ISSN 1946-7664. MCFNS 2013
Vol. 5, Issue 1, pp. 1–13
Mathematical and Computational Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Apr. 17, 2012
Accepted: Jan. 4, 2013
Published: Mar. 30, 2013
Last Correction: Mar. 30, 2013

COMBINING EXPERTS' RATINGS FOR PARTLY OVERLAPPING CONCEPT LISTS: A FEASIBILITY TEST WITH CULTURAL SUSTAINABILITY INDICATORS

S. Sironen[1*], T. Hujala[2], T. Myllyviita[1], J. Tikkanen[3], P. Leskinen[1]

[*] *Corresponding Author,* [1] *Finnish Environment Institute, Joensuu, Finland*

[2] *Finnish Forest Research Institute, Vantaa, Finland*

[3] *Oulu University of Applied Sciences, Oulu, Finland*

Abstract. Acquiring preference information from decision-makers and stakeholders may carry biasing effects due to question framing. In order to avoid unwanted distortions, respondent-driven querying methods are advisable to apply; however, concerning multiple stakeholders a challenge remains how to combine individually collected concepts and further on their individual valuations to an unified preference information. This paper introduces one solution: a semi-automatic stochastic simulation of joint preferences from partially overlapping individual concept lists and preference ratings. We used completed expert interview dataset of cultural sustainability indicators acquired for comparing bioenergy production chains. According to the results the approach seems generally applicable, but the feasibility may vary according to case characteristics. Combining concept list valuations with stochastic simulations may be more feasible the more similar the expected concept structures are. The presented method contributes particularly to planning processes in which democratic participation of a large number of stakeholders is needed in the goal setting phase. However, more tests with different decision problem types are needed to verify and refine the present findings.

Keywords: Computational analysis, Decision support systems, Methodology, Multi-objective, Simulation, Strategic planning.

1 Introduction

Goal and preference information elicited from decision-makers and stakeholders acts a key role in solving ill-defined societal-ecological decision problems such as watershed management (Kaplowitz and Witter 2008; Gooch and Stålnacke 2010), bioenergy impact assessments (Buchholz et al. 2008) or participatory planning of publicly owned forests (Nordström et al. 2009). In these kinds of decision problems the number of relevant stakeholders is typically large. There are some problem structuring methods for large-group interventions (see Shaw et al. 2004). They typically involve quantitative features (van der Lei and Thissen 2009), which can be realized in a form of statistical preference analysis, for example (Kainulainen et al. 2009).

In participative planning processes, attention should be paid to the constellation of goal queries, because preference elicitation may carry biasing effects owing to question framing or concepts given by the analysts (Tversky and Kahneman 1981; Morton and Fasolo 2009). Therefore respondent-driven querying methods, such as conceptual cognitive mapping (3CM) (Kearney and Kaplan 1997), are advisable in order to avoid unwanted distortions. However, when applying respondent-driven queries for large groups, detailed analysis of connections between given concepts appears unfeasible. A solution might be to divide the goal analysis in two subsequent phases: i) deriving the concept list and initial priorities, and ii) structuring the problem further with a focus on connections between the concepts. Phase i can be conducted flexibly with survey techniques or individual interviews, combined with numerical analysis (e.g. Hahn and Ahn 2005), while phase ii requires facilitated modelling (see Franco and Montibeller 2010) in a smaller group setting with the aid of cognitive mapping or one of its several variations such as causal mapping or reasoning maps (e.g. Eden 1988; Eden et al. 1992; Özesmi and Özesmi 2004; Siau and

Copyright © 2013 Publisher of the *Mathematical and Computational Forestry & Natural-Resource Sciences*

Tan 2005, Montibeller and Belton 2006, Montibeller et al. 2008).

The phase i contains the demanding task of how to combine individually collected information about concepts into a joint concept list, and how to derive an overall importance of those concepts as kind of a 'compromise weight' (Wei et al. 2000). Because of the known cognitive discrepancies of respondents, namely imperfect memory, selective attention, as well as constrained satisfaction (Festinger 1957; Simon et al. 2004), there is a reason to assume some importance for concepts that individual respondents simply forgot to mention. Therefore, ratings for non-overlapping items may be sought for. Some methods for dealing with incomplete preference data already exist (e.g. Hahn and Ahn 2005; Choi and Ahn 2009; Choi and Bae 2009), but there is a lack of procedures explicitly suitable for large-scale open-ended concept queries.

One example of respondent-driven elicitation process is sustainability assessment (see e.g. Xing and Dangerfield 2011). Sustainable development is rapidly changing from an abstract idea to measurable concept, after numbers of ecological, economic and social sustainability concepts or indicators have been identified. An indicator is a variable, which describes one characteristic of the state of a system, usually through observed or estimated data (OECD 2003). Sustainability assessments are mostly expert-driven processes (Buchholz et al. 2007; Phillis et al. 2011). Experts from various fields have described and selected relevant criteria and indicators in order to evaluate sustainability. Often criteria and indicators are adopted from literature or some other indicator lists; however, no universally accepted sustainability indicators are available, since sustainability is context-specific. Therefore it is advisable to identify sustainability indicators for each sustainability assessment separately. Cultural sustainability is the fourth pillar of sustainability (UNESCO 2002), but so far indicators of cultural sustainability are few in number. Therefore, defining indicators for cultural sustainability is fundamentally a typical task that needs several experts to think the matter over in an open-ended, respondent-driven way. The process of identifying sustainability indicators can be supported with various methods and tools which are qualitative in nature (Mendoza and Prabhu 2006). Participant can also evaluate the importance of sustainability criteria with various quantitative techniques such as Multi-criteria Decision Analysis (MCDA) (Mendoza and Prabhu 2000; Balana et al. 2010; Wolfslehner and Vacik 2011).

The purposes of this study were to present the procedure on how to combine ratings for non-overlapping concepts by using stochastic simulations and to analyze uncertainties related to this process. The aim was to combine the concept lists acquired from the expert interviews and generate the missing preferences, which result from the fact that all the experts have not defined and evaluated the same items. Uniform distributions with different distributional assumptions were tested in order to found out, whether this kind of generation method would provide results accurate enough to replace the second phase interviews needed to combine the different numerical expert judgments or concept lists. Assumptions were tested with a dataset provided by expert interviewees in the case of compiling cultural sustainability indicators in order to compare bioenergy production chains in eastern Finland. This particular dataset was chosen as it was already available, otherwise issues related to decision making or sustainability of bioenergy production chains were not considered in this study.

2 Material and Methods

2.1 Study material Expert interview data acquired to compile information on sustainability indicators (Myllyviita et al. 2013) were used as test data in this study. The expert interview data included two-phased data gathering of concept lists and rating of the concept lists. Data were gathered in the autumn 2010. Altogether 12 experts were interviewed twice during the process. At first phase, the experts were interviewed in order to construct concept list of items each of the experts themselves considered to be relevant when evaluating the role of four bioenergy production chains in supporting cultural sustainability. Then the experts expressed their preferences with an application of SMART (e.g. von Winterfeldt and Edwards 1986) by directly rating each of the items in their concept list in turn. The experts were requested to first select the most relevant item when considering the cultural sustainability, and assign 100 points to that particular item. Then the experts were asked to rank the other items correspondingly in the numerical scale from 0 to 100 including the possibility to assign the same value for several items. Completely irrelevant items were asked to be given 0 points. After all selected experts were interviewed a combined item list was composed from all the items the experts had defined in the 1^{st} phase. Each item identified in the interviews was included in the combined concept list once regardless of being defined by one or several experts. The combined concept list constructed after the 1^{st} phase interviews comprised altogether 49 items, i.e., indicators of cultural sustainability.

The second phase interviews were carried out after all the 1^{st} phase expert interviews were finalized. The combined list was re-evaluated by the same 12 experts. In the second phase interviews, each expert's personal statements of preferences for the different items defined

during the first phase were expressed in the combined item list, but the preferences stated by other experts were not revealed. Similarly to the 1^{st} phase, the experts were asked to evaluate each item in the combined list in terms of the relevancy of the item in question in assessing cultural sustainability of four bioenergy production chains. The experts were encouraged to utilize the values they had previously given to the items at the 1^{st} phase as reference points. Moreover, the experts were allowed to make changes to the valuations they had previously made. Contrary to the 1^{st} phase interviews, the experts were allowed to evaluate the different items with points over 100. A more detailed description of the process and the indicators of cultural sustainability may be found in Myllyviita et al. (2013). In this study, this data acquired from the 2^{nd} phase interviews were used as reference data, i.e., it included the real values where the generated values were compared in order to find out the accuracy of the tested generation methods.

2.2. Methods for generating the missing preferences In the 1^{st} phase, the experts had identified and evaluated the items each of them separately considered relevant with the applied direct rating method. In order to attain joined relative importance to each of the item, the expert level numerical evaluations need to be combined. However, usually all the experts do not define all the same items, thus there exists missing preferences. Therefore, second phase re-evaluation considering all the items is required. In this study, these missing preferences were generated by assuming probability distribution for the missing values, and producing a random realization for those values.

The continuous uniform distribution with different distributional assumptions and parameters was selected as the generation method in this study. A continuous uniform distribution has constant probability density on an interval (a, b) and zero probability density elsewhere**.** Thus the probability of any value from a continuous uniform distribution having a value between the minimum and maximum is equal.The distribution is specified by these two parameters a and b, and often abbreviated U(a,b). A probability density function for a continuous uniform distribution in interval (a, b) is defined as

$$f(x) = \begin{cases} 0 & \text{for} \quad x < a \\ \frac{1}{b-a} & \text{for} \quad a \leqslant x \leqslant b \\ 0 & \text{for} \quad x > b \end{cases} . \qquad (1)$$

Several assumptions related to the utilization of Eq(1) were tested. The first assumption was that the personal interview produces the best alternatives for each of the experts, i.e., the experts define all the items they consider relevant, therefore the new items that were originally missing from their concept list are less important. The other three tested distributions were based on the presumption that the interview process does not necessarily produce the most important items. Thus the second assumption was that the process produces most of the relevant items; however, the experts might consider the new items somewhat more important than the less important items in their concept lists. The third assumption was that the experts might consider the new items to be better or worse, and they will produce the values according to the original scale between 0 and 100. Finally, the fourth assumption was that the experts might consider the new items to be even much more important, and the missing items may receive values between 0 and 200. Particularly, the tested methods were:

UD1: Uniform distribution in an interval from 0 to the minimum value each expert had given for an item at the 1^{st} phase (U(0,min)). Since the minimum value varied from expert to expert, the parameters for the uniform distribution were different to each of the experts.

UD2: Uniform distribution in an interval from 0 to the mean value of the range each expert had given for all the items at the 1^{st} phase (U(0,mean)). Since the minimum and maximum values forming the range varied from expert to expert, the parameters for the uniform distribution were different to each of the experts.

UD3: Uniform distribution in an interval from 0 to 100 (U(0,100)). The parameters of the uniform distribution were similar to each of the experts.

UD4: Uniform distribution in an interval from 0 to 200 (U(0,200)). The parameters of the uniform distribution were similar to each of the experts.

The defined numerical weights of the 1^{st} phase interviews were used as basis of the generations. These values were kept fixed, and random generations were made only for the missing preferences. Random realizations were produced for every missing preference of each of the expert in turn. The random generations were made with the four selected method, and were repeated 1000 times.

After generating the expert level missing preferences, the expert level weights were rescaled to the scale of priorities, and combined by averaging over the 12 experts in order to determine the joined item level priorities. The assigned expert level numerical weights for different items were first rescaled to sum to 1, so that each item's weight was divided by the sum of weights that expert had given in total to all the 49 items. Thus the assigned expert level preferences in the scale of the priorities were

$$\hat{a}_i = \exp(\hat{\alpha}_i)/\Sigma_i \exp(\hat{\alpha}_i). \qquad (2)$$

Each expert received similar weight in the calculations. The total priority for each item in question was calculated as an arithmetic mean over the 12 experts.

The tested methods were compared to the true preferences acquired after the 2^{nd} phase re-evaluations, which were used as reference data. Normally, these true values are unknown. The real values were received by combining the elicited weights from both 1^{st} and 2^{nd} phase interviews. If the expert had changed the weight for a particular item during the 2^{nd} phase, the weight assigned at the 2^{nd} phase was used. The methods were evaluated both at the expert level and item level. The expert level and item level priorities were rescaled and calculated similarly to the above mentioned manner. The results were analyzed by applying order statistics and basic statistics; for example, by calculating mean, minimum, and maximum values, as well as standard deviations. In addition, 95% confidence intervals were calculated. Furthermore, probabilities for the items to attain a given rank, particularly, probability for an item to outperform all the other items, were calculated based on a Monte Carlo simulation technique (e.g. Alho et al. 2001)

3 Results

3.1 Interviews The first phase interviews resulted in altogether 49 items, i.e., indicators of cultural sustainability (Myllyviita et al. 2013). There was much variation between the expert opinions. Individual experts defined and evaluated between 3 to 19 items for their concept lists. The most common item was local raw-material defined by 9 experts (see Appendix). Only six items were defined and evaluated by at least half of the experts, particularly local raw-material, recreational uses, peatlands taken to peat production, stump removal, self-sufficiency and positive impacts of organisations to areas. Most of the items, i.e., 34% of all the 49 items, were defined by only one expert. Two experts defined 22% of all the items. Therefore, the amount of missing items to be generated in combining the concept lists was large.

The real true preferences used as a comparison point were calculated from the preferences acquired after combining the 1^{st} the 2^{nd} phase interviews. The differences between the real preferences in priority scale were not large especially considering the item level results. Twenty-three items received priorities between 2% to 3%, and 22 items priorities between 1% to 2%. However, three items had somewhat larger priorities than the other items, particularly local raw-material, significant change in scenery because of demand of wood, and perceptions of users. These items were defined and evaluated by 9, 4 and 2 experts in the 1^{st} phase interviews, respectively (see Table 1 and Appendix). The real priorities calculated after the 2^{nd} phase for these items were 4.4%, 3.8% and 3.2%, respectively. Although having the highest item level priority, the separate experts' opinions considering the importance of the local raw-material differed quite much. The priorities assigned to local raw-material varied from 1% to 23.3% between the experts. It had much larger standard deviation than the other items (see Appendix).

3.2 Generated missing preferences at expert level The missing values were generated altogether 1000 times for each of the experts. Firstly, the results were evaluated by calculating the minimum, mean and maximum priorities of the 1000 generations for all of the items of each expert. These results were plotted alongside the real priorities each expert had given. Experts 1, 5 and 11 were randomly selected as an example in this study (Fig. 1, 2 and 3). Mostly, the higher the real priority of a particular item, the further away it situated from the mean of the generated values. However, there were some exceptions. Considering the experts 1 and 11, none of the real priorities reached the range of the priorities generated according to the first assumption, i.e., the method UD1 (Fig. 1 and 3). The real priorities were mostly larger than the generated maximum priority. On the contrary, almost all the real priorities were in the range of the generated priorities for expert 5 (Fig. 2). Expert 5 defined and evaluated the smallest number of items for the concept list in the 1^{st} phase interviews.

Considering expert 1, the UD4 method produced the mean values closest to the real priorities. Expert 1 had the largest mean, range and standard deviation of the real evaluations at the original numerical scale after the 2^{nd} phase (Table 1). Therefore the uniform distribution having the largest minimum and maximum values produced the most accurate generations. In addition, the proportion of items having the real priority at the 95% confidence interval of the generated values was highest with the UD4 method. Expert 5 had evaluated only 3 items at the 1^{st} phase and quite large minimum weight as well. Thus, all the generation methods produced quite similar results. The proportion of items having the real priority at the 95% confidence interval of the generated values was 93.9% for the other methods, and 100% for UD3. Contrary to that expert 11 had defined small minimum weight at the 1^{st} phase, therefore UD1 produced quite poor results compared to the real values (Fig. 3). UD3 produced the best results considering the proportion of items reaching the 95% confidence interval. UD2 and UD3 methods seemed to produce mean values close to the real priorities of many items, exceptionally considering the items having the largest priorities as well. In general, considering the 95% confidence intervals, UD2 and UD3 seemed to produce the most accurate results for most of the experts. Considering expert 1, the proportion of items having the real priority at the 95% confidence interval was markedly larger for UD4. The results seemed to be somewhat better the less items the ex-

Table 1: Number and percentage of the items each of the expert defined at 1^{st} phase, real mean and standard deviation (SD) of the original numeric scale both after the 1^{st} and 2^{nd} phase interviews, and the proportion of items having the real priority reaching the 95% confidence interval of the generated priorities with different methods.

Expert	Items	Items, %	Mean 1^{st}	SD 1^{st}	Mean 2^{nd}	SD 2^{nd}	UD1	UD2	UD3	UD4
1	11	22.4	60.5	23.3	89.6	42.7	24.5	77.6	75.5	98.0
2	8	16.3	17.1	14.7	37.4	34.9	44.9	83.7	69.4	75.5
3	9	18.4	57.8	23.3	41.2	25.0	57.1	95.9	81.6	81.6
4	11	22.4	67.7	17.2	52.2	25.1	51.0	77.6	100.0	77.6
5	3	6.1	66.7	28.9	54.7	24.3	93.9	93.9	100.0	93.9
6	19	38.8	71.1	19.1	49.5	23.9	42.9	91.8	63.3	61.2
7	11	22.4	82.7	11.9	75.5	19.8	75.5	57.1	75.5	75.5
8	14	28.6	47.5	25.7	39.2	23.8	59.2	85.7	71.4	73.5
9	11	22.4	25.5	26.2	10.8	16.8	57.1	77.6	77.6	77.6
10	7	14.3	62.9	18.9	52.2	31.3	77.6	77.6	95.9	87.8
11	12	24.5	49.2	27.5	45.3	27.0	22.4	91.8	91.8	75.5
12	13	26.5	77.7	18.3	53.9	26.0	42.9	73.5	100.0	73.5

pert had defined at the 1^{st} phase, especially when UD1 method was considered. Expert 5, 10 and 9, who defined the least items, had quite large proportion of items having the real value at 95% confidence interval.

Furthermore, the probabilities of an item in question outperforming all other items were assigned. UD4 produced quite variable probabilities for every item and every expert. All the other methods were similar. Considering most of the experts, the probabilities were zero for other items, and mostly 1 for one item. This was mainly the item having the largest priority, except for expert 1 for whom it was the item having the 19^{th} largest priority.

Table 2: Correlation between the real priorities acquired after the 2^{nd} phase interviews and mean of the generated priorities for the different methods for each of the expert.

Expert	UD1	UD2	UD3	UD4
1	-0.178	-0.050	0.080	0.369
2	0.206	0.320	0.321	0.218
3	0.460	0.478	0.453	-0.085
4	0.433	0.450	0.439	-0.155
5	0.232	0.228	0.259	0.063
6	0.810	0.817	0.796	-0.223
7	0.245	0.231	0.266	0.050
8	0.496	0.532	0.502	-0.074
9	0.945	0.314	0.212	-0.259
10	0.251	0.251	0.257	-0.050
11	0.338	0.492	0.477	0.178
12	0.630	0.645	0.664	-0.166

Finally, the correlations between the real priorities and the mean of the generated priorities were calculated for each of the experts (Table 2.) For most of the experts, the correlations were similar for UD1, UD2 and UD3, and much smaller correlations were received between the real priorities and UD4. However, two experts differed somewhat. For expert 1, the highest correlations were received between the real priorities and priorities produced with UD4 method. This particular expert had the widest range for the real values at original numerical scale after the 2^{nd} phase re-evaluations (see Table 1.). On the other hand, expert 9 had much larger correlation between the real priorities and priorities produced with UD1. This expert had considerably smaller mean of the original numeric values after the 2^{nd} phase. In general, the correlations seemed to be somewhat better the more items the expert had defined at the first phase.

3.3. Results of the combined concept lists Considering the combined item level priorities, the correlations between the real and generated priorities were smallest for UD1 and UD4, that is, 0.118 and 0.164, respectively. Correlations for UD2 and UD3 were 0.446 and 0.428, respectively. Order statistics, particularly ordering the priorities of the four tested methods according to the rank order of the real priorities, revealed that UD1 produced priorities that seemed to differ most from the real priorities (Fig. 4). However, for the less important items UD1 produced the generated mean priorities closest to the real value. UD4, on the other hand, produced mean values closest to the real for most of the 20 items having the largest priorities. UD2 and UD3 produced closest means for the items in the middle of the rank order.

Generally, all the missing value estimation methods produced smaller priorities than the real ones for the most important items (Table 3). UD1 produced the mean closest to the real for the most important item,

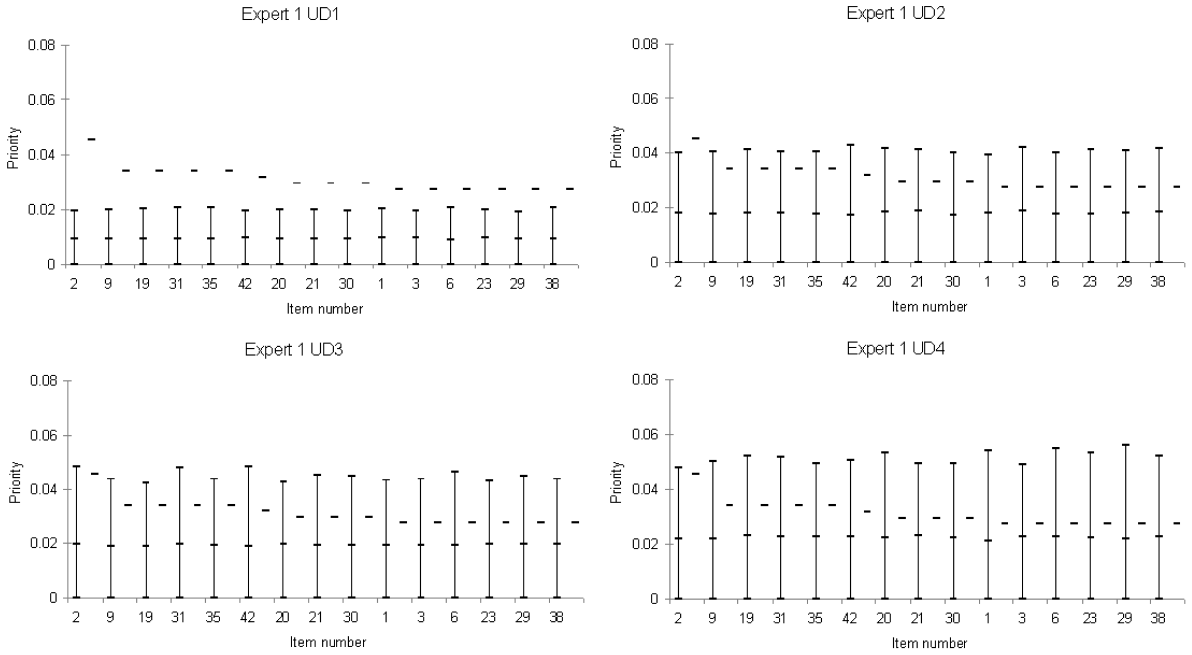

Figure 1: Minimum, mean and maximum priorities of the 1000 generations acquired with different generation methods, i.e., distributional assumptions (UD1, UD2, UD3 and UD4) for the 15 most important items of expert 1 plotted alongside the real priority expert 1 had given for the items after the 2^{nd} phase interviews.

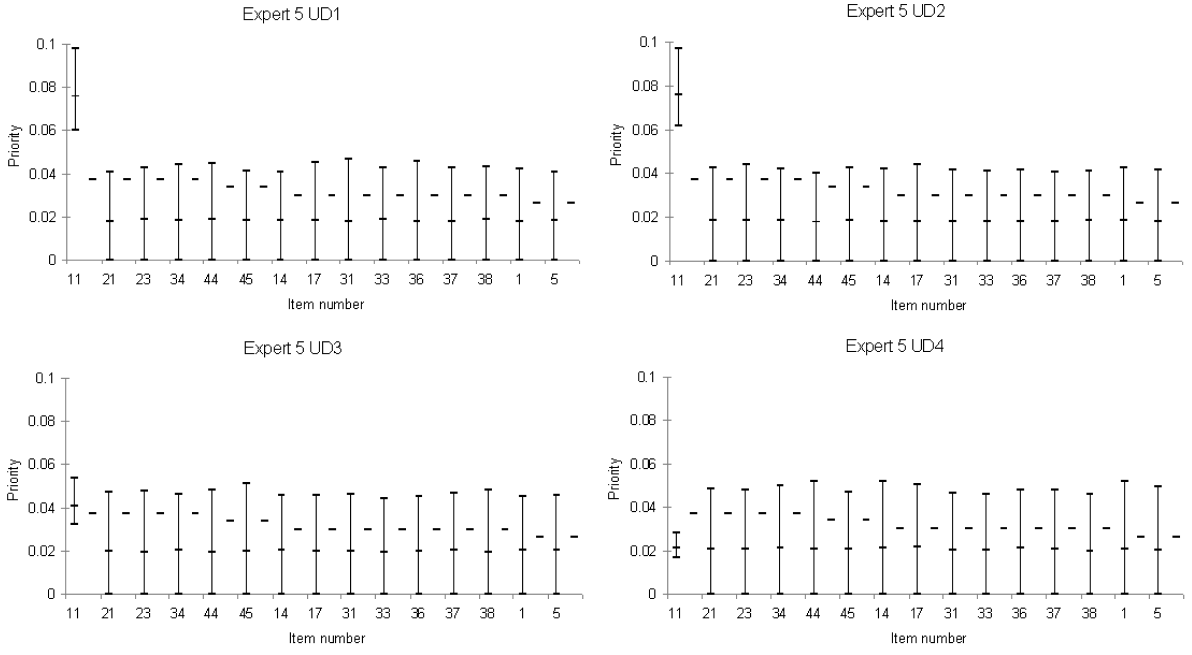

Figure 2: Minimum, mean and maximum priorities of the 1000 generations acquired with different generation methods, i.e., distributional assumptions (UD1, UD2, UD3 and UD4) for the 15 most important items of expert 5 plotted alongside the real priority expert 5 had given for the items after the 2^{nd} phase interviews.

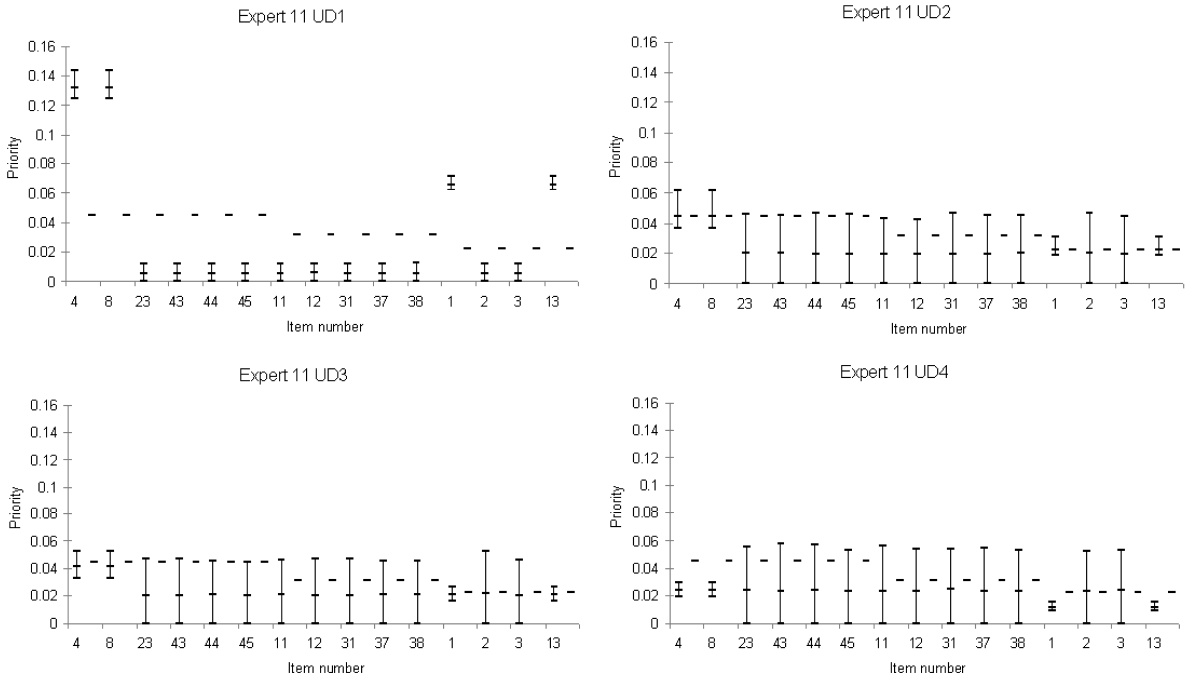

Figure 3: Minimum, mean and maximum priorities of the 1000 generations acquired with different generation methods, i.e., distributional assumptions (UD1, UD2, UD3 and UD4) for the 15 most important items of expert 11 plotted alongside the real priority expert 11 had given for the items after the 2^{nd} phase interviews.

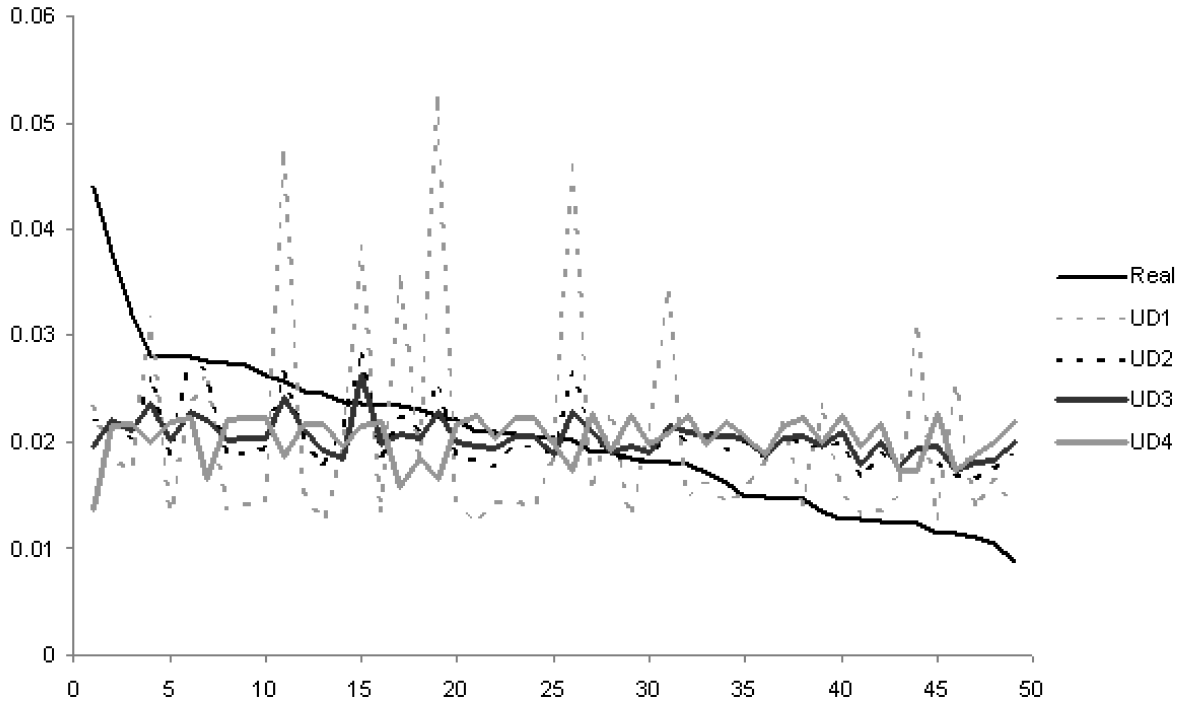

Figure 4: The item level priorities achieved with each of the generation method ordered in rank order of the real priorities.

i.e., local raw-material. However, the real value was not in the 95% confidence interval in any of the methods. Considering the 15 most important items, the real priorities of those items, were at 95% confidence interval most often applying the UD3 method. In addition, UD2 had only one item and UD4 two items less at 95% confidence interval. UD1 produced the poorest results considering the confidence intervals of the most important items; the real priorities of only 5 items reached the 95% confidence interval. Similar results were achieved by plotting minimum, mean and maximum values of the generated priorities alongside the real priorities (Fig. 5). None of the methods produced such values that the real priority of the most important item would have been in the range. In addition, the 2^{nd} most important items with the UD1 and UD2 methods were not in the range. In general, the real priorities were near the maximum of the generated priorities.

Furthermore, the probabilities of a particular item outperforming all other items were assigned at the item level as well. UD1 produced the highest probability for item 13, which was in the 19^{th} place in rank order of the real item level priorities. UD2 and UD3 produced the highest probability for item 14, which was the 15^{th} most important item in real. UD4 produced the highest probability for an item at 32^{nd} place in rank order. The UD2 method produced the highest probability for the most important item number 1. Considering the 15 most important items, UD2 seemed to more likely produce high priorities to those items. The results of UD3 were quite similar (Table 4).

4 Discussion

Practical experience has shown that the acquisition of knowledge from experts is costly and time-consuming task (Monti and Carenini 2000). Rarely an effective method is selected beforehand leading straightforwardly to the solution. Environmental problems may be complex, involve many parties, and have no easy solutions or right answers (Kearney and Kaplan 1997). In spite of this complexity, the decisions must be made. These kinds of decisions may be concluded by interdisciplinary teams or group of experts. Several approaches for combining individually collected information into joint conceptual model have been presented. In this study we focused on the problem on how to combine preference information collected from individuals separately into overall analysis about the decision indicators.

The indicator definition process relied on experts' knowledge, and was carried out by conducting expert interviews at two stages. The second phase re-evaluation was required, since all the experts had not defined and evaluated all the same items at first round of interviews. Methods for generating these missing preferences were tested in this study in order to acquire combined preferences without this re-evaluation. The uncertainty in combining the preferences and concept lists was assumed to result from the difference in expert opinions, and from the fact that all the possible items are not defined and evaluated by all the experts. The results of combining the concept lists were analyzed to determine, which issues affect the uncertainty most, and whether the results would be more accurate if there was less variation between the experts or more experts had evaluated the item in question.

In general, combining the concept lists through generating the missing preferences of each of the experts by assuming a distribution for the missing values seemed to produce quite accurate results, except for the items having the largest priorities. The differences in the real priorities of other items were quite small. Much difference was not acquired, since the number of the items defined altogether was large and therefore the average priority could not be large. In addition, the original evaluation scale was quite narrow, although the experts were allowed to give weights larger that 100 at the 2^{nd} phase re-evaluations. Moreover, the experts seemed to be quite careful in their evaluations, i.e., when they were uncertain about the importance of an indicator, they gave a moderate weight to that indicator. Most of the experts evaluated the items at the 2^{nd} phase such that the real mean value at the original numeric scale was near 50, and standard deviation near 25. Two of the experts had larger means, and one expert markedly lower mean for the evaluations. At the 2^{nd} phase the experts needed to re-evaluate all the items other experts had considered relevant, and the experts may seek for consensus.

According to Myllyviita et al. (2013), the experts considered that interaction with the other experts would be necessary. In addition, all the experts were able to construct their concept list; however, almost half of the experts considered that numerical valuation of the items with the SMART-application was artificial and demanding. This could be because most of the experts did not have experience on applying numerical MCDA methods, whereas most of the experts were acquainted with verbal judgments. One essential component of the cognitive mapping according to Kearney and Kaplan (1997) is that the participants should choose only those concepts that are meaningful to them to ensure that each individual's final sorting reflect only those object they own. The experts were encouraged to give zero weight to irrelevant items; however, most of them provided some weight for most of the items, although having defined only a few items themselves at the first phase. More interesting results might be achieved if there had been more difference between experts and the item level prior-

Table 3: The real mean and standard deviation for the 15 most important items in the concept list, as well as the generated values for the same items applying the different distributional assumptions. In addition, the item in question is marked with x if the real value reached the 95% confidence interval of the generated value.

Item number	Real Mean	Real SD	UD1 Mean	UD1 SD	95% CI	UD2 Mean	UD2 SD	95% CI	UD3 Mean	UD3 SD	95% CI	UD4 Mean	UD4 SD	95% CI
1	0.0441	0.0573	0.0236	0.0102	-	0.0219	0.0081	-	0.0197	0.0069	-	0.0138	0.0040	-
23	0.0379	0.0185	0.0183	0.0046	-	0.0222	0.0050	-	0.0222	0.0048	-	0.0215	0.0050	-
42	0.0321	0.0095	0.0173	0.0046	-	0.0203	0.0050	-	0.0211	0.0049	-	0.0216	0.0047	-
21	0.0282	0.0099	0.0318	0.0115	x	0.0264	0.0055	x	0.0237	0.0044	x	0.0201	0.0041	-
44	0.0282	0.0113	0.0134	0.0042	-	0.0187	0.0048	-	0.0201	0.0049	x	0.0218	0.0052	x
11	0.0279	0.0099	0.0228	0.0110	x	0.0283	0.0107	x	0.0228	0.0055	x	0.0223	0.0046	x
8	0.0276	0.0124	0.0267	0.0063	x	0.0263	0.0048	x	0.0221	0.0040	x	0.0167	0.0038	-
31	0.0275	0.0090	0.0136	0.0044	-	0.0187	0.0049	x	0.0202	0.0047	x	0.0222	0.0048	x
20	0.0272	0.0091	0.0141	0.0044	-	0.0191	0.0050	x	0.0204	0.0050	x	0.0224	0.0050	x
45	0.0263	0.0129	0.0143	0.0044	-	0.0194	0.0049	x	0.0204	0.0049	x	0.0224	0.0053	x
4	0.0256	0.0114	0.0475	0.0093	-	0.0266	0.0041	x	0.0241	0.0035	x	0.0187	0.0036	x
39	0.0248	0.0137	0.0154	0.0044	-	0.0203	0.0050	x	0.0213	0.0050	x	0.0217	0.0049	x
30	0.0245	0.0099	0.0124	0.0041	-	0.0177	0.0048	x	0.0193	0.0050	x	0.0216	0.0051	x
3	0.0238	0.0118	0.0192	0.0053	x	0.0212	0.0048	x	0.0184	0.0036	x	0.0196	0.0047	x
14	0.0237	0.0155	0.0385	0.0147	x	0.0287	0.0066	x	0.0263	0.0053	x	0.0216	0.0038	x

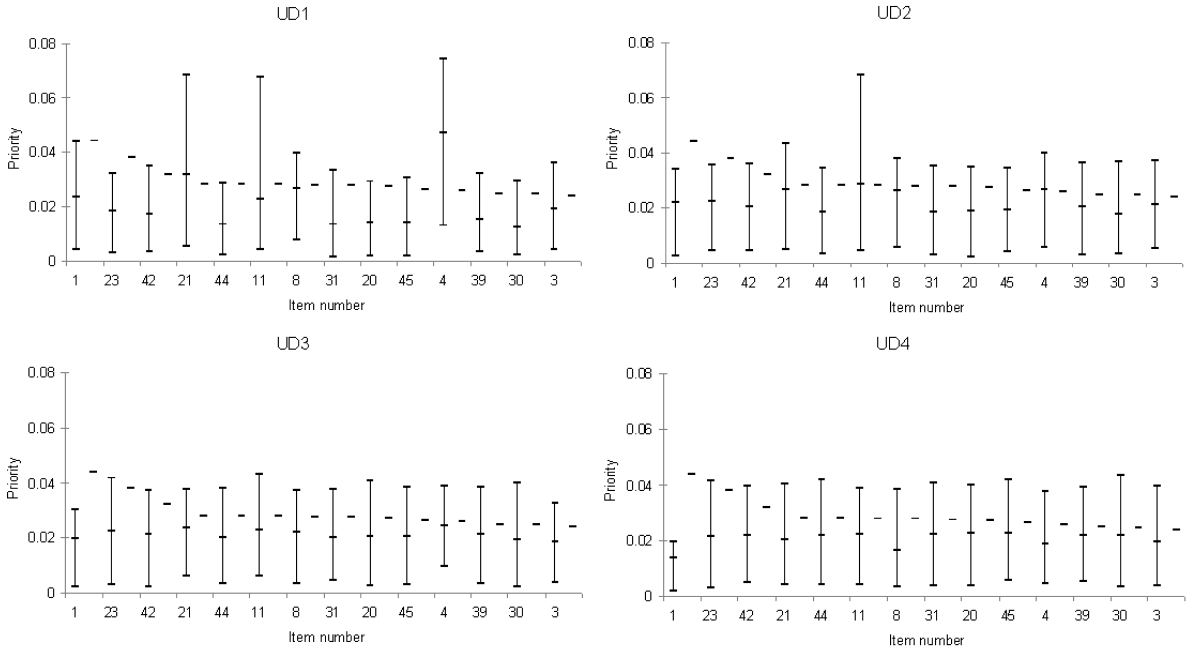

Figure 5: Minimum, mean and maximum for the combined item level priorities acquired with different generation methods, i.e., distributional assumptions (UD1, UD2, UD3 and UD4) for the 15 most important items plotted alongside the real combined item level priorities after the 2^{nd} phase interviews.

Table 4: The real priorities of the 15 most important items and the probability of an item in question to outperform all the other items arranged in the order of magnitude separately for each of the methods.

Item	Real priority,%	Item	UD1	Item	UD2	Item	UD3	Item	UD4
1	4.41	13	0.623	14	0.294	14	0.239	2	0.055
23	3.79	14	0.195	11	0.256	1	0.067	48	0.053
42	3.21	11	0.080	1	0.173	11	0.066	19	0.052
21	2.82	4	0.068	13	0.112	23	0.054	38	0.051
44	2.82	7	0.033	21	0.060	4	0.052	15	0.049
11	2.79	21	0.001	4	0.036	21	0.050	16	0.049
8	2.76	1	0	7	0.031	13	0.040	12	0.048
31	2.75	23	0	8	0.022	2	0.038	45	0.045
20	2.72	42	0	23	0.007	16	0.032	34	0.045
45	2.63	44	0	42	0.002	12	0.029	11	0.042
4	2.56	8	0	9	0.002	39	0.022	20	0.042
39	2.48	31	0	31	0.001	18	0.021	32	0.041
30	2.45	20	0	39	0.001	15	0.020	41	0.036
3	2.38	45	0	46	0.001	42	0.019	39	0.032

ities. However, in such a situation, the results might be different, and at least the accuracy of the methods and the best assumptions to form a uniform distribution and random realizations might differ from these.

Although the test data used in this study might not be the best data to test this kind of combining method, some results were achieved. The UD2 (uniform distribution in an interval from 0 to the mean value of the range each expert had given for all the items at the 1^{st} phase) and UD3 (uniform distribution in an interval from 0 to 100) methods produced the best results with least uncertainty. These methods were based on the assumption that the new missing items may receive almost as large or as large weight as the items each expert had first defined. The UD1 (uniform distribution in an interval from 0 to the minimum value each expert had given for an item at the 1^{st} phase) produced mainly lower priorities than the real ones were, and UD4 (uniform distribution in an interval from 0 to 200) had somewhat larger variation. The least accurate results were achieved assuming that the experts would define all the relevant items at first phase, and would consider the new items less important (UD1). In addition to the small variation between the experts, the limited number of the experts owing to the intensive method used in gathering the concept information and evaluations complicate the analysis of the results. One of the assumptions was, whether the results of combined item level priorities would be better if there were more experts defining that particular item at the 1^{st} phase. The results did not indicate such, although in other applications it might be possible. On the contrary, the results seemed to be somewhat less accurate the more experts had defined the item at the 1^{st} phase. The results indicated that combining of the concept lists would be more accurate, if the expert opinions were more similar and the variation between the expert preferences was smaller. However, it is quite difficult to say since the variation was quite small for all the other items than the most important item local raw-material. Moreover, the amount of experts defining the same particular item at the 1^{st} phase was small.

Furthermore, the results of couple of experts indicate that the method might be considered adjustable, and different distributional assumption could be used for different kind of experts. UD4 produced markedly better results for expert 1 than any other method, and similarly, UD 1 for expert 9. The mean, range and standard deviations of these experts especially at the 2^{nd} phase differed from the rest of experts, which had quite averaged real values. Some correspondence was found between the 1^{st} and the 2^{nd} phase results. There was quite high correlation between the mean values and the range each expert had given to all of the items at 1^{st} phase and at 2^{nd} phase. The first phase values may indicate the relevant distribution to be used for each expert as the generation method. Small mean weights at the 1^{st} phase may indicate that the expert in question may give small weights for the items at second phase as well, and vice versa, large weights in the 1^{st} phase may indicate large weights at the 2^{nd} phase as well. Therefore, a uniform distribution with narrow interval might be more suitable as a generation method at the former case, and a uniform distribution allowing large weights might be the best option at the latter case. However, to verify and refine these kinds of assumptions, more testing with

larger data and different types of decision problems is needed.

5 Conclusions

One of the most obvious results from combining preference information with missing preferences by generating them under different distributional assumptions was that the first assumption of experts being able to define all the items they consider relevant did not hold. Moreover, the experts did not consider the new items that were originally missing from their concept lists less important; which indicates the existence of cognitive biases (Festinger 1957; Simon et al. 2004) in this case. At least in a decision problem as complex as this, the expert may not know all the relevant items at first. Although the generation methods produced quite accurate results in this kind of situation with small variation between the experts, and small amount of items defined by one separate expert, but altogether large amount of items, it would be more appropriate to re-evaluate the combined item lists instead of generating the missing values. Somewhat averaged results were acquired from the experts as well; however, the number of the missing items to be generated may be large in challenging and difficult decision making problems such as this. To sum up, the present results from combining concept lists from a small number of respondents support the prime idea that an automatic stochastic analysis of concept sets could be a feasible and time-efficient approach in systematizing open-ended goal or preference queries of large number of stakeholders.

Acknowledgements

This study was funded by the Academy of Finland project "Bridging the gap between qualitative problem structuring and quantitative decision analysis in forestry" (decision number 127681).

References

Alho, J., M. Kolehmainen and P. Leskinen. 2001. Regression methods for pairwise comparisons data. P. 235-251 in Schmoldt, D. L., J. Kangas, G.A. Mendoza and M. Pesonen (eds). The analytic hierarchy process in natural resource and environmental decision making. Kluwer Academic Publishers, Dordrecht.

Balana, B.B., E. Mathijs and B. Muys. 2010. Assessing the sustainability of forest management: An application of multi-criteria decision analysis to community forests in northern Ethiopia. Journal of Environmental Management 91: 1294-1304.

Buchholz, T.S., T.A. Volk and V. Luzadis. 2007. A participatory systems approach to modeling social, economic, and ecological components of bioenergy. Energy Policy 35: 6084-6094.

Choi, S.H. and B.S. Ahn. 2009. IP-MAGS: an incomplete preference-based multiple attribute group support system. Journal of the Operational Research Society 60: 496–505.

Choi, S.H. and S.M. Bae. 2009. Strategic information systems selection with incomplete preferences: a case of a Korean electronics company. Journal of the Operational Research Society 60: 180-190.

Eden, C. 1988. Cognitive Mapping: a review. European Journal of Operational Research 36: 1-13.

Eden, C., F. Ackermann and S. Cropper. 1992. The analysis of cause maps. Journal of Management Studies 29: 309-324.

Festinger, L. 1957. A theory of cognitive dissonance. Stanford University Press, Stanford, CA.

Franco, L.A. and G. Montibeller. 2010. Facilitated modelling in operational research. European Journal of Operational Research 205: 489–500.

Gooch, G. and P. Stålnacke (eds.). 2010. Science, Policy and Stakeholders in Water Management: An Integrated Approach to River Basin Management. Earthscan, London.

Hahn, C. H. and B.S. Ahn. 2005. Interactive group decision-making procedure using weak strength of preference. Journal of the Operational Research Society 56: 1204-1212.

Kainulainen, T., P. Leskinen, P. Korhonen, A. Haara and T. Hujala. 2009. A statistical approach to assessing interval scale preferences in discrete choice problems. Journal of the Operational Research Society 60: 252-258.

Kaplowitz, M.D. and S.G. Witter. 2008. Agricultural and residential stakeholder input for watershed management in a mid-Michigan watershed. Landscape and Urban Planning 84: 20-27.

Kearney, A.R. and S. Kaplan. 1997. Toward a methodology for the measurement of knowledge structures of ordinary people: The conceptual content cognitive map (3CM). Environment and Behavior 29: 579-617.

van der Lei, T.E. and W.A.H. Thiessen. 2009. Quantitative problem structuring methods for multi-actor problems: an analysis of reported applications. Journal of the Operational Research Society 60: 1198-1206.

Mendoza, G.A. and R. Prabhu. 2000. Multiple criteria decision making approaches to assessing forest sustainability using criteria and indicators: a case study. Forest Ecology and Management 131: 107-126.

Mendoza, G.A. and R. Prabhu. 2006. Participatory modeling and analysis for sustainable forest management: Overview of soft system dynamics models and applications. Forest Policy and Economics 9: 179-196.

Monti, S., and G. Carenini. 2000. Dealing with the experts inconsistency in probability elicitation. IEEE Transactions on Knowledge and Data Engineering 12(4): 499-508.

Montibeller, G. and V. Belton. 2006. Causal maps and the evaluation of decision options – a review. Journal of the Operational Research Society 57: 779-791.

Montibeller, G., V. Belton, F. Ackermann and L. Ensslin. 2008. Reasoning maps for decision aid: an integrated approach for problem-structuring and multicriteria evaluation. Journal of the Operational Research Society 59: 575-589.

Morton, A. and B. Fasolo. 2009. Behavioural decision theory for multi-criteria decision analysis: a guided tour. Journal of the Operational Research Society 60: 268-275.

Myllyviita, T., K. Lähtinen, L.A. Leskinen, T. Hujala, L. Sikanen and P. Leskinen. 2013. Identifying cultural sustainability indicators for wood-based bioenergy production An application of qualitative mapping technique and Multi-criteria decision analysis (MCDA). Submitted to Environment, Development and Sustainability.

Nordström, E.M., C. Romero, L.O. Eriksson and K. Öhman. 2009. Aggregation of preferences in participatory forest planning with multiple criteria: an application to the urban forest in Lycksele, Sweden. Canadian Journal of Forest Research 39:1979-1992.

OECD (Organisation for Economic Co-operation and Development). 2003. OECD Environmental Indicators: development, measurement, and use. Organisation for Economic Co-operation and Development, Paris.

Özesmi, U. and S. Özesmi. 2004. Ecological models based on people's knowledge:a multi-step fuzzy cognitive mapping approach. Ecological Modelling 176: 43-64.

Phillis, Y.A., E. Grigoroudis and V.S. Kouikoglou. 2011. Sustainability ranking and improvement of countries. Ecological Economics 70: 542-553.

Shaw, D., M. Westcombe, J. Hodgkin and G. Montibeller. 2004. Problem structuring methods for large group interventions. Journal of the Operational Research Society 55: 453-463.

Siau, K. and X. Tan. 2005. Improving the quality of conceptual modeling using cognitive mapping techniques. Data & Knowledge Engineering 55: 343-365.

Simon, D., C.J. Snow and S.J. Read. 2004. The Redux of Cognitive Consistency Theories: Evidence Judgments by Constraint Satisfaction. Journal of Personality and Social Psychology 86: 814-837.

Tversky, A. and D. Kahneman. 1981. The framing of decisions and the psychology of choice. Science 211: 453-458.

UNESCO. 2002. Unesco Universal declaration of cultural diversity. Adopted by the 31^{st} Session of the General Conference of UNESCO, Paris, 2 November 2001. http://unesdoc.unesco.org/images/0012/001271/127160m.pdf.

Wei, Q., H. Yan, J. Ma and Z. Fan. 2000. A compromise weight for multi-criteria group decision making with individual preference. Journal of the Operational Research Society 51: 625-634.

von Winterfeldt, D. and W. Edwards. 1986. Decision Analysis and Behavioral Research. Cambridge University Press, Cambridge, UK. 624 p.

Wolfslehner, B. and H. Vacik. 2011. Mapping indicator models: From intuitive problem structuring to quantified decision-making in sustainable forest management. Ecological Indicators 11: 274-283.

Xing, Y. and B. Dangerfield. 2011. Modelling the sustainability of mass tourism in island tourist economies. Journal of the Operational Research Society 62: 1742-1752.

Appendix:

Table 5: Full list of identified cultural sustainability indicators and their rating information.

	Indicator of cultural sustainability	Min	Mean	Max	SD	*N
1	Local raw-material	0.0099	0.0441	0.2326	0.0573	9
2	Possibilities for uneven-aged stand management	0.0000	0.0180	0.0456	0.0147	1
3	Scenery change because of collecting harvesting residue	0.0000	0.0238	0.0469	0.0118	2
4	Stump removal	0.0052	0.0256	0.0465	0.0114	7
5	Peatlands taken to peat production	0.0054	0.0234	0.0465	0.0115	8
6	Discomfort caused by peat production	0.0081	0.0221	0.0465	0.0115	2
7	Self sufficiency	0.0000	0.0203	0.0436	0.0132	7
8	Recreational uses	0.0054	0.0276	0.0465	0.0124	8
9	Green values	0.0112	0.0230	0.0342	0.0066	5
10	Soundscape	0.0000	0.0105	0.0260	0.0085	2
11	Timeline	0.0163	0.0279	0.0545	0.0099	1
12	Organisations' culture	0.0000	0.0190	0.0407	0.0109	1
13	Positive impacts of organisations to areas	0.0000	0.0225	0.0371	0.0094	6
14	Conflict over raw-material	0.0000	0.0237	0.0545	0.0155	3
15	Permanency of an organisation	0.0000	0.0205	0.0381	0.0099	1
16	Long tradition of utilization of wood for heating	0.0000	0.0146	0.0396	0.0116	1
17	Increase of first thinnings	0.0000	0.0180	0.0347	0.0104	2
18	Efficient utilization of raw-material and by-products	0.0000	0.0149	0.0303	0.0086	3
19	Importance of supporting traditional silviculture	0.0090	0.0211	0.0436	0.0111	1
20	Refinement of raw-material	0.0104	0.0272	0.0417	0.0091	1
21	Acceptability	0.0000	0.0282	0.0396	0.0099	5
22	Depletion of scenery because of storing raw-material	0.0000	0.0114	0.0292	0.0086	3
23	Significant change in scenery because of increased demand of wood	0.0149	0.0379	0.0930	0.0185	4
24	Increased traffic because of transportation of raw-material	0.0000	0.0134	0.0248	0.0088	2
25	Needs for new education	0.0054	0.0203	0.0285	0.0056	2
26	New, efficient and comfortable forestry machinery	0.0000	0.0111	0.0233	0.0079	3
27	Improvements in roads to rural areas because of transporting raw-material	0.0000	0.0147	0.0297	0.0099	1
28	Importance of securing culture of peat production	0.0000	0.0087	0.0244	0.0075	1
29	Ownership of raw-material	0.0000	0.0234	0.0490	0.0129	1
30	Ownership of companies	0.0114	0.0245	0.0490	0.0099	1
31	Utilization of nature near to settlement	0.0099	0.0275	0.0417	0.0090	1
32	Changes required to current production chains	0.0000	0.0128	0.0367	0.0123	1
33	Traditional knowledge related to forests	0.0000	0.0125	0.0303	0.0104	0
34	Spiritual values of forests	0.0000	0.0207	0.0545	0.0148	1
35	Technical challenges of pellet utilization	0.0000	0.0124	0.0342	0.0115	4
36	Authority of contractors	0.0000	0.0209	0.0465	0.0110	2
37	Impact of large companies to the area	0.0041	0.0181	0.0315	0.0083	2
38	Large, supranational companies do not support local culture	0.0000	0.0186	0.0490	0.0147	0
39	Replacing fossil fuels	0.0041	0.0248	0.0465	0.0137	2
40	Utilization of new raw-material	0.0000	0.0123	0.0208	0.0061	4
41	Balance in consumption and production	0.0000	0.0162	0.0347	0.0114	1
42	Perceptions of users	0.0224	0.0321	0.0521	0.0095	2
43	Harvesting of logging residues	0.0000	0.0190	0.0450	0.0126	4
44	Participation of stakeholders	0.0000	0.0282	0.0450	0.0113	1
45	Stakeholders are informed	0.0000	0.0263	0.0450	0.0129	1
46	Positive effects of export in terms of culture transfer	0.0000	0.0171	0.0265	0.0082	4
47	Acceptability of export	0.0000	0.0128	0.0260	0.0080	2
48	Homogeneity of practitioners	0.0000	0.0115	0.0545	0.0148	0
49	Appreciation of labour	0.0000	0.0148	0.0292	0.0091	3

*No. of experts defined at 1^{st} phase.