A Qualitative Decision-Support Model for Evaluating Researchers

Katerina Taškova
Department of Computer Systems
Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: katerina.taskova@ijs.si

Daniela Stojanova
Slovenian Forestry Institute, Ljubljana, Slovenia
E-mail: daniela.stojanova@gozdis.si

Marko Bohanec and Sašo Džeroski
Department of Knowledge Technologies
Jožef Stefan Institute, Ljubljana, Slovenia
E-mail:[marko.bohanec, saso.dzeroski]@ijs.si

Keywords: evaluation of researchers, research performance indicators, qualitative multi-attribute model, method DEXi, sensitivity analysis

Received: June 14, 2007

The evaluation of research work is an essential element of the scientific enterprise. In general, the evaluation of researchers and their work is highly dependent on the social and economic condition of the country in which the researchers work. The most commonly used form of evaluation is based on peer review. In Slovenia, a quantitative model for evaluating researchers has been developed and used by the Slovenian Research Agency, which has been criticized by the public. In order to alleviate some of the problems with this model and motivate further discussion on this issue, we propose an alternative qualitative model. The model belongs to the paradigm of hierarchical multi-attribute models and has been developed after a literature survey on existing models in foreign countries.

Povzetek: Članek prikazuje kvalitativni večparametrski model za ocenjevanje raziskovalcev in primer njegove uporabe pri ocenjevanju raziskovalcev s področja računalniških znanosti.

1 Introduction

Evaluation is an essential characteristic of human activity and is perhaps the single most important and sophisticated cognitive process in the repertoire of human reasoning and logic. It is also the one that has defied adequate explanation for nearly two millennia [1]. Without evaluation there is simply no means for distinguishing the bad from the good, the worthwhile from the worthless, and the significant from the insignificant.

In science, evaluation has been an essential element of the scientific enterprise, even before the appearance of the first scientific journals, usually in the form of peer review. In the past few decades, the evaluation of scientific research, and in particular researcher performance, has changed substantially in terms of scale and scope [2], as well as methodology. In part, these changes have occurred as a result of attempts to guide, regulate and control research agendas and priorities, not only in regard to distributing research funds, but also to influence the system of scientific research itself [3]. Therefore, criteria other than strictly scientific ones (e.g., social and political criteria) have been introduced, which further increase the complexity of evaluation. The new pillars of research evaluation include: governments; politicians; the media; social movements; and non-governmental organizations.

Typically, the evaluation takes place at a national level and each country has its own national model for evaluating research and allocating research funds. In Slovenia, such a model has been built by the Slovenian Research Agency (ARRS) [5]. This model uses data from the Slovenian Current Research Information System (SICRIS) [4] and the on-line bibliographic database COBISS [6] that maintain detailed information for every registered researcher in Slovenia. It also relies on the Web of Science database for citation information [9]. Different variants of the model are used for different disciplines and for evaluating applications to different ARRS's calls (for young researchers, for project leaders etc). Several of the variants of the model have been published together with the specific calls; these use several scientific performance indicators and combine them in a linear fashion. In the past decade, the model has been a very popular topic of discussion and criticism from the public and especially from researchers and research organizations.

Motivated by this situation, we have developed a system for evaluating researchers that uses the paradigm of qualitative multi-attribute modeling (see section 2). Our qualitative model was built on the basis of literature survey and takes into account existing foreign models for evaluation of researchers and their work. The model proposes a new methodological approach, based on qualitative multi-attribute modeling, and uses some new indicators. We propose it as an initial alternative to the existing model used by ARRS and hope to motivate further discussion on this important topic in Slovenia.

2 Methodology

Methodologically, we have taken the approach of model-based decision support [10]. We used the software tool DEXi [11] to construct a qualitative multi-attribute model aimed at evaluating and analyzing researchers. DEXi is particularly suitable for a hierarchical decomposition of evaluation problems that require judgment and qualitative reasoning.

A DEXi model is characterized by the following:

- Each model consists of a number of hierarchically structured variables called *attributes*.
- *Input attributes* are terminal nodes of the hierarchy.
- Attributes are aggregated through several levels of *aggregate attributes* into the overall assessment, which is represented by the *root attribute* of the hierarchy.
- All the attributes in the model are *qualitative*: they can take only discrete symbolic values.
- The aggregation of values in the model is defined by *decision rules*.

An example of decision rules is shown later in Fig. 3. There, each row represents a decision rule that maps two attributes, *Quality* and *Relevance*, to the *Evaluation of Researchers*. For instance, rule 1 in Fig. 3 states that if *Quality* is “Very Low” and *Relevance* is “Medium” or worse, then the *Evaluation* is “Unsatisfactory”.

The model is gradually hand-crafted through four steps [10]: (1) identifying attributes, (2) structuring attributes, (3) defining attribute scales, and (4) defining decision rules. If necessary, these steps can be iterated. The model-building process is supported by the software tool DEXi, which facilitates the development of attribute trees, definition of decision rules, evaluation and analysis of options, and graphical output.

DEXi is freely available for download from http://kt.ijs.si/MarkoBohanec/dexi.html.

Usually, DEXi models are developed in collaboration between decision analysts and experts in the given field. Typically, experts suggest attributes and decision rules, while decision analysts conduct the process and define components of the model. The decision rules can be defined explicitly in tabular form or implicitly by specifying the relative importance (weight) of the contributing attributes.

The importance of attributes is most often modeled by weights in conventional multi-attribute models [12]. Each attribute is given a weight that defines the impact of that attribute to the final evaluation: the higher the weight, the more important the attribute. In DEXi, the relationship between attributes is modeled by decision rules and, in principle, there is no need for weights. However, for comparison with conventional methods, DEXi does use attribute weights; it can approximately transform decision rules to weights and vice versa:

- *From decision rules to weights* [13]: DEXi regards the currently defined decision rules as if they were points in a multi-dimensional space and approximates them by a hyperplane, using a least-squares linear regression method. From the hyperplane, it estimates approximate average weights of attributes. In Fig. 3, the weights of *Quality* and *Relevance*, obtained in this way, are 71 % and 29 %, respectively.
- *From weights to decision rules* [11]: In this case, the given weights define a multi-dimensional hyperplane, which is used to construct a complete table of decision rules. Again, each decision rule is considered to be a point in the space whose value is approximated from the hyperplane. The table constructed in this way is typically used to provide an initial ruleset, which is then reviewed by the decision-maker and possibly modified on a rule-by-rule basis.

3 Indicators of research performance

Applying research performance indicators in practice is not a straightforward task. It is important to clarify what role the indicators will play in the assessment of research, which indicators should be selected, and what possible unintended consequences could arise from their application. The problem with all quantitative indicators is that research practices vary across different fields, and it is necessary to determine what level of aggregation is to be used and the form in which the results will be presented. The vast majority of the literature discusses these issues for bibliometric indicators only. However, they affect all quantitative indicators.

A literature review was undertaken to examine quantitative performance indicators used in the evaluation of research. Quantitative evaluations of research have generally been conducted by scientometricians, bibliometricians, information and library scientists, and used indicators of quantity, quality, impact, or influence of research [7].

The indicators can be easily divided into bibliometric and non-bibliometric. *Bibliometric* indicators are based on published literature in all of its forms – journal articles, monographs, book chapters, conference papers, patents, and citations. *Non-bibliometric* measures encompass all other readily quantifiable indicators, such as the ability to attract external funding and measures of esteem: honors and awards, editorship of journals, membership of major national and international professional societies, keynote addresses, PhD students data, etc. [8]. However, we should be careful when using non-bibliometric indicators because they can be a poor reflection of research activities in areas of applied research. Patents are regarded as a much better indicator of the output in these disciplines. Patent indicators are often used to measure the economic or innovative strength of a country in a certain area; thus, many analyses are undertaken on the macro-level in a cross-country comparison. The simplest patent indicator is the number of patents.

The number of citations is a measure of the strength of influence of a body of research, when applied to sufficiently large aggregates. Citation analyses are more difficult to undertake than publication analyses. The citations used in standard bibliometric analyses are the references contained in selected journals to other journals in the Web of Science (WoS) framework [9].

In using the citation indicator, we follow ARRS and use a citation indicator slightly different from the standard definition. Namely, due to the difficulty in obtaining detailed and consistent necessary information on citations (which is stored in different databases, using different formats of records, etc.), we use a combined approach. This approach defines the citation indicator as a weighted sum of the number of cited papers and the normalized number of citations, taking into account only plain citations (without self-citations).

We will use this citation indicator in the models together with the normalized number of citations. The former is important in situations where only one paper from a researcher's bibliography has been cited, while the rest of the work is unknown. Self-citations are excluded from the analysis.

The input to the present ARRS' model does not cover the entire space of quantitative indicators. Most notably, information, such as citations of books and citation of papers published at international conferences, are not included. Also, citation analyses are based only on the number of citations and not the number of cited papers. For some research areas (e.g. computer science), this can have a high influence on the overall evaluation.

Another important attribute is the ability to attract external funding. ARRS collects detailed information on the level of external funding attracted by individual researchers, which however is not publicly available. In our model we use the information on the number of European and National research projects as a proxy for this information.

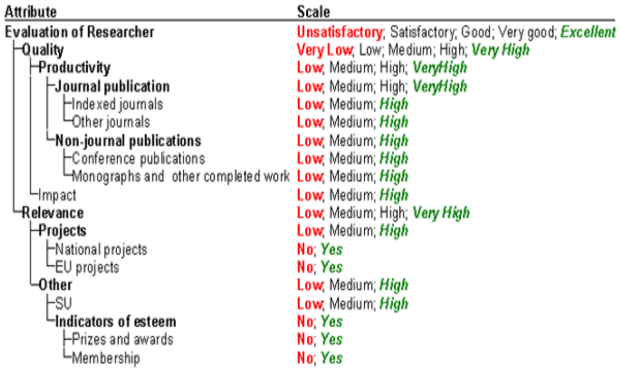
4 Definition of the model

To define a DEXi model, we first need to identify the input attributes. We then have to specify the hierarchical structure of the model and the scales of the attributes. Finally, we need to define the decision rules for each internal node in the hierarchy. Below, we discuss each of these issues for our model for evaluating researchers.

The input attributes in our case are the performance indicators discussed in Section 3. More specifically, at the lowest level of the hierarchy we have the following attributes: *Indexed journals, Other journals, Conference publications, Monographs and other completed work, Impact, National projects* and *EU projects, SU, Prizes and awards* and *Membership*. In some of the models (M2, M2a), *Impact* is not an input attribute, but rather an aggregated one, which takes as input *Norm. num. citations* and *Num. cited Papers* (as discussed in Section 3).

The hierarchy of attributes is defined as follows (Fig. 1): At the top is the root attribute *Evaluation of Researcher*. It is decomposed into two descendants: *Quality* and *Relevance*. *Quality* aggregates *Productivity* and *Impact*. *Productivity* reflects the bibliometric indicators and is decomposed into *Journal publications* and *Non-journal publication*. *Relevance* incorporates mainly non-bibliometric indicators and is divided into *Projects* and *Other,* decomposed into *SU* (COBISS Index of professional success) and *Indicators of esteem*. Fig. 1 also gives the scales of the attributes.

The input attributes under *Relevance* have two values (Yes, No), the intermediate ones have three (Low, Medium, High) and the attribute *Relevance* has four discrete ordered values (from Low to Very High). The input attributes under *Quality* have three values (Low, Medium, High), the intermediate four (from Low to Very High). *Quality* and *Evaluation* have five ordered values each (from Very Low to Very High, and from Unsatisfactory to Excellent, respectively).

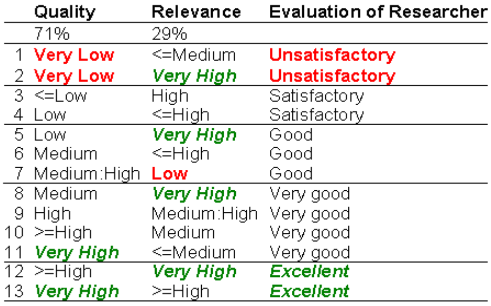


Attribute	Scale
Evaluation of Researcher	Unsatisfactory; Satisfactory; Good; Very good; ***Excellent***
├**Quality**	Very Low, Low; Medium; High; ***Very High***
│├**Productivity**	Low, Medium; High; ***VeryHigh***
││├**Journal publication**	Low, Medium; High; ***VeryHigh***
│││├Indexed journals	Low, Medium; ***High***
│││└Other journals	Low, Medium; ***High***
││└**Non-journal publications**	Low, Medium; ***High***
││ ├Conference publications	Low, Medium; ***High***
││ └Monographs and other completed work	Low, Medium; ***High***
│└Impact	Low, Medium; ***High***
└**Relevance**	Low, Medium; High; ***Very High***
├**Projects**	Low, Medium; ***High***
│├National projects	No; ***Yes***
│└EU projects	No; ***Yes***
└**Other**	Low, Medium; ***High***
├SU	Low, Medium; ***High***
└**Indicators of esteem**	No; ***Yes***
├Prizes and awards	No; ***Yes***
└Membership	No; ***Yes***

Figure1: The structure and scales of the evaluation model

The input attributes are discrete with preferentially ordered scales (values are listed in Fig. 1 from the least to the most desirable one, e.g., Unsatisfactory to Excellent). Such qualitative inputs could be obtained, for example, through a peer review process, where the quality of a researcher is assessed on a qualitative scale along each individual dimension (indicator). Another approach to obtaining qualitative input values is the discretization of continuous values (which in our case are readily available for most indicators, e.g., *Num. cited Papers*).

For each aggregate attribute, decision rules were defined so that all the combinations of the input attributes' values are mapped into values of the corresponding aggregate attribute. An example ruleset is represented in Fig. 2. Each row of the table specifies the value of the aggregate attribute for one combination of input attributes values. In this way, each row can be interpreted as an if-then rule.



	Quality	Relevance	Evaluation of Researcher
	71%	29%	
1	Very Low	<=Medium	Unsatisfactory
2	Very Low	*Very High*	Unsatisfactory
3	<=Low	High	Satisfactory
4	Low	<=High	Satisfactory
5	Low	*Very High*	Good
6	Medium	<=High	Good
7	Medium:High	Low	Good
8	Medium	*Very High*	Very good
9	High	Medium:High	Very good
10	>=High	Medium	Very good
11	*Very High*	<=Medium	Very good
12	>=High	*Very High*	*Excellent*
13	*Very High*	>=High	*Excellent*

Figure 2: The topmost decision rules

The decision rules were not specified explicitly in tabular form, but rather implicitly by specifying the weight of the input attributes. As explained in Section 2, decision rules in tabular form are derived from the weights.

In fact, we developed five variants of the model, with slight variations of the tree structure and decision rules (attribute weights). The models M1, M1k, and M1a have ten, while M2 and M2a have 11 inputs (*Impact* decomposes into *Norm. num. citations* and *Num. cited Papers*). The (global) attribute weights for the models are given in Fig. 3, where the basic attributes are in bold and the difference in the tree structure are given in italics.

In all models, the contributions of *Quality* and *Relevance* to the overall evaluation are 75 % and 25 %, respectively. In the first model, M1, the local weights of the *Impact* and *Productivity* attributes to *Quality* are 70 % and 30 %. These are approximately translated into global weights of 48 % and 27 %, respectively (summing to 75 %). The 70–30 % local weights of *Impact* and *Productivity* have been changed to 60–40 % in the model M1a and to 50–50 % in the model M2a.

The 25 % global weight of *Relevance* is divided into 17 % for *Projects* (9 % for *National* and 8 % for *EU*) and (approx.) 8 % for *Other* (of which 6 % for *SU*, 1 % for *Prizes and awards* and 1 % for *Membership*). We have 70–30 % local weight of *SU* and *Indicators of esteem* to *Other*, except in M1k, where we have equal local weights (50–50 %).

As evident from the discussion above, both local and global attribute weights are defined. The local weights always refer to a single aggregate attribute, so the sum of the weights of the immediate descendants of an aggregate attribute is 100 %. Global weights, on the other hand, take into account the structure of the evaluation model and relative importance of aggregate attributes. The sum of the global weights of all input attributes (up to rounding errors) is also equal to 100 %. At the topmost level of the hierarchy, local weights are equal to global weights.

Evaluation	**Mk1**	**M1**	**M1a**	**M2**	**M2a**
Quality	75	75	75	75	75
Productivity	27	27	36	27	43
Journal	17	17	23	17	27
Indexed	**14**	**14**	**18**	**14**	**22**
Other	**3**	**3**	**5**	**3**	**5**
Non-journal	10	10	14	10	17
Conference	**5**	**5**	**7**	**5**	**8**
Monographs	**5**	**5**	**7**	**5**	**8**
Impact	**48**	**48**	**39**	48	32
Norm. Num citations	-	-	-	***34***	***23***
Num. Cited papers	-	-	-	***14***	***9***
Relevance	25	25	25	25	25
Projects	17	17	17	17	17
National	**9**	**9**	**9**	**9**	**9**
EU	**9**	**8**	**8**	**8**	**8**
Other	7	7	7	7	7
SU	**5**	**6**	**6**	**6**	**6**
Indicators of esteem	3	2	2	2	2
Prizes & Awards	**1**	**1**	**1**	**1**	**1**
Membership	**1**	**1**	**1**	**1**	**1**

Figure 3: Global attribute weights in the evaluation models

5 Using and evaluating the models

In this section, we discuss how to use the developed model to evaluate researchers and illustrate its use on two sets of researchers from Slovenia. We first describe the data used, i.e., the two sets of researchers and the procedures used to discretize the input attributes. We then present a graphical description of the evaluation of four researchers. Finally, we analyze the distribution of the evaluation grades for the two sets of researchers for the several variants of the evaluation model that we have developed (the five variants, M1, M1a, M1k, M2, and M2a are described in Section 4).

5.1. Data, discretization and example use

The data we used to illustrate the use of our model(s) and analyze their behaviour concerned two batches of data on Slovenian researchers. The data were extracted by querying the COBISS database that maintains detailed data about the work of researchers in Slovenia [6]. The extracted data covered the performance of researchers in the time interval 2002–2006, which was consistent with several current ARRS's calls at the time of preparing this publication. The smaller dataset included 14 researchers, 12 of which were from the area of computer science, while the larger comprised 171 researchers, all from computer science.

The original data did not contain information on all basic attributes used in our models, hence there are undefined values for the basic attributes that are descendants of the *Indicators of esteem* attribute in all models. Actually, we used the feature of DEXi, which allows for the values of input attributes to be incompletely defined or undefined altogether. In this context completely defined means that a specific single value from the corresponding scale is given, incompletely defined means that a range of values is given instead of a specific one, and undefined means that the entire range of possible values is allowed for the attribute.

The input values of the basic attributes of the evaluated researchers (the smaller dataset is given in Fig. 4) were obtained by applying discretization to the continuous values space across which all basic attributes were initially defined. Discretization was applied to all input attributes, except for *Prizes and Awards* and *Membership*. Two different discretization approaches were taken.

The first discretization approach used a threshold. Above the threshold, the best qualitative value (High) was assigned. The interval below the threshold was divided into three equal subintervals, the first of which was also mapped to High, the remaining intervals were mapped to Medium and Low, correspondingly. As a threshold, we used the performance of the top 1 % of the 171 researchers in the first case and the top 10 % of the 14 researchers in the second case (This means the top two researchers). Ranking and discretization was performed for each indicator separately.

The second approach to discretization was based on calculating percentiles. Values belonging to the interval between the 25th and 75th percentile were classified as “Medium”. Values below the 25th percentile were classified as “Low” and values above the 75th percentile as “High”.

We have applied the developed model (and its variants) to the researchers from the two datasets. Let us first look at an illustration of using the basic model (M1) on four researchers from the smaller dataset. The evaluations of researchers X3, X6,Y1 and Y2 from the table in Fig 4 are depicted in graphical form in Fig 5.

The four dimensions depict the values of the attributes *Relevance, Impact, Journal publications,* and *Non-journal publications.* These are given for each of the researchers, together with the overall evaluation. We can see, for example, that X3 is evaluated as Satisfactory primarily because of the low impact. On the other hand, Y1 is evaluated as very good because of the high relevance, (very) high publications and medium impact. We can visually asses that Y1 is evaluated better than X6 as the corresponding rectangle for X6 is subsumed by the rectangle for Y1.

5.2 Distribution of the evaluation outcomes produced by the different DEXi models

After the illustrative use of the basic DEXi model on some of the researchers from the smaller set, we systematically applied the different variants of the model to both sets of researchers. The goal we had in mind was to study the similarities and differences of the different models. In particular, we compare the models in terms of the distribution of their outcomes, i.e., percentage of researchers obtaining each of the evaluation grades.

For the set of 14 researchers, we have evaluated each researcher with all the five variations of the model. By default, percentile discretization is used, but we also consider linear discretization with thresholds 1 % for M1 and M2, and 10 % for M1 (M1_Lin1, M2_Lin1, M1_Lin10). The distributions of the model outputs for each of the eight cases are depicted in the first 8 barcharts in Fig. 6.

In the case of 171 researchers, we did not consider the M2 model, due to the missing data about the number of citations per paper that this model takes as input in the calculations of the impact value. Data on *Projects* and *Membership* were also not available: we assumed completely undefined values for chart M1_D and values *Projects* = *Medium*, and *Membership* = *yes* for M1_D_PM. Finally, for the latter case, we also consider linear discretization with thresholds 1 % and 10 %.

The barcharts in Fig. 6 depict the percentage of researchers evaluated as Unsatisfactory to Excellent for each of the models and discretization approaches considered, as outlined above. The percentage evaluated as Unsatisfactory is given at the top, Excellent at the bottom of each chart.

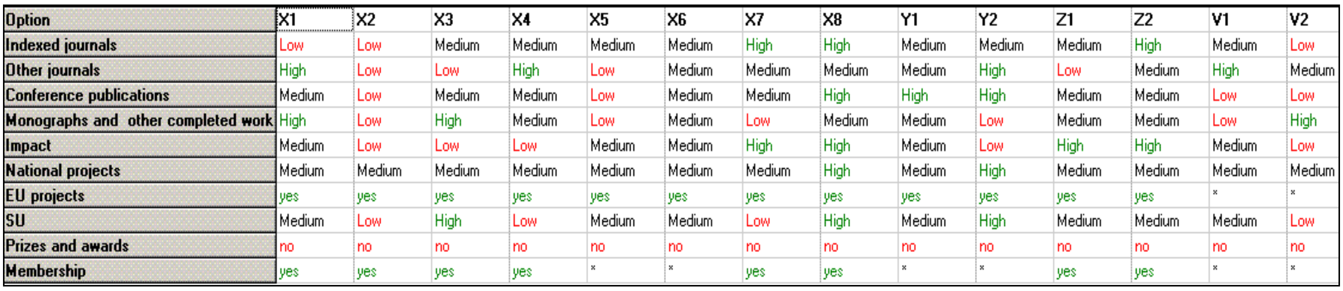
Looking at the different charts, we can draw the following conclusions:

1. *There is a strong connection between the weights and the model output.*

   Decreasing the weight of the *Impact* attribute in favour of the *Productivity* attribute from model M1 to M1a, leads to a larger proportion of Excellent and Very Good evaluations and a smaller proportion of Good evaluations by model M1a (as compared to M1). Model M2 exhibits a similar behaviour: when we decrease the weight of attribute *Impact,* the area covered by classes Very Good and Good shrinks and the area of class Excellent extends in model M2a (as compared to M2).

2. *The model output depends strongly on the discretization applied in the pre-processing phase.*

   As mentioned above, we used two discretization techniques in order to prepare the available data in the appropriate input form for DEXi model evaluation. The barcharts M1_Lin1, M1_Lin10, M2_Lin1, M1_PM_Lin1 and M1_PM_Lin10 concern model evaluations on data obtained



Option	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2	Z1	Z2	V1	V2
Indexed journals	Low	Low	Medium	Medium	Medium	Medium	High	High	Medium	Medium	Medium	High	Medium	Low
Other journals	High	Low	Low	High	Low	Medium	Medium	Medium	Medium	High	Low	Medium	High	Medium
Conference publications	Medium	Low	Medium	Medium	Low	Medium	Medium	High	High	High	Medium	Medium	Low	Low
Monographs and other completed work	High	Low	High	Medium	Low	Medium	Low	Medium	Medium	Low	Medium	Medium	Low	High
Impact	Medium	Low	Low	Low	Medium	Medium	High	High	Medium	Low	High	High	Medium	Low
National projects	Medium	Medium	Medium	Medium	Medium	Medium	Medium	High	Medium	High	Medium	Medium	Medium	Medium
EU projects	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	*	*
SU	Medium	Low	High	Low	Medium	Medium	Low	High	Medium	High	Medium	Medium	Medium	Low
Prizes and awards	no	no	no	no	no	no	no	no	no	no	no	no	no	no
Membership	yes	yes	yes	yes	*	*	yes	yes	*	*	yes	yes	*	*

Figure 4: Input values of 14 evaluated researchers obtained by percentile discretization

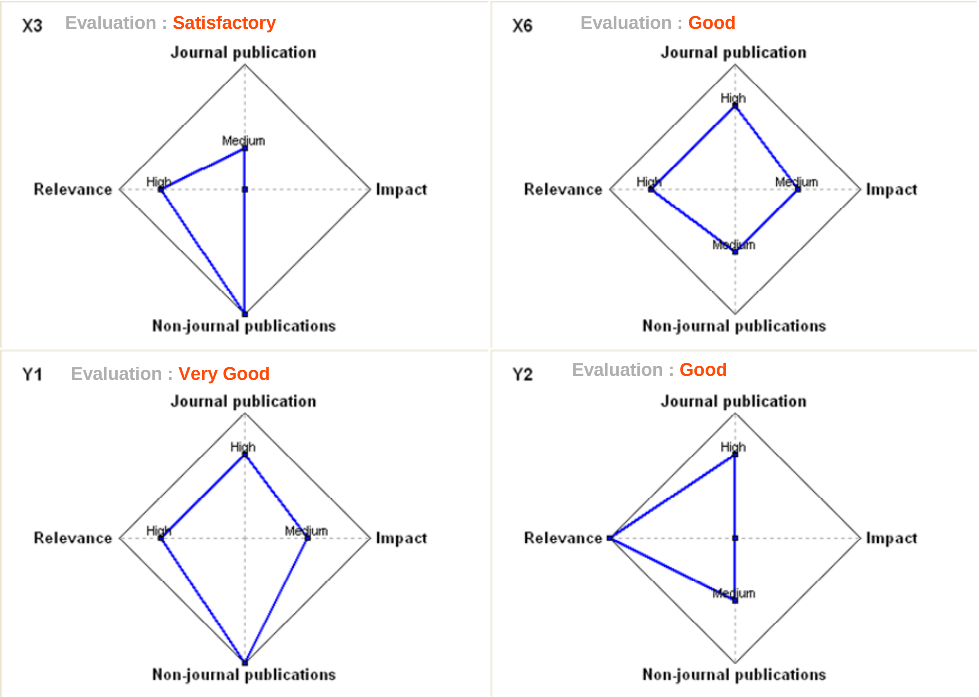


Figure 5: Chart illustration of M1 model evaluation of four researcher, using percentile based discretized data

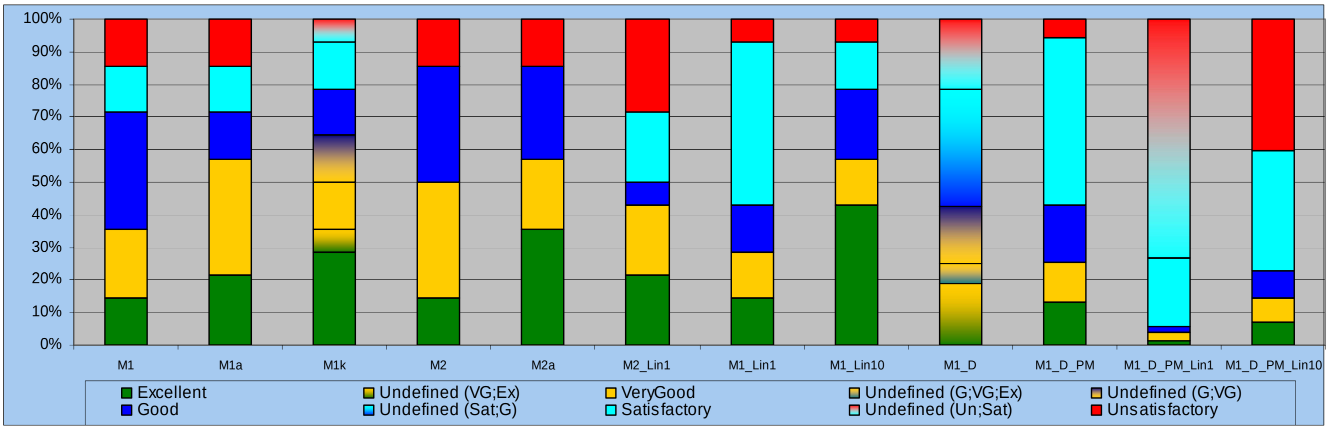


Figure 6: Distribution of evaluation results from different DEXi models, percent of classified researchers per model per class of research

by equidistant interval discretization (with thresholds 1 % and 10 % respectively). The other barcharts concern model evaluations performed on data obtained by percentile based discretizations.

Overall, percentile based discretization leads to a more balanced distribution of the researchers, with most of researchers classified in the classes Good and Very Good (M1, M1a, and M2) and Satisfactory (M1_D_PM). When equidistant interval dicretization is applied, an imbalance is visible in favour of classes Unsatisfactory (M2_Lin1, M1_D_PM_Lin10) and Excellent (M2_Lin1 and M1_Lin10).

3. *The choice of the threshold value in equidistant interval discretization can exhibit significant influence on the model behaviour.*

   Using a threshold of 10 % instead of 1 % in model M1 leads to enormous shrinkage of class Satisfactory in favour of class Excellent (M1_Lin10 vs. M1_Lin1) for the small set of researchers. A similar behaviour is visible for the larger set of researchers (comparing the barcharts M1_D_PM_Lin1 and M1_D_PM_Lin10), where most of the researchers are classified as Unsatisfactory (40 %) and Satisfactory (~35 %).

The above indicates that the evaluations produced by the model are highly sensitive to the (relative) weights given to the individual attributes and to the discretization of the continuous input variables. Some of the sensitivity could be avoided by using qualitative values of the input variables directly. Such values could be derived, e.g., in a peer review process where reviewers evaluate each indicator on a qualitative scale.

The fact that the evaluations produced by the model are sensitive to the weights of attributes and to the discretization procedures also means that we can adapt the model to meet different goals without changing its structure.

Based on the overall evaluation goals and the considered field of research, we can select appropriate values for the weights and discretization approaches. For example, if we want to be strict and evaluate as excellent only a few examples of the whole population of researchers being evaluated (e.g., in the case of very limited funding), we can use model M1, or even the more strict model M1 with a linear discretization scale. On the other hand, if we want to select a larger subset for funding, we can select a model like M2 that classifies most researchers as Very Good or Good. This depends also on the set of researchers at hand, and the model can be tuned for a given set of evaluated researchers.

6 Discussion and conclusions

We have developed a hierarchical multi-attribute model for evaluating the performance of researchers and applied it to two sets of computer science researchers in Slovenia. In contrast to the current approach taken by the Slovenian Research Agency, which is quantitative and calculates a weighted sum of performance indicators, our model is qualitative and combines indicators in a sounder manner. Namely, in the case of summation we can get very high overall scores, even with very low scores along some dimensions, which is not desirable.

The model we have constructed encompasses knowledge from a wide range of studies carried out in the literature. These include researcher evaluation methods from several countries, such as the United Kingdom, the Netherlands and Australia. It is based on performance indicators that are also used in these countries.

The model that we propose can be further developed and evaluated along a number of dimensions. We have currently used weights to specify the decision rules for aggregating attributes. The intended use of weights is to provide initial rules that are reviewed and modified by a decision analyst; this was not done in our case. Manual development of decision rules is a worthwhile investment that would clearly distinguish the proposed model from quantitative linear models.

In addition, the decision support framework in which we have implemented the model has many other desirable properties. It produces evaluations for each of the intermediate levels of evaluation (such as *Quality* or *Relevance*) and provides explanations at several levels of detail. It also produces several graphical representations of the evaluations.

The proposed models are a possible alternative to the model used by ARRS and we hope it will motivate further discussion on this important topic in Slovenia.

References

[1] Scriven, M. (2001). An overview of evaluation theories. Evaluation Journal of Australasia, 1(2), 27-29.

[2] Frederiksen, L. F., Hannson, F., & Wennberg, S. B. (2003). The Agora and the role of research evaluation. Evaluation: The International Journal of Theory, Research and Practice, 9(2), 149-172.

[3] Hansson, F. (2006). Organizational use of evaluation:Governance and control in research evaluation. Evaluation: The International Journal of Theory, Research and Practice, 12 (2), 159-178.

[4] SICRIS (2007). Slovenian Current Research Information System. http://sicris.izum.si/

[5] ARRS (2007). Slovenian Research Agency. http://www.arrs.si

[6] COBISS (2007). Co-operative Online Bibliographic System & Services. http://www.cobiss.si/

[7] Coryn, C. L. S., & Hattie, J. A. (2006). The transdisciplinary model of evaluation. Journal of MultiDisciplinary Evaluation, 4, 107-114.

[8] Quantitative indicators for research assessment-a literature review

[9] Web of Science, *http://scientific.thomson.com/-products/wos/*

[10] Bohanec, M. (2003). Decision support. In: D. Mladenić, N. Lavrač, M. Bohanec, S. Moyle (Editors), Data mining and decision support: Integration and collaboration. Kluwer Academic Publishers, 23–35.

[11] Bohanec, M. (2007). DEXi: Program for Multi-Attribute Decision Making, User's Manual, Version

2.00. IJS Report DP-9596, Jožef Stefan Institute, Ljubljana, 2007.
http://www-ai.ijs.si/MarkoBohanec/dexi.html

[12] Bouyssou, D., Marchant, T., Pirlot, M., Tsoukias, A., Vincke, P. (2006). Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst. Springer.

[13] Bohanec, M., Zupan, B. (2004). A function-decomposition method for development of hierarchical multi-attribute decision models. Decision Support Systems 36, 215–233.