,000 randomly generated matrices) suggests that both solutions (geometric means and the principal eigenvector) are reasonable and do not differ much for 'not-so-inconsistent' (NSI) matrices, as demonstrated in Herman and Koczkodaj (1996).

A further investigation of the selection of the norm (or distance) is beyond the scope of this study. In fact, it may require many years of research before any conclusions could be made and probably the pairwise comparisons may be helpful in it. Unfortunately, not much can be analytically proven for non-transitive matrices. In data processing, it is well expressed by the popular computer concept GIGO (Garbage In—Garbage Out). GIGO summarizes what is known for a long time: getting good results from 'dirty data' is unrealistic and certainly cannot be guaranteed.

Research entities evaluation: the official procedure

The evaluation procedure officially adopted in *Poland* for assessment of research units consists of six steps. Some of them are more or less informal and based largely on the work of experts, whilst the other ones are precisely defined with extensive use of mathematical formulas. In particular the final results of the algorithm highly depends on subjectively defined weights $W_1, \ldots, W_4$ describing importance of each of the criteria $c_1, \ldots, c_4$, as presented in "Preliminaries of the PC method" Section.

Note that due to diversification of the research activities in different areas of science, all the units are divided into relatively small groups of similar entities (e.g. Faculties of Electrical Engineering). Hence, all the 963 units were divided into similarity groups (GWO) of a limited number of units (for example, around 50 in a typical GWO). The procedure was performed independently for each group.

Assessment procedure

1. At the beginning experts proposed weights $W_1, \ldots, W_4$ for each group of mutually comparable entities.
2. Then, each scientific entity X is assigned numerical values with respect of the four criteria as defined in (Table 1). As a result, a vector of four values $O_1(X), \ldots, O_4(X)$ defining how good is unit X with respect to $c_1, \ldots, c_4$ is X is prepared.
3. The experts proposed two artificial entities A_1 and A_2 which will be used as reference units in order to assign every real research unit an appropriate funding level. A_1 and A_2 become part of a ranked group.
4. All the entities are mutually compared within its GWO of comparable entities with respect to all four criteria (Table 1). The result of a single comparison of $X, Y \in U$, where U is the GWO for X and Y, with respect to the i-th criterion is given as:

$$P_i(X,Y) = sgn(O_i(X) - O_i(Y)) \cdot \begin{cases} 0 & \text{if} \quad \Delta O < D \\ \frac{\Delta O - D}{G - D} & \text{if} \quad D \leq \Delta O < G \\ 1 & \text{if} \quad G \leq \Delta O \end{cases} \tag{9}$$

where

$$\Delta O = |O_i(X) - O_i(Y)| \tag{10}$$

$$D = max\left\{\frac{min\{O_i(X), O_i(Y)\}}{10}, \frac{\sum_{Z \in U} O_i(Z)}{10 \cdot card(U)}\right\} \tag{11}$$

$$G = \max\left\{\frac{3 \cdot \min\{O_i(X), O_i(Y)\}}{10}, 3 \cdot D\right\} \tag{12}$$

5. During the currently adopted ranking procedure by the Ministry of Science and Higher Education (2012), the value $V(X, Y)$ is computed according to the formula:

$$V(X,Y) = \sum_{i=1,\ldots,4} W_i P_i(X,Y) \tag{13}$$

where $V(X, Y)$ is the total comparison score of the scientific unit X versus Y, W_i is the rank (importance) of the i-th criterion, and $P_i(X, Y)$ is the result of the pairwise comparisons between X and Y with respect to the i-th criterion.
6. The final rank of the scientific entity $X \in U$ is computed as:

$$R(X) = \frac{1}{card(U) - 1} \left(\sum_{Y \in U \setminus \{X\}} V(X, Y) \right) \tag{14}$$

where $U = \{X_1, \ldots, X_{card(U)}\}$ is the set of the scientific and the two reference (artificial) units to be assessed.

Numerical example

To better understand the algorithm, let us assume that U for some specific type of scientific entities (GWO) consists of six elements $U = \{X_1, \ldots, X_4, A_1, A_2\}$. The result vectors of every scientific unit including the two referential ones determined by experts[4] are given in (Table 2).

The original weights as proposed for this comparisons group in the original algorithm are $W_1 = 0.65$, $W_2 = 0.1$, $W_3 = 0.15$ and $W_4 = 0.1$.

To determine the ranking value $R(X_1)$ every $V(X_1, X_2), \ldots, V(X_1, X_4), V(X_1, A_1)$ and $V(X_1,A_2)$ need to be computed, then the average according to (Eq. 14) need to be computed. For example, to determine $V(X_1,X_3)$ every single $P_i(X_1,X_3)$ need to be calculated. Thus, following the procedure ("Assessment procedure" Section) it is easy to see that the comparisons are $P_1(X_1, X_3) = 0.711$, $P_2(X_1, X_3) = -0.052$, $P_3(X_1, X_3) = 1$ and $P_4(X_1,X_3) = 0$. Since, the rank value of X_1 and X_3 is defined as

$$V(X_1, X_3) = W_1 P_1(X_1, X_3) + \cdots + W_4 P_4(X_1, X_3) \tag{15}$$

hence we have $V(X_1, X_3) = 0.65 \cdot 0.711 - 0.1 \cdot 0.052 + 1 \cdot 0.15 + 0 = 0.607$, and, simultaneously, $V(X_3,X_1) = -0.607$.

After consecutive repeating the procedure for every pair it can be calculated that $V(X_1, X_2) = -0.3$, $V(X_1, X_4) = 1$, $V(X_1, A_1) = 0.736$ and $V(X_1,A_2) = 1$. Thus, the final score for X_1 is $R(X_1) = \frac{1}{5}(-0.3 + 0.607 + 1 + 0.736 + 1) = 0.609$. After calculating R for all other entities algorithm stops. The obtained rank is as follows: $R(X_1) = 0.609$, $R(X_2) = 0.318$, $R(A_1) = 0.046$, $R(X_3) = 0.028$, $R(X_4) = -0.21$ and $R(A_2) = -0.791$.

The final results of the evaluation procedure are given by the following linear ordering: X_1, X_2, A_1, X_3, X_4, A_2. Since there are two referential units, apart of the linear ordering the units are assigned to three categories: A—for the leading ones (here: X_1,X_2, B—for the medium class (here: X_3,X_4), and C—for ones which must improve (here: empty).

The need for a better method for the weight selection

Needless to say that the weights $W_1, \ldots, W_4$ (see step 5 of the procedure) have a significant influence on the final results of the ranking process. Their values determine what kind of achievements (and to what extent) are preferred. Hence, the choice of these weights determines the required policy of the development of scientific entities in *Poland* (the rank position translates into an appropriate funding level).

[4] Data used are real and taken from http://www.nauka.gov.pl. For the purposes of example the number of entities in the group SI1EA has been reduced from 44 to arbitrary selected 4.

Table 2 Group of four mutually comparable scientific entities

Id.	Entity name	O_1	O_2	O_3	O_4
1	X_1	51.97	525	12.64	84.5
2	X_2	84.07	127	1.02	20
3	X_3	41.11	583	4.22	88
4	X_4	33.79	246	7.21	60
5	A_1	39.07	455.40	12.12	59.76
6	A_2	18.99	221.37	5.89	29.05

Recall that these weights were defined by experts in an arbitrary way. Due to the significance of their values, we propose to adopt the selection procedure by computing values of the weighting coefficients from their pairwise comparisons.

There are several reasons to this approach be considered acceptable. One of them is intangibility of the compared achievement assigned to each of the evaluation criteria. Tangible things can be easily measured with reference to some specific unit. Thus, the measure determines the levels of desired features. The intangible factors can be compared in pairs Saaty (2013) without an a priori measurement. In the domain literature, there is considerable evidence indicating that the pairwise comparisons method works when the intangible objects need to be compared Subramanian and Ramanathan (2012).

The algorithm criteria as mentioned in (Table 1) reflect intangible achievements. Experts need a method that would allow them to assess all the objects. The pairwise comparisons method simplifies it by reducing the comapred objects to only two at a time.

Note that the weights tuning mechanism is used also in AHP (Analytic hierarchy process)—another decision making scheme based on the comparing objects in pairs Saaty (1977). In the context of the *AHP* method such weights are often referred to as the criteria with respect to the goal evaluation. Of course, the *AHP* uses the weights in a bit different way. However, the regularity that, the higher the weight of the criterion is, the greater is its impact on the final result, is preserved.

Selection of such important factors as weights in the scientific entity evaluation procedure should be based on transparent, well justified mechanisms. The results should gain a widespread acceptance among the members of the evaluated units. Here again, the pairwise comparisons method can be helpful.

As it is shown in the experiment (see "An experimental survey procedure" Section in the evaluation process may attend any number of experts from different research centers. The pairwise comparisons method also addresses the inconsistency problem. It allows the experts to measure Saaty (1977), to localize Koczkodaj (1993) and to reduce Koczkodaj and Szarek (2010) the inconsistency of the results of comparisons in pairs.

An experimental survey procedure

As with any anticipated change, the proposed CERU[5] conceptual model of the assessment process was vigorously debated in the scientific community (see Kistryn 2013). In particular, the weights $W_1, \ldots, W_4$ corresponding to the importance of the criteria $c_1, \ldots, c_4$

[5] A Polish Committee for Evaluation of Research Units; Polish acronym is KEJN

Table 3 Comparison scale—the values 1,2 and 3 are assigned to the appropriate definitions of intensity or importance. Intermediate judgments are also possible. E.g. value 1.4 corresponds to the situation when one criterion is slightly more preferred than the other

Value	Definition of intensity or importance	Explanation
1	Equal importance	Two criteria equally contribute to the objective
2	Essential or strong importance	Experience and judgments favor one criterion over another
3	Absolute importance	The highest affirmation degree of favoring one criterion over another
1.4	Intermediate judgments	An expert prefers slightly one criterion over another

(Table 1) were the subject of debate and criticism since they were established in an arbitrary way.

The goal of the experimental survey was to provide the weighting coefficients assuming that:

- the PC method is the core of the experiment; so the values are better justified,
- any arbitrarily large number of experts can express their preferences,
- the expert judgment consistency should be evaluated and kept at a possible low level.

The relative preference criteria are established on the basis of partial values (pairwise comparisons) provided by experts in form of the matrix M (see Eq. 1). All the matrices for which inconsistency index (Eq. 7) is higher than 0.1 are excluded from the ranking as not reliable enough.

The final weights are computed according to the PC methodology given as (Eq. 3). The authors invited members of the academic community who know the specificities of Polish research units to provide expert assessments by an Internet survey. The surveyed scientists used the reference scale (Table 3) that helps them in translating intuitive meanings of assessments into numbers.

The choice of scale is also a challenging problem. It has been extensively discussed in the literature Fülöp et al. (2010); Dong et al (2008); Ji and Jiang (2003); Salo and Hämäläinen (1997); Triantaphyllou et al. (1994). There is no "one fits all" scale, although some studies argue that a certain scale should give more reliable results than another. In Fülöp et al. (2010), a small scale from 1 to 3 (Table 3) is shown to have the best mathematical properties (relted to the convexity)for the PC method. It was adopted by the authors for this study. After all, practically all modern languages have only three levels of gradation in the grammar (e.g., good -> better -> the best).

Despite the scale recommendation the Internet survey application allowed respondents to set any value (except 0) of the m_{ij} ratio between $\frac{1}{99}$ and 99. Introduction (suggestion) the scale while allowing almost the free choice of ratio is an attempt to find a compromise between the desire to give an intuitive interpretation for some numerical values (the scale), and allowing the experts to the greatest possible precision in expressing beliefs. Moreover, thanks to introducing the scale all the experts share the same correspondence between the numerical values and the intuitive descriptions of importance. This helps to minimize the risk of situation in which two experts sharing the belief that c_i is absolutely more important than c_j assign two essentially different (although greater than one) values of m_{ij}. The scale introduction allows for identification all such cases, and excluding the identified outliers from the ranking. The candidates for outliers are experts whose answers are significantly off the scale. Usually their response also has a large inconsistency.

The meaning of the adopted scale is quite intuitive. For example, if an expert assigns W_i/W_j to 1, this means that the criteria i and j are of equal importance. On the other hand if, for instance, $2 < W_i/W_j < 3$ then, according to the adopted textual interpretation (Table 3), the i-th criterion was recognized as essentially more important than the j-th one.

In the ideal case, there should be always $W_i/W_j \cdot W_j/W_k = W_i/W_k$. However, because each of the three ratios are determined independently, in practice this is often not the case. Hence, very often there are some triads of ratios which do not meet this equality. This situation is related to the problem of data inconsistency in the PC matrix, which is discussed more thoroughly in "Data inconsistency and how to deal with it" Section.

Survey results

Survey data

The survey involved *37* researchers from *17* Polish and foreign scientific institutions engaged in research in the field of technical and engineering sciences. Most of them are tenured faculty members at Universities in Poland, USA, Canada, and Australia although some of them declared employment in research institutes. The vast majority of respondents declared the position of a full professor or equivalent.[6] A few persons held the prestigious title of distinguished professor.

Every participant of the survey had to answer six questions and, thus, determine six ratios: $\frac{w_1}{w_2}, \frac{w_1}{w_3}, \frac{w_1}{w_4}, \frac{w_2}{w_3}, \frac{w_2}{w_4}, \frac{w_3}{w_4}$. The answers allowed for formation of the partial PC matrix M_r in the form:

$$M_r = \begin{bmatrix} 1 & \frac{w_1}{w_2} & \frac{w_1}{w_3} & \frac{w_1}{w_4} \\ \frac{w_2}{w_1} & 1 & \frac{w_2}{w_3} & \frac{w_2}{w_4} \\ \frac{w_3}{w_1} & \frac{w_3}{w_2} & 1 & \frac{w_3}{w_4} \\ \frac{w_4}{w_1} & \frac{w_4}{w_2} & \frac{w_4}{w_3} & 1 \end{bmatrix} \tag{16}$$

To synthesize the final results the authors used almost all the gathered matrices M_r. The only exceptions were five result sets with the very high inconsistency index $\mathcal{K}(M_r)$ (over 0.836), and the inconsistency index $Ic(M_r)$ higher than 0.1. Although all the rejected cases differ in detail, most of the rejected authors indicated very significant importance of the first criterion (scientific and/or creative achievements) over other arbitrarily chosen criteria. Unfortunately, due to the large inconsistency (in the literature $Ic(M_r)$ higher than 0.1 is considered as unacceptable Saaty 2005) their opinions have not been taken into account[7] in the synthesized matrix $\widehat{M}$.

All the 32 admissible partial results M form the following final output matrix $\widehat{M}$, which looks as follows:

$$\widehat{M} = \begin{bmatrix} 1 & 1.813 & 1.503 & 1.784 \\ 0.552 & 1 & 0.952 & 1.296 \\ 0.666 & 1.05 & 1 & 1.302 \\ 0.561 & 0.772 & 0.768 & 1 \end{bmatrix} \tag{17}$$

[6] In Poland there are professor extraordinarius and professor ordinarius.

[7] On the other hand, even if two rejected cases were taken into account their impact on the final result would be negligible.

The normalized weight vector ω derived from $\widehat{M}$ using the geometric mean method is as follows:

$$\omega = [0.36 \quad 0.22 \quad 0.236 \quad 0.184]^T \tag{18}$$

which means that the invited experts found that *rank*(c_1)—the relative importance of *scientific and/or creative achievements* criterion is 0.36, *rank*(c_2)—*scientific potentiality* criterion is 0.22, *rank*(c_3)—*tangible benefits of the scientific activity* criterion is 0.236, and finally *rank*(c_4)—*intangible benefits of the scientific activity* criterion is 0.184.

The inconsistency indices for $\widehat{M}$ are low. The more sensitive for local perturbations Koczkodaj's index $\mathscr{K}(\widehat{M}) = 0.241$ whilst $Ic(\widehat{M}) = 0.002$. The standard geometric deviation for the appropriate $\widehat{m}_{ij}$ defined as:

$$\sigma_g(\widehat{m}_{ij}) = \exp\left(\sqrt{\frac{\sum_{k=1}^{32}\left(\ln m_{ij}^{(k)} - \ln \widehat{m}_{ij}\right)^2}{32}}\right) \tag{19}$$

forms the matrix $\widehat{M}_\sigma = [\sigma_g(\widehat{m}_{ij})]$ as follows:

$$\widehat{M}_\sigma = \begin{bmatrix} 1 & 2.093 & 2.03 & 1.787 \\ 2.093 & 1 & 1.913 & 1.831 \\ 2.03 & 1.913 & 1 & 1.61 \\ 1.787 & 1.831 & 1.61 & 1 \end{bmatrix} \tag{20}$$

It is easy to see (Eq. 20) that the most controversial (with the highest standard geometric deviation) comparison is between the *scientific and/or creative achievements* c_1, and the *scientific potentiality* c_2. On the other hand experts were most unanimous comparing c_3 and c_4 (the standard geometric deviation of $\sigma_g(\widehat{m}_{34}) = \sigma_g(\widehat{m}_{43}) = 1.61$ is the closest to 1).

Results: different perspectives

Experts were chosen at random among those who know the specificity of Polish technical scientific units. Most of the experts are affiliated at the Polish universities or research institutes. Three experts are affiliated at foreign universities, although they worked at Polish universities in the past. Since the aim of the survey was to propose the weights $w_1, \ldots, w_4$ for technical scientific units (including such units as the departments of mathematics, physics or computer science), hence most experts is working or has worked in such institutions. On the other hand, it was important for the authors of the survey that the experts came from different research centers. The best represented university is *AGH UST*—the place of work of the second and the third author. The representations of other *16* scientific units count from one to three experts. Out of the all respondents the authors chose the *VIP* group of six the most influential people consisting of distinguished professors and former or current members of official governmental and scientific bodies, including CERU. The overall results taking into account two special groups: *VIP* group and experts employed at the best represented *AGH UST* are shown below (Fig. 1).

Among the experts whose opinion have been taken into account there may be distinguished a group of full professors (or professor ordinarius)—18 persons, associate professors (or doctors with habilitation)—6 persons, assistant professors (or doctors)—7 persons, assistants—1 person. The ranking result with respect of these groups (assistant professors and assistants are treated as a single group) are shown below (Fig. 2).

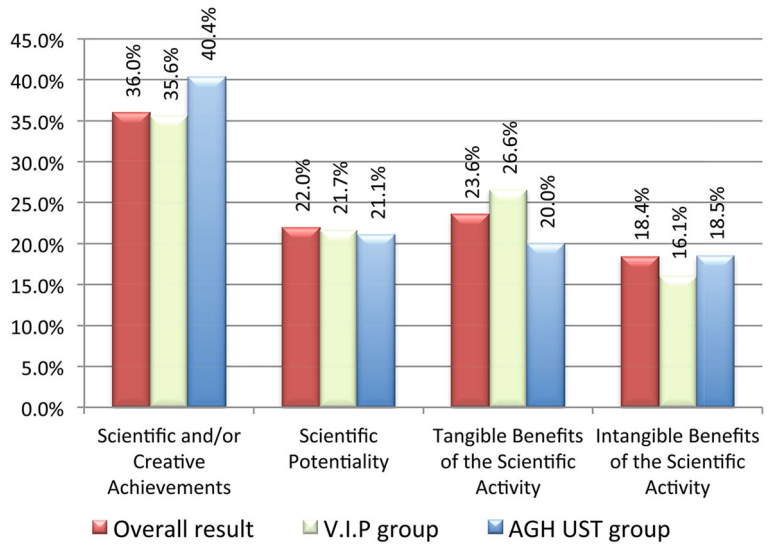

Fig. 1 The overall survey results with the VIP and AGH UST employee groups

Montecarlo discrepancy validation

As a validation method for the survey data the authors adopt ten times repeated twofold cross-validation procedure Kohavi (1995). In every repetition the survey sample is randomly split into two disjoint sets $S_1 = \{M_1, \ldots, M_{16}\}$ and $S_2 = \{M_{17}, \ldots, M_{32}\}$. Both groups are used to synthesize matrices $\widehat{M}_1$ and $\widehat{M}_2$, next two ranking vectors $a = [a_1, \ldots, a_4]^T$ and $b = [b_1, \ldots, b_4]^T$ are computed. The vector a is called the reference rank vector, whilst b is called the validation rank vector. For each pair of vectors a and b the discrepancy vector $d = [|a_1 - b_1|, \ldots, |a_4 - b_4|]^T$ is computed.

The values $d_1, \ldots, d_4$ are adopted as a measures of fit. They provide information on how much the synthesized ranking values for the criteria $c_1, \ldots, c_4$ provided by the first group of experts differ from the ranking values provided by the second group. Intuitively speaking the adopted procedure simulates the situation where two disjoint group of experts provide two competitive rankings. Then the first ranking is validated by the second one. The validation procedure has been repeated ten times, so there are ten vectors $a^{(1)}, \ldots, a^{(10)}$ and ten vectors $b^{(1)}, \ldots, b^{(10)}$. The final reference rank a^{avg} and the discrepancy (fit indicator) vector d^{avg} are computed as arithmetic means:

$$a^{\text{avg}} = \left[\frac{1}{10}\sum_{i=1}^{10} a_1^{(i)}, \ldots, \frac{1}{10}\sum_{i=1}^{10} a_4^{(i)}\right] \tag{21}$$

and

$$d^{\text{avg}} = \left[\frac{1}{10}\sum_{i=1}^{10} d_1^{(i)}, \ldots, \frac{1}{10}\sum_{i=1}^{10} d_4^{(i)}\right] \tag{22}$$

As a result of the conducted experiment, the following numerical values are obtained:

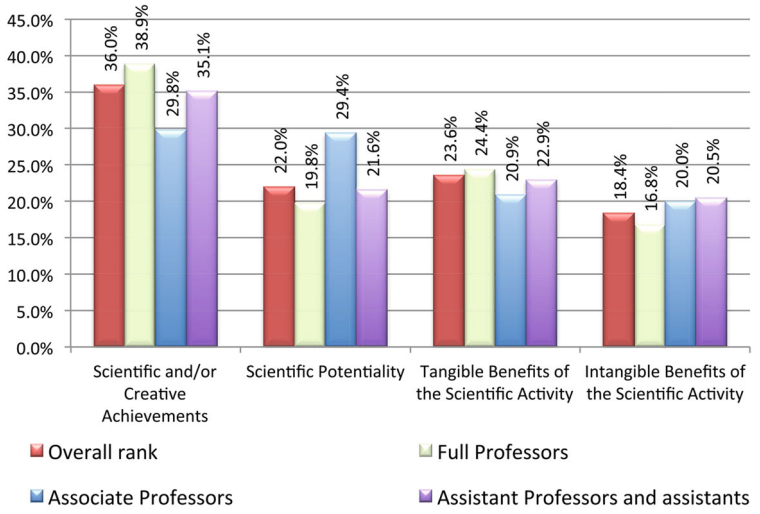

Fig. 2 The survey results in groups of full professors, associate professors, and assistant professors and assistants (the overall result was left as the reference)

$$a^{\text{avg}} = [0.364 \quad 0.219 \quad 0.234 \quad 0.183]^T \tag{23}$$

$$d^{\text{avg}} = [0.053 \quad 0.037 \quad 0.043 \quad 0.023]^T \tag{24}$$

It is easy to see, that the obtained rank result is (on average) similar to the overall result of the survey (Eq. 18). In particular both vectors a_{avg} (Eq. 23) and ω (Eq. 18) propose the same order of criteria importance. Their individual numerical values are also close to each other. The absolute average absolute difference between individual values in vectors $a^{(i)}$ and $b^{(i)}$ seem to be reasonably small since they are almost an order of magnitude less than the values in a^{avg}. They suggest that regardless of the selection of the group criterion c_1 should be the most important one $a_1^{\text{avg}} - d_1^{\text{avg}} > a_i^{\text{avg}} + d_i^{\text{avg}}$. Unfortunately there is no similar guarantee in the case of any other criterion. The values $a_1^{\text{avg}} \pm d_1^{\text{avg}}, \ldots, a_4^{\text{avg}} \pm d_4^{\text{avg}}$ indicate the discrepancy intervals in which the weights of criteria $c_1, \ldots, c_4$ established by the competitive team of experts are expected to be found.

The results a^{avg} and d^{avg} (Eqs. 23, 24) were calculated on the assumption that S_1 and S_2 are equal in size. Thus, in our case both of them count 16 elements. Of course when the size of the set of experts is changing the values a^{avg} and d^{avg} may get changed. For example, according to the intuition (confirmed in tests), the smaller set S_1 (and the larger S_2) the higher discrepancies $d_1^{\text{avg}}, \ldots, d_4^{\text{avg}}$. For example for $|S_1| = 6$ and $|S_2| = 26$ the sample reference rank and the discrepancy vectors are:

$$a_6^{\text{avg}} = [0.336 \quad 0.221 \quad 0.252 \quad 0.189]^T \tag{25}$$

$$d_6^{\text{avg}} = [0.083 \quad 0.048 \quad 0.051 \quad 0.027]^T \tag{26}$$

The adopted Montecarlo discrepancy validation procedure tries to model a realistic situation in which one group of experts provides one rank, whilst the other group (disjoint with the first one) creates another rank. Both groups call into question the results of its

opponent. As demonstrated by the tests carried out when both groups are composed of experts with a similar scientific background the discrepancies might not be to high.

Also the further research on the inconsistency of synthesized PC matrix $\widehat{M}$ seem to be interesting. In particular the relationship between the values of inconsistency indices $Ic(\widehat{M})$ and $\mathscr{K}(\widehat{M})$ and the deviations of the individual expert judgements in matrices $M_1, \ldots, M_r$ need better explanation.

Discussion

The survey concerned the basic scientific units at universities in the field of technical and engineering sciences. Thus, the gathered results do not apply to social sciences or the arts. The surveyed researchers have made six comparisons between the four criteria $c_1, \ldots c_4$ (Table 1). They could almost freely choose between ratios from $\frac{1}{99}$ to 99, thus indicating which criterion is more (and how much) important. However, a small scale was recommended following the theory proved in Fülöp et al (2010).

Comparing the survey results (Fig. 1) with the weights adopted in the official government regulation Ministry of Science and Higher Education (2012) (they are: c_1—0.65, c_2—0.1, c_3—0.15, and c_4—0.1) it should be noted that they differ in the intensity of preferences, although they tend to be similar with regard to the order of preferences. In both rankings, the criterion designated as most important is c_1 and the second most important criterion is c_3. However, the weight of c_1 resulting from the survey is almost two times less than the one assumed in the regulation. On the other hand, c_2 obtained from the survey is a bit higher than the one adopted in the official document. According to the survey, the criterion c_2 is slightly less important than c_3 but more important than c_4, whilst the regulation assumes that the weights of c_2 and c_4 are the same. In both these cases the weights obtained from the survey are higher than the ones adopted in the regulation.

The regulation retains the dominant criterion c_1, whilst the other criteria are less important. In fact, it is enough for the scientific entity to be strong in c_1 to avoid having to worry about the other criteria. The survey participants were in favor of a more balanced model in which c_1 is still the most important criterion, but is not predominant. They also appreciate the importance of other criteria with particular emphasis to c_3 (tangible benefits of the scientific activity). Hence, in the model proposed by the surveyed researchers the predominant position of only one criterion c_1 has been replaced by the predominant position of the pair (c_1,c_i), where c_i is any other criterion out of c_2, c_3, c_4 (please note that the rank of c_1 and the rank of any other criterion is more than 0.5). Therefore, based on the survey results, such a model would be recommended in which the evaluated scientific entity is good in terms of c_1 but is also good in terms of at least one other criterion, c_2, c_3 or c_4. Of course, the appropriate selection of weights will not solve all the problems related to the scientific entity evaluation algorithm. In particular, it does not prevent the "displacement" of good results in the most important category c_1 by outstanding (in the number but not in the quality nor originality of achievements) results in the less important categories.

The present work tackles many problems and can be a starting point for further research in various areas. In particular, although the new algorithm weights deriving in the official scientific units evaluation procedure is proposed, there are also other highly subjective parts of the algorithm where the PC methods might help. One of them is choosing by experts the so-called reference scientific units.

Conclusions

The identification of major criteria is a key issue for building a conceptual evaluation model. Once it is done, the final weights are computed from the relative pairwise comparisons by synthesizing them. The model demonstrated in this paper has been used in Poland for evaluating scientific entities consistent with the one proposed by the Ministry of Science and Higher Education (2012). However, the presented method is flexible and can accommodate all criteria at hand, including both quantitative and qualitative factors. No model is ideal and usually undergoes evolution as time passes. It is anticipated that CERU will be improving the model to evaluate academic entities at the national level. Using our approach to compute the weights is a time consuming but necessary exercise since it will benefit the entire country when the weights are computed (as opposed to arbitrary assignment). In particular, the success-index could improve the performance evaluation methods Franceschini et al (2012) in the evolved model.

Acknowledgements The following faculty members have kindly provided their input and agreed to be listed (according to the Polish tradition with the scientific titles; the country is listed if outside Poland): Prof. dr hab inż. Ryszard Tadeusiewicz (AGH UST), Prof. dr hab. Stanisław Kistryn (The Jagiellonian University), Prof. dr Bogdan Denny Czejdo, Belk Distinguished Professor (Fayetteville State University, USA), Prof. dr hab. Stan Matwin (IPI PAN, Poland, Dalhousie University, Canada), Prof. Eugene Eberbach (Rensselaer Polytechnic Institute, USA), dr hab. Krzysztof Oprzędkiewicz, Prof. AGH (AGH UST), Prof. dr hab. Maria Mach-Król (University of Economics in Katowice), Prof. dr hab. Halina Kwaśnicka (Wroclaw University of Technology) and Prof. Witold Kwaśnicki (Administration and Economics University of Wroclaw), dr inż. Jarosław Wąs (AGH UST), dr inż. Radosław Klimek (AGH UST), dr inż. Paweł Skrzyński (AGH UST), mgr inż. Krzysztof Kluza (AGH UST), Weronika Adrian (AGH UST). The authors would like to thank all respondents of the Internet survey. The authors would like to thank T. Kakiashvli, MD, Amanda Dion—Groleau (Laurentian University), Grant O. Duncan (student at Laurentian University; Team Lead at Health Sciences North, Sudbury, Ontario, Canada), and Mr Ian Corkill for the editorial improvements. Mr Karol Wójcik (student at AGH UST) has developed and maintained the Internet surveying tool.

References

Aczél, J., & Saaty, T. L. (1983). Procedures for synthesizing ratio judgements. *Journal of Mathematical Psychology 27*(1):93–102. doi:10.1016/0022-2496(83)90028-7.

Anholcer, M., Babiy, V., Bozóki, S., & Koczkodaj, W. W. (2010). A simplified implementation of the least squares solution for pairwise comparisons matrices. *Central European Journal of Operations Research 19*(4):439–444.

Bozóki, S., & Rapcsak, T. (2008). On Saaty's and Koczkodaj's inconsistencies of pairwise comparison matrices. *Journal of Global Optimization 42*(2):157–175.

Bozóki, S., Fülöp, J., & Rónyai, L. (2010). On optimal completion of incomplete pairwise comparison matrices. *Mathematical and Computer Modelling 52*(1–2):318 – 333, doi:10.1016/j.mcm.2010.02.047, URL http://www.sciencedirect.com/science/article/pii/S0895717710001159.

Condercet, M. (1785). *Essay on the Application of Analysis to the Probability of Majority Decisions*. Paris:Imprimerie Royale.

Crawford, G. B. (1987). The geometric mean procedure for estimating the scale of a judgement matrix. *Mathematical Modelling 9*(3–5):327 – 334 doi:10.1016/0270-0255(87)90489-1, URL http://www.sciencedirect.com/science/article/pii/0270025587904891.

Dong, Y., Xu, Y., Li, H., & Dai, M. (2008). A comparative study of the numerical scales and the prioritization methods in AHP. *European Journal of Operational Research 186*(1):229–242.

Fechner, G. T. (1966). *Elements of psychophysics, vol 1. Holt*, Rinehart and Winston, New York.
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2012). Evaluating research institutions: The potential of the success-index. *Scientometrics 96*(1):85–101.
Fülöp, J., Koczkodaj, W. W., & Szarek, S. J. (2010). A different perspective on a scale for pairwise comparisons. *Transactions on Computational Collective Intelligence 1*:71–84.
Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva 41*(4):277–304.
Geuna, A., of Sussex SPRU : Science U, Research TP (1999). The Changing Rationale for European University Research Funding: Are There Negative Unintended Consequences? Electronic working paper series, University of Sussex, SPRU, URL http://books.google.pl/books?id=lBpuMwEACAAJ.
Grzybowski, A. Z. (2012). Note on a new optimization based approach for estimating priority weights and related consistency index. *Expert Systems with Applications 39*(14):11,699–11,708.
Herman, M. W., & Koczkodaj, W. W. (1996). A monte carlo study of pairwise comparison. *Inf Process Lett 57*(1):25–29 doi:10.1016/0020-0190(95)00185-9.
Jensen, R. E. (1984). An alternative scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology 28*(3):317 – 332. doi:10.1016/0022-2496(84)90003-8, URL http://www.sciencedirect.com/science/article/pii/0022249684900038.
Ji, P., & Jiang, R. (2003). Scale transitivity in the AHP. *Journal of the Operational Research Society 54*(8):896–905 doi:10.1057/palgrave.jors.2601557.
Kistryn, S. (2013). Mission of CEAE – how easy is to evaluate the quality? (Misja KEJN – czy łatwo ocenić jakość?). URL http://forumakademickie.pl/fa/2012/05/misja-kejn-czy-latwo-ocenic-jakosc/.
Koczkodaj, W. W. (1993). A new definition of consistency of pairwise comparisons. *Math Comput Model 18*(7):79–84. doi:10.1016/0895-7177(93)90059-8.
Koczkodaj, W. W., & Szarek, S. J. (2010). On distance-based inconsistency reduction algorithms for pairwise comparisons. *Logic Journal of the IGPL 18*(6):859–869.
Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. San Mateo: Morgan Kaufmann. pp 1137–1143.
Kułakowski, K. (2013). A heuristic rating estimation algorithm for the pairwise comparisons method. *Central European Journal of Operations Research* pp 1–17, doi:10.1007/s10100-013-0311-x.
Ministry of Science and Higher Education. (2012). Regulation on principles of science financing (Polish: Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego w sprawie kryteriów i trybu przyznawania kategorii naukowej jednostkom naukowym). *Dziennik Ustaw Rzeczypospolitej Polskiej 877*, URL http://www.bip.nauka.gov.pl/_gAllery/19/31/19319/poz._877.pdf.
Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology 15*(3):234 – 281, doi:10.1016/0022-2496(77)90033-5, URL http://www.sciencedirect.com/science/article/pii/0022249677900335.
Saaty, T. L. (2005). The analytic hierarchy and analytic network processes for the measurement of intangible criteria and for decision-making. In: Multiple Criteria Decision Analysis: State of the Art Surveys, International Series in Operations Research and Management Science, vol 78, Springer New York, pp 345–405. doi:10.1007/0-387-23081-5_9.
Saaty, T. L. (2008) Relative Measurement and Its Generalization in Decision Making. Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors. The Analytic Hierarchy/Network Process. *Estadística e Investigación Operativa / Statistics and Operations Research (RACSAM) 102*:251–318.
Saaty, T. L. (2013). On the measurement of intangibles. A principal eigenvector approach to relative measurement derived from paired comparisons. *Notices of the American Mathematical Society 60*(02):192.
Salo, A. A., & Hämäläinen, R. P. (1997). On the measurement of preferences in the analytic hierarchy process. *Journal of Multi-Criteria Decision Analysis 6*(6):309–319. doi:10.1002/(SICI)1099-1360(199711)6:6<309::AID-MCDA163>3.0.CO;2-2.
Subramanian, N., & Ramanathan, R. (2012). A review of applications of analytic hierarchy process in operations management. *International Journal of Production Economics 138*(2):215–241.
Thurstone, L. L. (1994). A law of comparative judgment, reprint of an original work published in 1927. *Psychological Review 101*:266–270.
Triantaphyllou, E., Lootsma, F. A., Pardalos, P. M., & Mann, S. H. (1994). On the evaluation and application of different scales for quantifying pairwise comparisons in fuzzy sets. *Journal of Multi-Criteria Decision Analysis 3*(3):133–155.
Wang, X., Liu, D., Ding, K., & Wang, X. (2011). Science funding and research output: A study on 10 countries. *Scientometrics 91*(2):591–599.