Group Decis Negot
DOI 10.1007/s10726-016-9494-6

# A Robustness Study of State-of-the-Art Surrogate Weights for MCDM

**Mats Danielson**[1,2] · **Love Ekenberg**[1,2]



**Abstract** A vast number of methods for solving multi-criteria decision problems have been suggested for assessing criteria weights requiring more exact input data than users normally are able to provide. In particular, the selection of adequate criteria weights is difficult and in order to be realistic, other methods must be introduced. One class of such methods is to introduce so called surrogate weights, where numerical weights are assigned to each criterion based on a cardinal or ordinal rank ordering, assumed to represent the information extracted from the user. One essential problem is the robustness of such methods. In this article, we compare state-of-the-art methods based on surrogate weights from the literature and, utilising a simulation approach, discuss underlying assumptions and robustness properties. This results in a quantitative measurement of these weighting methods and a methodology applicable also to forthcoming methods.

**Keywords** Multi-criteria decision analysis · Surrogate criteria weights · Robustness · Criteria ranking · Rank order

✉ Love Ekenberg
lovek@dsv.su.se; ekenberg@iiasa.ac.at

Mats Danielson
mats.danielson@su.se

[1] Dept. of Computer and Systems Sciences, Stockholm University, Postbox 7003, 164 07 Kista, Sweden

[2] International Institute for Applied Systems Analysis, IIASA, Schlossplatz 1, 2361 Laxenburg, Austria

Published online: 19 July 2016

## 1 Introduction

In multi-criteria decision analysis (MCDM), the most common underlying measurement mechanism is Multi-Attribute (Value or) Utility Theory (MAVT / MAUT). Within MAUT, a common form of evaluation function is the additive model $V(a) = \sum_{i=1}^{m} w_i v_i(a)$, where $V(a)$ is the overall value of alternative $a$, $v_i(a)$ is the value of the alternative under criterion $i$, and $w_i$ is the weight of this criterion. One of the problems with the additive model as well as other models is that in real-life decision making, numerically precise information is seldom available, and when it comes to providing reasonable weights for the criteria, most decision-makers experience difficulties due to most humans seemingly do not have the required granulation capacity and also suffer from other cognitive deficiencies pertinent to the specification of a decision problem. To somewhat facilitate eliciting weights from decision-makers, some of the approaches in the literature utilise ordinal or imprecise importance information to determine criteria weights and sometimes values of alternatives. Other approaches instead make use of surrogate weights which represent the most likely interpretation of the preferences expressed by a decision-maker or a group of decision-makers. This paper deals with the latter approach to eliciting preferences or importance information.

However, it is not obvious how to determine the decision quality of a multi-criteria surrogate weighting method. Methods were mostly assessed in case studies until Barron and Barrett (1996a) introduced a process utilising systematic simulations. The basic idea is to generate surrogate weights as well as "true" reference weights from some underlying distribution and investigate how well the result of using surrogate numbers match the result of using the "true" results. The idea in itself is good, but the methodology is vulnerable since the validation result is heavily dependent on the distribution used for generating the weight vectors. Barron and Barrett themselves 1996a argue that the elicitation of exact weights demands an exactness which does not exist in the mind of the decision-maker, and already von Winterfeldt and Edwards (1986) claim that "*the precision of numbers is illusory*". And, for example, ratio weight procedures can be difficult to accurately employ due to response errors (Jia et al. 1998). The common lack of reasonably complete information increases this problem significantly. Several attempts have been made to resolve this issue. Methods allowing for less demanding ways of ordering the criteria, such as ordinal rankings or interval approaches for determining criteria weights and values of alternatives, have been suggested. The idea is, as far as possible, not to force decision-makers to express unrealistic, misleading, or meaningless statements, but at the same time being able to utilise the information the decision-maker is able to supply. An approach of this type is to use surrogate weights, which are derived from ordinal importance information (Barron and Barrett 1996a, b; Katsikopoulos and Fasolo 2006). In such methods, the decision-maker provides information on the rank order of the criteria, i.e. supplies ordinal information on importance. Thereafter, this information is converted into numerical weights consistent with the extracted ordinal information. Several proposals on how to convert the rankings into numerical weights exist in the literature, e.g., rank sum (RS) weights and rank reciprocal (RR) weights (Stillwell et al. 1981), and centroid (ROC) weights (Barron 1992). However, the use of only ordinal information

is often perceived as being too vague or imprecise, resulting in a lack of confidence in the alternatives' final rankings.

Furthermore, it is not obvious how "correct" a surrogate weight method is, since the "real" weights are unknown or even inexistent (in some objective sense). The decision quality of a method was at first mostly assessed in case studies until (Barron and Barrett 1996a) introduced a process utilising systematic simulations. The basic idea is to generate surrogate weights as well as "true" reference weights from some underlying distribution and investigate how well the result of using surrogate numbers match the result of using the "true" numbers. The idea is good, but is nevertheless vulnerable since the validation result is heavily dependent on the distribution used for generating the weight vectors.

In this article, we discuss a spectrum of methods for increasing the expressive power of user statements, with a particular aim at how the weight function(s) still can be reasonably elicited while preserving the comparative simplicity and correctness of ranking approaches. Below we discuss and compare some important aspects of robustness of a set of ranking methods for weights as well as their relevance and correctness. After having briefly recapitulated some ordinal ranking methods in the Sect. 2, we continue with state-of-the-art ranking methods taking strength into account and discuss a spectrum of interesting candidates as well as cognitive models of decision-makers. Thereafter, using simulations, we investigate robustness properties of the methods and conclude with pointing out, according to the results, a particularly attractive method for weight elicitation.

## 2 Ordinal Ranking Methods

In multi-criteria decision making (MCDM), different elicitation formalisms have been proposed by which a decision-maker can express preferences. Such formalisms are sometimes based on scoring points, as in point allocation (PA) or direct rating (DR) methods.[1] In PA, the decision-maker is given a point sum, e.g. 100, to distribute among the criteria. Sometimes, it is pictured as putty with the total mass of 100 being divided and put on the criteria. The more mass, the larger weight on a criterion, and the more important it is. When the first $N - 1$ criteria have received their weights, the last criterion's weight is automatically determined as the remaining mass. Thus, in PA, there is $N - 1$ degrees of freedom (DoF) for $N$ criteria. DR, on the other hand, puts no limit to the total number of points to be allocated. The decision-maker allocates as many points as desired to each criterion. The points are subsequently normalized by dividing by the sum of points allocated. When the first $N - 1$ criteria have received their weights, the last criterion's weight still has to be assigned by the decision-maker. Thus, in DR, there are $N$ degrees of freedom for $N$ criteria. Regardless of elicitation method, the assumption is that all elicitation is made relative to a weight distribution held by the decision-maker.[2]

[1] PA and DR are akin to elements of the SAW approach (Danielson and Ekenberg 2007).

[2] For various cognitive and methodological aspects of imprecision in decision making, see, e.g., Danielson and Ekenberg (2007), Danielson et al. (2013) and other papers by the same authors.

One very early idea in MCDM was to just skip the criteria elicitation and assign equal weights to every criterion, but the information loss is then very large. It is therefore worthwhile to at least rank the criteria when applicable, since rankings are normally easier to provide than precise numbers. From the ranking, so called surrogate weights can then be derived. This technique is utilised in Barron and Barrett (1996a, b), Katsikopoulos and Fasolo (2006), and many others. Needless to say, for practical decision making, surrogate weights can sometimes be perceived as a peculiar way of motivating a method. Nevertheless, validation in this field is very difficult, due to difficulties regarding elicitation, and the surrogate methods are quite widely used and can be considered as attempts of trying to motivate the various generation methods. The crucial issue is then rather how to assign surrogate weights while losing as little information as possible and preserving the "correctness" when assigning the weights. Stillwell et al. (1981) discuss the weight approximation techniques rank sum (RS) and rank reciprocal (RR) weights. They are suggested in the context of maximum discrimination power, and are both alternatives to ratio based weight schemes. The rank sum is based on the idea that the rank order should be reflected directly in the weights. For a set of $N$ criteria weights ($i = 1, \ldots, N$) assume a simplex $S_w$ generated by $w_1 > w_2 > \ldots > \quad w_N, \Sigma w_i = 1$ and $0 \leq w_i$. Assign an ordinal number to each item ranked, starting with the highest ranked item as number 1. Denote the ranking number $i$ among $N$ items to rank. Then the RS weight (Eq. 1) for all $i = 1, \ldots, N$ becomes

$$w_i^{\mathrm{RS}} = \frac{N + 1 - i}{\sum_{j=1}^{N} (N + 1 - j)} \tag{1}$$

Another idea, also discussed in Stillwell et al. (1981) is rank reciprocal weights. They have a similar origin as RS weights, but are based on the reciprocals (inverted numbers) of the rank order for each item ranked. These are obtained by assigning an ordinal number to each item ranked, starting with the highest ranked item as number 1. Denote the ranking number $i$ among $N$ items to rank. Then the rank reciprocal (RR, Eq. 2) weight becomes

$$w_i^{\mathrm{RR}} = \frac{1/i}{\sum_{j=1}^{N} \frac{1}{j}} \tag{2}$$

A decade later, Barron Barron (1992) suggested a weight method based on vertices of the simplex of the feasible weight space. The ROC (rank order centroid) weights are the centroid vector components of the simplex $S_w$. That is, ROC is a function based on the average of the corners in the polytope defined by the simplex $S_w = w_1 > w_2 > \ldots > w_N, \Sigma w_i = 1$, and $0 \leq w_i$. The weights then become the centroid (mass point) of $S_w$. The ROC weights for the ranking number $i$ among $N$ items to rank are given by Eq. 3.

$$w_i^{\mathrm{ROC}} = 1/N \sum_{j=i}^{N} \frac{1}{j} \tag{3}$$

Examining the weights, ROC resembles RR more than RS but is, particularly for lower dimensions, more extreme than both in the sense of weight distribution, especially for the largest and smallest weights.

As discussed in Danielson and Ekenberg (2014), RS, RR, and ROC perform well only for specific assumptions on decision-maker behaviour. If we assume that the decision-maker in his/her mind stores his/her criteria preferences in a way similar to a given point sum, for example pictured as putty with the fixed total mass, there are consequently $N-1$ degrees of freedom (DoF) for $N$ criteria. On the other hand, if we assume that the decision-maker stores his/her criteria preferences in a way that puts no limit to the total number of points (or mass) allocated, then there are $N$ degrees of freedom for $N$ criteria. Those two models of decision-maker behaviour yield very different results in assessing surrogate weights. The RS weight model is tailored to the assumption of $N$ degrees of freedom and the RR and ROC models are tailored to the $N-1$ DoF assumption. Since the models RS and RR are, in this sense, opposites, and in reality the preferences are reasonably stored in either one of the above ways or somewhere in between, a weight function combining the properties of RS and RR was proposed in Danielson and Ekenberg (2014). The SR weight method is an additive combination of Sum and Reciprocal weight functions as shown in Eq. 4.

$$w_i^{\mathrm{SR}} = \frac{1/i + \frac{N+1-i}{N}}{\sum_{j=1}^{N}\left(1/j + \frac{N+1-j}{N}\right)} \tag{4}$$

In our previous work Danielson and Ekenberg (2014), we carried out a set of simulations of the above ordinal methods and confirmed some previous results as well as discussed some new results regarding a mixed model of decision-maker behaviour that takes into account the different possible degrees of freedom available. Of the above methods in this section, SR was found to be the most robust and will, together with ROC, be used as references in the following comparative study.

## 3 Preference Strength Ranking Methods

Providing ordinal rankings of criteria seems to avoid some of the difficulties associated with the elicitation of exact numbers. It puts fewer demands on decision-makers and is thus, in a sense, effort-saving. Furthermore, there are techniques such as those above for handling ordinal rankings with some success. However, decision-makers might in many cases have more knowledge of the decision situation, even if the information is not precise. For instance, importance relation information containing strengths may implicitly exist.[3] However, these cannot be taken into account in the transformation of an ordinal rank order into weights. This entails that the surrogate weights may not really reflect what the decision-maker actually means by his/her ranking. Some form of strengths often exist and this information should reasonably be used when transforming orderings into weights to utilise all the information the decision-maker is able to supply. Below, we will therefore investigate whether the above (ordinal) methods can be successfully extended to accommodate some information regarding

[3] For example, for three criteria A, B and C: "A is slightly more important than B while B is vastly more important than C" must, in an ordinal ranking, be expressed as "A is more important than B which is more important than C".

relational strengths as well, i.e. to handle ordinal information together with strength relations information, while still preserving the property of being less demanding and more practically useful than other types of methods. The idea is that instead of using a predetermined conversion method (as in, e.g., ROC weights) to obtain surrogate weights from an ordinal criteria ranking, the decision-maker will be able to express and utilise known differences in importance between the criteria.

### 3.1 Preference Strength

Assume that there exists an ordinal ranking of $N$ criteria. In order to make this order into a stronger ranking, information should be given about how much more or less important the criteria are compared to each other. Such rankings also take care of the problem with ordinal methods of handling criteria that are found to be equally important, i.e. resisting pure ordinal ranking. In this paper, we will use the following notations for the strength of the rankings between criteria as well as some suggestions for a verbal interpretation of these:

$>_0$ equally important
$>_1$ slightly more important
$>_2$ more important (clearly more important)
$>_3$ much more important

While being more cognitively demanding than ordinal weights, they are still less demanding than, for example, AHP weight ratios (usually employing nine ratios, i.e. 1/9, 1/7, 1/5, 1/3, 1, 3, 5, 7, and 9) or point scores like SMART (usually employing several integers). In an analogous manner as for ordinal rankings, the decision-maker statements can be converted into weights.

### 3.2 Weights of Preference Strength

In analogy with the ordinal weight functions above, counterparts using the concept of preference strength can straightforwardly be derived.

1. Assign an ordinal number to each importance scale position, starting with the most important position as number 1.
2. Let the total number of importance scale positions be $Q$. Each criterion $i$ has the position $p(i) \in \{1, \ldots, Q\}$ on this importance scale, such that for every two adjacent criteria $c_i$ and $c_{i+1}$, whenever $c_i >_{s_i} c_{i+1}$, $s_i = |p(i+1) - p(i)|$. The position $p(i)$ then denotes the importance as stated by the decision-maker. Thus, $Q$ is equal to $\Sigma s_i + 1$, where $i = 1, \ldots, N-1$ for $N$ criteria.

Then the cardinal counterparts to the ordinal ranking methods above can be found as follows. To begin with, we consider the counterpart to RS weights (Stillwell et al. 1981). The concept of cardinal rank sum (CRS) weights is based on the idea that the rank order strength should be reflected directly in the weights. Then the CRS weights are obtained by Eq. 5

$$w_i^{\mathrm{CRS}} = \frac{Q + 1 - p\,(i)}{\sum_{j=1}^{N} (Q + 1 - p\,(j))}, \tag{5}$$

based on the importance positions $p(i)$as stated by the decision-maker. The counterpart to ordinal rank reciprocal weights (Stillwell et al. 1981) is analogously defined. According to step 2, let the total number of importance scale positions be $Q$. Each criterion $i$ has the position $p(i)$ on the importance scale such that $p\,(i)\;p\,(j)$ if $i < j$. Then the corresponding rank reciprocal (CRR) weights are obtained by Eq. 6

$$w_i^{\mathrm{CRR}} = \frac{\frac{1}{p(i)}}{\sum_{j=1}^{N} \frac{1}{p(j)}} \tag{6}$$

with the usual property that a higher weight is assigned to lower ranking numbers. ROC weights (Barron 1992) are generalised in the same way. The ordinal ROC weights, given by Eq. 3 in Sect. 2

$$w_i^{\mathrm{ROC}} = 1/N \sum_{j=i}^{N} \frac{1}{j} \tag{7}$$

could be interpreted as candidate weights for positions on the importance scale. Then the corresponding preference strength rank order centroid weights (CRC, Eq. 7) are obtained as

$$w_i^{\mathrm{CRC}} = \frac{\sum_{j=p(i)}^{Q} \frac{1}{\mathrm{j}}}{\sum_{k=1}^{N} \left( \sum_{j=p(k)}^{Q} \frac{1}{\mathrm{j}} \right)} \tag{8}$$

Finally, generalising the SR weights (Danielson and Ekenberg 2014) is done in the same way. The ordinal SR weights were given by the Eq. 4

$$w_i^{\mathrm{SR}} = \frac{1/i + \frac{N+1-i}{N}}{\sum_{j=1}^{N} w_j^{\mathrm{SR}}} \tag{9}$$

which will now be interpreted as candidate weights for positions on the importance scale. Using steps 1–3 above, the corresponding preference strength SR weights (CSR, Eq. 8) are obtained as

$$w_i^{\mathrm{CSR}} = \frac{1/p\,(i) + \frac{Q+1-p(i)}{Q}}{\sum_{j=1}^{N} \left( 1/p\,(j) + \frac{Q+1-p(j)}{Q} \right)} \tag{10}$$

which is a similar generalisation as the other weights. Thus, using the idea of importance steps, ordinal weight methods are easily generalised to their respective counterparts. Having obtained weights for preference strength relationships, we now proceed by assessing them together with ordinal weights.

Another class of MCDM methods is the ELECTRE family of methods. In that context, Simos proposed a simple procedure, using a set of cards, trying to indirectly determine numerical values for criteria weights (Simos 1990a, b). The Simos method is, however, a bit different from the methods discussed above. It is a relatively simple method for easily expressing criteria hierarchies while introducing some cardinality if

needed. It has been widely applied and has been well-received by real decision-makers. When applying this method, a group of decision-makers are provided with a set of coloured cards with the criteria names written on them. Furthermore, the decision-makers are provided with a set of white (blank) cards. Thereafter, the non-blank cards are ranked from the least important to the most important, where criteria of equal importance are grouped together. Furthermore, the decision-makers are asked to place the white cards in between the coloured cards to express preference strengths. Then the surrogate numbers can be computed. A constant value difference, $u$, between two consecutive cards is here assumed. A white card between two consecutive coloured ones means a difference of $2 \cdot u$ and two white cards means a difference of $3 \cdot u$, etc. The normalised surrogate weights are then determined from this ordering. This method is referred to as S1 in the assessment in Sect. 4. One problem with the Simos method is that it is not robust when the preferences are changed (Scharlig 1996) and that it has some other contra-intuitive features, such as that it only picks one of the weight vectors satisfying the model, while there can of course be an infinite number of them. Furthermore, because of the weights being determined differently depending on the number of cards in the subsets of equally ranked cards, the differences between the weights also change in an uncontrolled way when the cards are reordered. This is why Figueira and Roy (2002) suggested a revised version, where a more robust proportionality when using these white cards is provided. This is accomplished by requesting the decision-makers to state how many times more important the most important criterion or criteria group is compared to the least important. This addition seemingly solves some problems, but introduces the complication to require the decision-maker to reliably and correctly estimate a proportional factor $z$ between the largest and the smallest criteria weights. The revised method is referred to as S2 in the assessment below.

## 4 Generalised Assessment of Models for Weights

Given that we have a set of methods as in the previous section, how can they be validated? For ordinal weights, simulation studies similar to Barron and Barrett (1996a), Arbel and Vargas (1993), Stewart (1993), Ahn and Park (2008), and others have become a kind of de facto standard for comparing multi-criteria surrogate weight methods. The underlying assumption of most studies is that there exist a set of 'true' weights in the decision-maker's mind which are inaccessible in its pure form by any elicitation method. We will utilise the same technique for determining the efficacy, in this sense, of the ranking approaches suggested above. The modelling assumptions regarding decision-makers' mind-sets we discussed above are mirrored in the genera-tion of decision problem vectors by a random generator. Thus, following an $N - 1$ DoF model, a vector is generated in which the components sum to 100 %, i.e., a process with $N - 1$ degrees of freedom. Following an $N$ DoF model, a vector is generated keeping components within [0, 100 %] and subsequently normalising, i.e., a process with $N$ degrees of freedom. Other distributions modelling actual decision-makers would of course be possible, and could maybe be elicited in one way or another. However, this is not the main point herein. The important observation is that these validation meth-

ods are highly dependent of the model of decision-makers and this yields significant effects on the reliability of the validations. The degree of freedom is only one type of dichotomy, but one actually expressing a meaningful semantics for discriminating cognitive models in this respect.

When following an $N-1$ DoF model, a vector is generated in which the components sum to 100 %. This simulation is based on a homogenous $N$-variate Dirichlet distribution generator. Details on this kind of simulation can be found, e.g., in Rao and Sobel (1980). On the other hand, following an $N$ DoF model, a vector is generated without an initial joint restriction, only keeping components within [0, 100 %] yielding a process with $N$ degrees of freedom. Subsequently, they are normalised so that their sum is 100 %. Details on this kind of simulation can be found, e.g., in Roberts and Goodwin (2002). We will call the $N-1$ DoF model type of generator an $N-1$-*generator* and the $N$ DoF model type an *N-generator*. Depending of the simulation model used (and consequently the background assumption of how decision-makers assess weights), the results become very different. For instance, ROC weights in $N$ dimensions coincide with the mass point for the vectors of the $N-1$-generator over the polytope $S_w$. In our earlier work Danielson and Ekenberg (2014), the close relationships between ROC weights and the $N-1$-generator as well as between RS weights and the $N$-generator were discussed, and we concluded that the choice of degrees of freedom for the random number generator significantly affects the results.

In reality, though, we cannot know whether a specific decision-maker (or even decision-makers in general) adhere more to $N-1$ or $N$ DoF representations of their preferences. Both as individuals and as a group, they might use either or be anywhere in between. A *robust* rank ordering mechanism (in a reasonable sense) must therefore employ a surrogate weight function that handles both styles of representation and anything in between. Thus, the evaluation of surrogate weights in this paper will use both types of generators and combinations thereof to find the most efficient and robust weights.

Barron and Barrett (1996a) compared RS, RR, and ROC, where the idea was to measure the validity of the method by simulating a large set of scenarios utilising surrogate weights and see how well different methods provided results similar to scenarios utilising "true" weights. Again, note that the notion of a "true" weight is dependent on the decision-maker model. The Barron and Barrett study obviously assumes an $N-1$ DoF model and presents a computer simulation consisting of four steps, assuming the problem is modelled as the simplex $S_w$.

**Generation Procedure**

1. For an $N$-dimensional problem, generate a random weight vector $\boldsymbol{t}$ with $N$ components. This is called the true weight vector. Determine the order between the weights in the vector $\boldsymbol{t}$. For each method $\mathbf{X}'$, use the order to generate a weight vector $w^{X'}$.
2. Given $M$ alternatives, generate $M \times N$ random values with value $v_{ij}$ belonging to alternative $j$ under criterion $i$.
3. Let $w_i^X$ be the weight from weighting method $\mathbf{X}$ for criterion $i$ (where $\mathbf{X}$ is either $\mathbf{X}'$ or $\boldsymbol{t}$). For each method $\mathbf{X}$, calculate $V_j^X = \sum_i w_i^X v_{ij}$. Each method produces a preferred alternative $\mathrm{A_X}$, i.e. the one with the highest $V_j^X$.

4. For each method $\mathbf{X}'$, assess whether $\mathbf{X}'$ yielded the same decision (i.e. the same preferred alternative $A_X$) as $\boldsymbol{t}$. If so, record a hit.

This is repeated a large number of times (simulation rounds). The hit rate (or frequency) is defined as the number of times a weighting method made the same decision as TRUE. The study also used two other measures of efficacy, average value loss and average proportion of maximum value range achieved. The two latter measures are strongly correlated to the hit ratio and do not add much insight into method performance. The results of the original study in Barron and Barrett (1996a) were that ROC outperformed the other two weighting methods. Of the two other, RR was slightly superior to RS. Since the three methods require equally much input from the decision-maker, the conclusion was made that ROC was to be preferred among the surrogate weights. Using an $N-1$-generator simulation model over the simplex $S_w$, the results of the Barron and Barrett study can easily be verified. However, note again that this distribution favours the ROC method since the centroid of the generated "true" weights is the same as the vector of the corresponding ROC weights.

It should also be noted that most simulation studies to date arrive at the same conclusions regarding ROC, RS, and RR. A study by Roberts and Goodwin (2002), though, came up with a different result where RS performed better than ROC with RR in third place. The random weight distribution is in most other simulations (in step 1 of the generation procedure above) generated by an $N-1$ procedure, thus generating a vector with $N-1$ DoF. Instead, Roberts and Goodwin employ a different distribution generating function where a fixed number, say 100, is given to the most important criterion and the others are uniformly generated as U[0, 100], i.e. an $N$-generator. As explained above, this $N$-generator is not the same as $N-1$-generators based on a Dirichlet distribution and thus, their simulation study instead yields the result that RS outperforms ROC with RR in third place. This is also confirmed in Danielson and Ekenberg (2014), i.e. given an $N$-generator RS outperforms ROC and RR while ROC is marginally better than RR. While yielding a different "best" weighting method, this result is consistent with the other study results considering it is merely a consequence of choice of DoF in the simulator generator. The Simos family of weighting methods have not been previously assessed in this way. In the assessment below, S1 is the original method suggested by Simos (1990a, b). S2 is the revised method from Figueira and Roy (2002) with the additional parameter $z$ estimated in two ways. It is a severe complication for the decision-maker to have to make this estimate and two different approaches are employed in this study. Both approaches are in actual use. In S2A, $z$ is assumed to be a suitable fixed number, in this case 20. In S2B, $z$ is assumed to be proportional to $Q$, the number of steps ('>'-symbols), in this case $Q+1$. There is no other way for the decision-maker to obtain $z$ but to estimate it.

### 4.1 Comparing Weight Methods

Our comparative simulations were carried out with a varying number of criteria and alternatives. There were four numbers of criteria $N = \{3,\ 6,\ 9,\ 12\}$ and five numbers of alternatives $M = \{3,\ 6,\ 9,\ 12,\ 15\}$ creating a total of 20 simulation scenarios. Each scenario was run 10 times, each time with 10,000 trials, yielding a total of 2,000,000

**Table 1** The winner frequency for the methods using an $N-1$ generator

| N − 1 DoF | ROC | SR | CRC | CRS | CRR | CSR | S1 | S2A | S2B |
|---|---|---|---|---|---|---|---|---|---|
| 3\|3 | 90.2 | 89.3 | 92.6 | 93.2 | 92.7 | *93.8* | 93.3 | 91.9 | 93.0 |
| 3\|15 | 79.1 | 76.9 | 82.4 | 84.5 | 82.4 | *85.3* | 83.8 | 80.3 | 83.9 |
| 6\|6 | 84.8 | 83.1 | 88.0 | 85.8 | 83.6 | *88.5* | 85.0 | 87.0 | 85.5 |
| 6\|12 | 81.3 | 78.9 | 85.4 | 82.7 | 79.4 | *85.5* | 81.7 | 83.9 | 82.2 |
| 9\|9 | 83.5 | 81.2 | *86.0* | 80.7 | 78.3 | 85.2 | 79.9 | 82.4 | 80.2 |
| 12\|6 | *86.4* | 84.1 | 86.1 | 81.1 | 78.7 | 85.5 | 80.6 | 81.7 | 81.0 |
| 12\|12 | 83.4 | 80.2 | *83.7* | 77.5 | 74.5 | 82.5 | 75.9 | 78.1 | 77.0 |

**Table 2** The winner frequency for the methods using an $N$ generator

| N DoF | ROC | SR | CRC | CRS | CRR | CSR | S1 | S2A | S2B |
|---|---|---|---|---|---|---|---|---|---|
| 3\|3 | 87.3 | 89.1 | 89.4 | 92.6 | 90.8 | 92.5 | 92.5 | 89.2 | *93.3* |
| 3\|15 | 77.9 | 80.6 | 78.9 | 85.7 | 82.5 | 85.6 | 85.8 | 76.9 | *86.9* |
| 6\|6 | 80.1 | 85.1 | 84.7 | 90.2 | 81.4 | 89.1 | 90.5 | 88.8 | *90.9* |
| 6\|12 | 76.4 | 82.0 | 81.5 | 88.0 | 77.7 | 86.8 | *88.3* | 86.9 | 88.2 |
| 9\|9 | 76.3 | 83.0 | 81.6 | 88.0 | 74.5 | 85.9 | *89.1* | 88.3 | 88.2 |
| 12\|6 | 77.5 | 84.6 | 83.0 | 89.0 | 75.0 | 86.4 | *90.8* | 89.4 | 88.9 |
| 12\|12 | 73.4 | 81.7 | 79.9 | 86.5 | 70.7 | *83.9* | 88.9 | 86.8 | 85.4 |

decision situations generated. An $N$-variate joint Dirichlet distribution was employed to generate the random weight vectors for the $N-1$ DoF simulations and a standard round-robin normalised random weight generator for the $N$ DoF simulations. Similar to Barron and Barrett (1996a), unscaled value vectors were generated uniformly, and no significant differences were observed with other value distributions.[4]

In Table 1,[5] using an $N-1$-generator, it can be seen that all four preference strength methods generally outperform the ordinal ones as expected and CSR is the best one, except for the last three rows, where CRC and ROC, respectively perform the best. This is because the cardinality loses some meaning when the decision situation is denser, and ROC benefits from the type of generator.

The frequencies have changed in Table 2, according to expectations, since we employ a model with $N$ degrees of freedom. Still the preference strength methods perform better than the ordinal ones. S1 and S2 improve and, e.g., CRC generally fares a bit worse. In general, strength methods perform clearly better than ordinal ones.

[4] Success measures we used were (a) "winner", having the same preferred alternative, (b) matching of the three highest ranked alternatives ("podium"), and (c) matching of all ranked alternatives ("overall"), the number of times all evaluated alternatives using a particular method coincide with the true ranking of the alternatives. The two latter sets correlated strongly with the first and are not shown in this paper.

[5] In this and the following tables, the leftmost column contains the notation $N|M$, denoting a decision situation having $N$ criteria and $M$ alternatives.

**Table 3** The winner frequency for the methods using a combined generator

| Combined | ROC | SR | CRC | CRS | CRR | CSR | S1 | S2A | S2B |
|---|---|---|---|---|---|---|---|---|---|
| 3\|3 | 88.8 | 89.2 | 91.0 | 92.9 | 91.8 | *93.2* | 92.9 | 90.6 | *93.2* |
| 3\|15 | 78.5 | 78.8 | 80.7 | 85.1 | 82.5 | *85.5* | 84.8 | 78.6 | 85.4 |
| 6\|6 | 82.5 | 84.1 | 86.4 | 88.0 | 82.5 | *88.8* | 87.8 | 87.9 | 88.2 |
| 6\|12 | 78.9 | 80.5 | 83.5 | 85.4 | 78.6 | *86.2* | 85.0 | 85.4 | 85.2 |
| 9\|9 | 79.9 | 82.1 | 83.8 | 84.4 | 76.4 | *85.6* | 84.5 | 85.4 | 84.2 |
| 12\|6 | 82.0 | 84.4 | 84.6 | 85.1 | 76.9 | *86.0* | 85.7 | 85.6 | 85.0 |
| 12\|12 | 78.4 | 81.0 | 81.8 | 82.0 | 72.6 | *83.2* | 82.4 | 82.5 | 81.2 |

**Table 4** Mean over all simulations

| Total correct | ROC | SR | CRC | CRS | CRR | CSR | S1 | S2A | S2B |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 81.3 | 82.8 | 84.5 | 86.1 | 80.2 | *86.9* | 86.2 | 85.1 | 86.0 |
| Rank | 8 | 7 | 6 | 3 | 9 | 1 | 2 | 5 | 4 |

**Table 5** Spread over different DoF

| Spread | ROC | SR | CRC | CRS | CRR | CSR | S1 | S2A | S2B |
|---|---|---|---|---|---|---|---|---|---|
| 3\|3 | *0.2* | 3.2 | 0.6 | 1.9 | 1.3 | 0.8 | 2.7 | 0.3 | |
| 3\|15 | 1.2 | * 3.7 | 3.5 | 1.2 | *0.1* | 0.3 | 2.0 | 3.4 | 3.0 |
| 6\|6 | 4.7 | 2.0 | 3.3 | 4.4 | 2.2 | *0.6* | 5.5 | 1.8 | 5.4 |
| 6\|12 | 4.9 | 3.1 | 3.9 | 5.3 | 1.7 | *1.3* | 6.6 | 3.0 | 6.0 |
| 9\|9 | 7.2 | 1.8 | 4.4 | 7.3 | 3.8 | *0.7* | 9.2 | 5.9 | 8.0 |
| 12\|6 | 8.9 | *0.5* | 3.1 | 7.9 | 3.7 | 0.9 | 10.2 | 7.7 | 7.9 |
| 12\|12 | 10.0 | 1.5 | 3.8 | 9.0 | 3.8 | *1.4* | 13.0 | 8.7 | 8.4 |

In Table 3, the $N$ and $N-1$ DoF models are combined with equal emphasis on both. Cardinal methods consequently perform better than the ordinal ones and we can see that in total CSR performs the best. S2B still performs reasonable, at least for lesser number of criteria. As expected, it is also clear that the CRC, CRR, and CSR methods outperform the best ordinal methods under varying assumptions of decision-maker weight generation, indicating that the added information is put to good use.

Table 4 shows the average of the respective columns of Table 3. As we saw, CSR performs the best followed by the original SIMOS, CRS and S2B as basically equal.

It is very important that a surrogate method not only has good precision, it also needs to be robust in the sense that it performs well regardless of if the decision-maker in his mind uses a cognitive model where the representation has $N$ or $N-1$ DoF or any combination thereof. Table 5 shows the differences in results between the $N$ and $N-1$ DoF simulations and Table 6 shows the standard deviation of these differences. The most robust method in this sense is obviously CSR. The other methods perform

**Table 6** Standard deviation of spread

| Total spread | ROC | SR | CRC | CRS | CRR | CSR | S1 | S2A | S2B |
|---|---|---|---|---|---|---|---|---|---|
| SD | 6.4 | 2.2 | 3.6 | 5.9 | 2.8 | *1.0* | 7.9 | 5.4 | 6.2 |
| Rank | 8 | 2 | 4 | 6 | 3 | 1 | 9 | 5 | 7 |

**Table 7** Final score

| Final score | ROC | SR | CRC | CRS | CRR | CSR | S1 | S2A | S2B |
|---|---|---|---|---|---|---|---|---|---|
| Score | 74.9 | 80.7 | 80.9 | 80.2 | 77.4 | *85.9* | 78.3 | 79.8 | 79.8 |
| Rank | 9 | 3 | 2 | 4 | 8 | 1 | 7 | 6 | 5 |

worse, even worse than the ordinal SR method, and notably the SIMOS varieties are in this respect not performing very well.

The final score for the surrogate weight methods are computed as Final score = Mean result − Spread, taking both precision and robustness into account. Table 7 shows the final scores of the comparisons. CSR is significantly better than the others, with CRC and SR far behind. The original SIMOS and the refined are quite equal and all of them are worse than SR.

Since the CSR method performed the best both in precision and robustness, it is top of the form in the final score table and consequently it is the method that this study recommends for use as a surrogate weight method.

## 5 Concluding Remarks

Elicitation methods available today are often too cognitively demanding for normal real-life decision-makers and there is a clear need for weighting methods that do not require formal decision analysis knowledge. We have investigated a spectrum of methods, including state-of-the-art approaches for asserting surrogate weights with the possibility to supply information regarding preference strength as well as have found some interesting results of mixed models of decision-maker behaviour considering which degree of freedom that is adequate. We have compared these models and propose the so called CSR method, which extends the rank order weighting procedure SR from Danielson and Ekenberg (2014) by also taking strength preference into account in a more straightforward way than previously suggested in Danielson et al. (2014). CSR has several desired robustness properties and is comparatively stable under reasonable assumptions and is also usable for multi-stakeholder decision making. Figure 1 shows of a multi-criteria multi-stakeholder tool developed on CSR targeting infrastructure policy making in Swedish municipalities.

We conclude that to be robust, a rank ordering method should fare well under both of these assumptions and others. In the assessment, we also include the well-known and popular Simos methods, see e.g., Morais et al. (2014). We have found that the other methods analysed here are clearly behind the CSR weights in performance

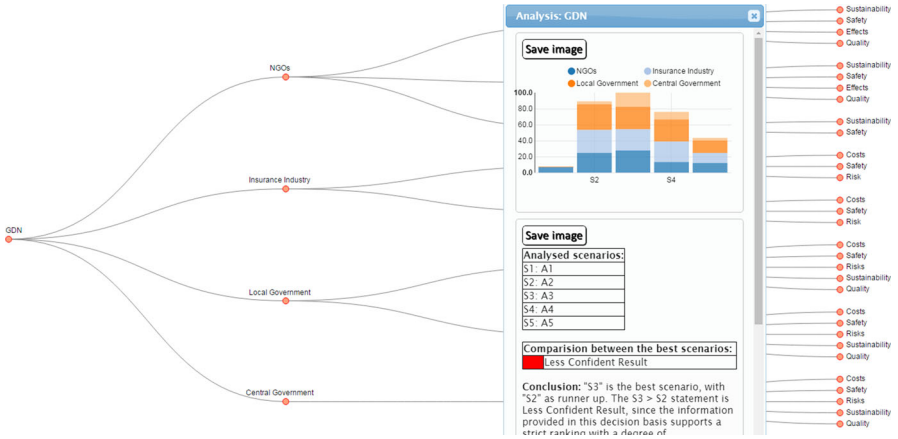


**Fig. 1** The Group Decision tool Decision Wizard

considering both precision and robustness of the results and, despite their relative popularity, neither of the original nor refined Simos methods improve much on CRS (Eq. 5), which they resemble the most.

There exist also a number of MCDA methods suggested and all of these have not been compared systematically against each other. Next step in this work is to compare with some other approaches suggested over the years, in particular the dominance rules suggested in Sarabando and Dias (2009, 2010), Aguayo et al. (2014), Mateosetal. (2014). Furthermore, the idea with this approach is that it should combine realistic decision making with a reasonable degree of simplicity so that it can be used by real life decision makers. The above mentioned Decision Wizard tool is supposed to, at least partly, accomplish this, but it remains to test whether this is accepted at a broad basis by the stakeholders it is intended for, i.e., public servants and politicians in the Swedish municipalities. Another development is to put this in a context of a more formalised and acceptable decision process as discussed in, e.g., Riabacke et al. (2012) for multi-stakeholder decision making.

**Acknowledgments** This research was funded by the Swedish Research Council FORMAS, project number 2011-3313-20412-31, as well as by Strategic funds from the Swedish government within ICT—The Next Generation.



## References

Aguayo EA, Mateos A, Jiménez-Martín A (2014) A new dominance intensity method to deal with ordinal information about a DM's preferences within MAVT. Knowl Based Syst 69:159–169

Ahn BS, Park KS (2008) Comparing methods for multiattribute decision making with ordinal weights. Comput Oper Res 35(5):1660–1670

Arbel A, Vargas LG (1993) Preference simulation and preference programming: robustness issues in priority derivation. Eur J Oper Res 69:200–209
Barron FH (1992) Selecting a best multiattribute alternative with partial information about attribute weights. Acta Psychol 80(1–3):91–103
Barron F, Barrett B (1996b) Decision quality using ranked attribute weights. Manage Sci 42(11):1515–1523
Barron F, Barrett B (1996a) The efficacy of SMARTER: simple multi-attribute rating technique extended to ranking. Acta Psychol 93(1–3):23–36
Danielson M, Ekenberg L, He Y (2014) Augmenting ordinal methods of attribute weight approximation. Decis Anal 11(1):21–26
Danielson M, Ekenberg L (2014) Rank ordering methods for multi-criteria decisions. In: Proceedings of the 14th group decision and negotiation—GDN 2014. Springer
Danielson M, Ekenberg L (2007) Computing upper and lower bounds in interval decision trees. Eur J Oper Res 181(2):808–816
Danielson M, Ekenberg L, Larsson A, Riabacke M (2013) Weighting under ambiguous preferences and imprecise differences in a cardinal rank ordering process. Int J Comput Intell Syst
Figueira J, Roy B (2002) Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. Eur J Oper Res 139:317–326
Jia J, Fischer GW, Dyer J (1998) Attribute weighting methods and decision quality in the presence of response error: a simulation study. J Behav Decis Mak 11(2):85–105
Katsikopoulos K, Fasolo B (2006) New tools for decision analysis. IEEE Trans Syst Man Cybern A Syst Hum 36(5):960–967
Mateos A, Jiménez-Martín A, Aguayo EA, Sabio P (2014) Dominance intensity measuring methods in MCDM with ordinal relations regarding weights. Knowl Based Syst 70:26–32
Morais DC, de Almeida AT, Figueira JR (2014) A sorting model for group decision making: a case study of water losses in Brazil. Group Decis Negot 23:937–960
Rao JS, Sobel M (1980) Incomplete Dirichlet integrals with applications to ordered uniform spacing. J Multivar Anal 10:603–610
Riabacke M, Danielson M, Ekenberg L (2012) State-of-the-art in prescriptive weight elicitation. Adv Decis Sci. doi:10.1155/2012/276584
Roberts R, Goodwin P (2002) Weight approximations in multi-attribute decision models. J Multi-Criteria Decis Anal 11:291–303
Sarabando P, Dias L (2009) Multi-attribute choice with ordinal information: a comparison of different decision rules. IEEE Trans Syst Man Cybern A 39:545–554
Sarabando P, Dias L (2010) Simple procedures of choice in multicriteria problems without precise information about the alternatives' values. Comput Oper Res 37:2239–2247
Scharlig A (1996) Pratiquer Electre et PROMETHEE Un complement à decider sur plusieurs critères. Collection Diriger L'Entreprise, Lausanne: Presses Polytechniques et Universitaires Romandes
Simos J (1990a) Evaluer l'impact sur l'environnement: Une approche originale par l'analyse multicriteere et la negociation. Presses Polytechniques et Universitaires Romandes, Lausanne
Simos J (1990b) L'evaluation environnementale: Un processus cognitif neegociee. Theese de doctorat, DGF-EPFL, Lausanne
Stewart TJ (1993) Use of piecewise linear value functions in interactive multicriteria decision support: a Monte Carlo study". Manag Sci 39(11):1369–1381
Stillwell W, Seaver D, Edwards W (1981) A comparison of weight approximation techniques in multiattribute utility decision making. Org Behav Hum Perform 28(1):62–77
von Winterfeldt D, Edwards W (1986) Decision analysis and behavioural research. Cambridge University Press, Cambridge