



Delft University of Technology

ANDURIL - A MATLAB toolbox for ANALysis and Decisions with UnceRtalnty: Learning from expert judgments

Leontaris, George; Morales Napoles, Oswaldo

DOI

[10.1016/j.softx.2018.07.001](https://doi.org/10.1016/j.softx.2018.07.001)

Publication date

2018

Document Version

Final published version

Published in

SoftwareX

Citation (APA)

Leontaris, G., & Morales Napoles, O. (2018). ANDURIL - A MATLAB toolbox for ANALysis and Decisions with UnceRtalnty: Learning from expert judgments. *SoftwareX*, 7, 313-317.
<https://doi.org/10.1016/j.softx.2018.07.001>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



ANDURIL – A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty: Learning from expert judgments

Georgios Leontaris*, Oswaldo Morales-Nápoles

Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Article history:

Received 26 April 2018

Received in revised form 4 July 2018

Accepted 4 July 2018

Keywords:

Structured expert judgment

Cooke's classical model

MATLAB toolbox

EXCALIBUR software

ABSTRACT

The Classical model (or Cooke's model) for elicitation and combination of expert judgments has been used in science and engineering since at least the early 1990's. The most widely used program for applications of this model is EXCALIBUR. However, its code is not available for practitioners, which limits the accessibility and potential of the method. In this paper, we discuss a MATLAB toolbox (ANDURIL¹) intended to fill in this gap. The software has been tested in a recent real-life application reproducing the results of EXCALIBUR. We discuss different advantages for the users from having the developed source code available for practice.

© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Code metadata

Current code version	ANDURIL v1.0
Permanent link to code/repository used for this code version	https://github.com/ElsevireSoftwareX/SOFTX_2018_39
Legal Code License	GNU General Public License
Code versioning system used	None
Software code languages, tools, and services used	MATLAB (including the Statistics and Machine Learning Toolbox), EXCALIBUR
Compilation requirements, operating environments dependencies	MATLAB (including the Statistics and Machine Learning Toolbox)
If available Link to developer documentation/manual	https://github.com/ElsevireSoftwareX/SOFTX_2018_39
Support email for questions	G.Leontaris@tudelft.nl

1. Motivation and significance

In practice, engineers, scientists and decision makers are often confronted with problems where sufficient relevant field data (measurements) are not available. In these cases, expert judgments can become an alternative source of valuable data to support

uncertainty analysis in particular. The subject of treating expert judgments as an alternative source of data has been extensively discussed [1–3]. However, the question of how to combine these judgments remains an active research topic. Cooke's classical model for structured expert judgment (SEJ) [1] is a method to aggregate expert judgments based on performance measures. It is widely accepted and has been used in many fields including the nuclear sector, chemical and gas industry, hydraulic engineering, aerospace and aviation, occupational safety, health, banking and volcanology. Up to 2008 a total of 45 applications were collected in a database [4]. Since then, at least 33 more applications have been performed [5].

A complete description of the method is presented in the Supplementary Information (SI). In Cooke's Classical model, the experts assess their uncertainty over two types of continuous quantities. They do so by providing estimates of pre-defined quantiles of these uncertain quantities. Typically the 5th, 50th and 95th percentiles of experts' uncertainty distributions are elicited. The first type of uncertain quantity queried from experts corresponds to *target variables*. These are variables whose uncertainty cannot

* Corresponding author.

E-mail address: g.leontaris@tudelft.nl (G. Leontaris).

¹ In order to avoid confusion of the minority of people, who are not familiar with the universe of Lord of the Rings by J.R.R. Tolkien, the authors would like to clarify the inspiration for the name of the developed Matlab toolbox. Andúril was the name of the sword of Aragorn, the son of Arathorn, which was reforged from the shards of Narsil (the sword that was used by Isildur to cut the One Ring from Sauron's hand). Excalibur is also the name of the legendary sword of King Arthur. Similarly to the sword, the source code of EXCALIBUR software remained accessible only to a few worthy ones. Therefore, the researchers and practitioners could only admire and use the software without being able to further investigate and explore developments of the method. To change this, the existing software had to be "broken to pieces" and then "reforged". Naturally, the name of the resulting new open-source Matlab toolbox is ANDURIL. Hopefully, this will help in bringing peace to troubled researchers and practitioners of Cooke's classical model.

be sufficiently described using current models or field data and hence expert judgments are required. The second type of variables queried in the classical model are the so called *seed* or *calibration variables*. These are variables from the experts' field which are known to the analyst(s) at the moment of the elicitation (or will be known post hoc) but whose true values are not known to the experts at the moment of the elicitation. Experts are thus scored according to their performance in assessing uncertainty over seed variables. Their opinions are weighted and later combined on the basis of their performance. This resulting combined uncertainty distribution is called the *Decision Maker* (DM). According to [1], any methodology for structured expert judgment that aims at enabling rational consensus should be *scrutable*, subject to *empirical control*, *neutral* and *fair*.

In the majority of past studies, the closed source software EXCALIBUR² that is only available for Windows OS, has been used for the analysis and aggregation of expert judgments. Recently, a number of cross validation studies have been conducted using Eggstaff's MATLAB code [5,6]. However, this code is not publicly available and it still does not implement important features of the model such as the item weighting scheme [5]. Precisely in the spirit of contributing to warranty that the condition of scrutability is further met, the MATLAB toolbox presented in this paper was developed. We believe that it is important for researchers to have open access to a code that makes transparent the calculations of performance measures and the aggregation of expert judgments. In this way, the current methods can be made more accessible and different approaches or extensions to current methods can be further explored. Therefore, the purpose of ANDURIL toolbox is to assist researchers or practitioners who are interested in Cooke's classical model, in applying the method or investigating further developments to it, irrespective of their choice of operating system.

2. Software description

ANDURIL is a MATLAB toolbox that consists of different functions and supports the majority of the features of EXCALIBUR. Although the value of EXCALIBUR is undeniable, there are some limitations that stem from the fact that it is a closed source software. First, the understanding of the method is more difficult and time consuming for researchers who are recently introduced to the method. Moreover, it is impossible to modify it in order to expand its features or investigate different approaches for combination of expert judgments.

Cooke's method is essentially a linear pooling method (see Eq. (1)). For details regarding the method we refer the reader to the supplement of this paper and references therein. Here we sketch the main features of the method. Assume we have answers from $e = 1, \dots, E$ experts on $i = 1, \dots, N$ seed variables and target variables. The uncertainty distribution $f_{DM,i}$ per item for a DM is computed as:

$$f_{DM,i} = \frac{\sum_{e=1}^E w_{\alpha}(e) f_{e,i}}{\sum_{e=1}^E w_{\alpha}(e)} \quad (1)$$

where $f_{e,i}$ are expert's e probability densities constructed with her assessments of predefined quantiles per item i . The weights $w_{\alpha}(e)$ depend on two measures of performance computed with experts' answers to calibration questions. The first one is the *statistical accuracy* of experts' assessments. An expert will receive a non-zero weight if her score for statistical accuracy is above a certain confidence level α . Otherwise, the judgments of this particular expert e are not taken into account, and thus the probability densities of expert e do not contribute to the DM. The second performance

measure is the *information score* which indicates how "spread out" are experts' assessments with respect to a background measure. In general, we want experts with a high score for statistical accuracy and a high score for informativeness. ANDURIL's main function is to compute $f_{DM,i}$ by using a performance-based combination of individual judgments.

ANDURIL does not have a user interface yet, but there is a main script named ANDURIL_Main that can be used to enter the data obtained from expert judgments in order to conduct the desired analysis. The supported functionalities of Cooke's classical model in ANDURIL_Main are: (i) Calculation of DM using global weights; (ii) Calculation of DM using item weights; (iii) Calculation of DM using equal or user defined weights; (iv) Optimization of DM; (v) Robustness check itemwise; (vi) Robustness check expertwise; (vii) Plotting assessments itemwise; and (viii) Plotting robustness results. A description of the main functions of ANDURIL is given in Table 1. A more detailed explanation of every function can be found in the supplementary material.

3. Illustrative examples

ANDURIL has been validated with EXCALIBUR. For this purpose a recent structured expert judgment (SEJ) study concerning the estimation of GHG emissions in Mexico for 2020 and 2030 was used as a test case [7]. The part of the study that is used to validate ANDURIL is the one concerning the estimation of Gross Domestic Product. In this study 9 experts participated and provided the 5th, 50th and 95th percentiles of their uncertainty distribution regarding 13 seed variables and 6 target variables. The results obtained from applying ANDURIL to the test case are presented and compared with those obtained from EXCALIBUR. These results can be reproduced by using the ANDURIL_example script and the .dtt and .rls files of EXCALIBUR provided as a supplement.

Five different DMs were calculated using ANDURIL: (i) The global weight decision maker (DM_1), calculated using the function calculate_DM_global, (ii) the item weight (DM_2) using the function calculate_DM_item, (iii) the equal weight (DM_3) calculated using the function calculate_DM_global with equal weights for every expert, (iv) the optimized global weight decision maker (DM_4) which was calculated using the function DM_optimization and (v) the user weight (DM_5) which was calculated using the function calculate_DM_global while giving to expert 5 and 6 weights equal to 0.4 and 0.6 respectively. It should be noted that the background measure for every item was chosen as uniform. However, the same DMs were calculated and validated when the log-uniform background measure was used for every item.

The comparison of the obtained quantiles using ANDURIL and EXCALIBUR is presented in Table 2. As it can be seen, there are very small differences between the output of EXCALIBUR and ANDURIL due to differences in the precision of the calculating engine. Particularly, the maximum difference is 0.0005 in absolute value across the quantiles of the DMs of interest.

Furthermore, Fig. 1 shows the comparison of the obtained plots for every individual expert and DMs (DM_1 , DM_2 and DM_3) concerning seed item 5. The plots of ANDURIL were produced using the function plotting_itemwise and show that the same results are obtained with EXCALIBUR.

4. Impact

ANDURIL can be used by practitioners and researchers to apply and investigate aspects of Cooke's classical model. Some limitations of the existing closed-source software EXCALIBUR have been investigated by using ANDURIL and are presented in this section. In

² EXCALIBUR is freely available at <http://www.lighttwist.net/wp/excalibur>.

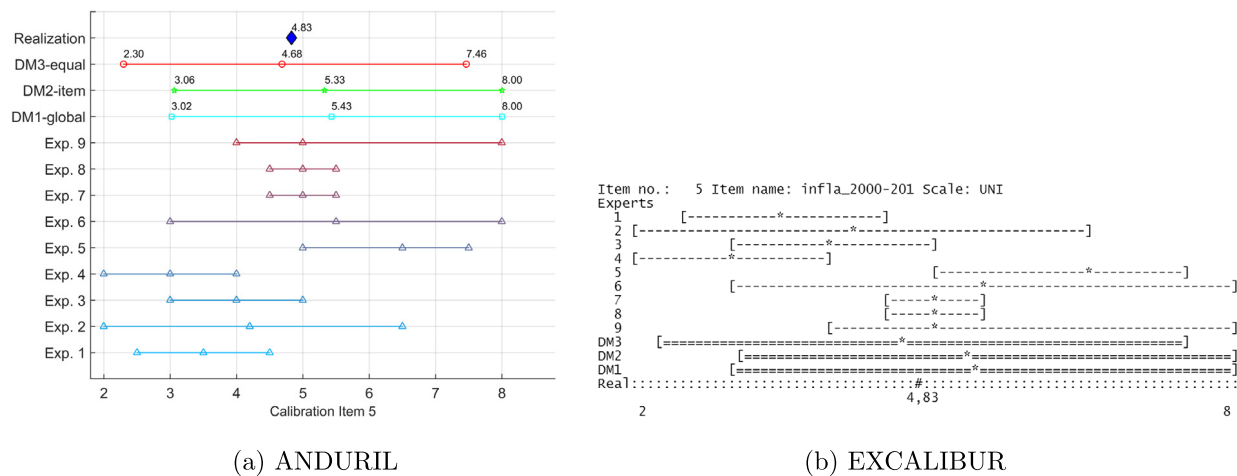


Fig. 1. Comparison of obtained plots for the assessments of all experts and DMs concerning seed item 5.

Table 1

Main functions of ANDURIL.

Function's name and description
<code>calscore</code> : Calculates the statistical accuracy (or calibration score) of expert e over the set of <i>seed items</i> (eq. 2 in supplement).
<code>calculate_information</code> : Calculates the relative information (or information score) of expert e over the set of <i>seed items</i> as well as the information score of every expert over all items (eq. 3 in supplement).
<code>global_weights</code> : Calculates the calibration score, the information score over the seed items and subsequently the weight of every expert e .
<code>calculate_DM_global</code> : Calculates the distribution of the DM for every item, using the global weights or equal weights weighting schemes.
<code>item_weights</code> : Calculates the weights of every expert e for every item. The main difference with the global weights weighting scheme is that the weights are different for every item. In this way, the opinion of every expert has a different weight for every item. This is achieved by using the relative information of every particular item.
<code>calculate_DM_item</code> : Calculates the distribution of the DM for every item using the item weights weighting scheme.
<code>DM_optimization</code> : Calculates the distribution of the DM for every item using the significance level α that optimizes the DM in terms of statistical accuracy.
<code>Checking_Robustness_items</code> : Calculates the performance measures (calibration score, information score over seed variable and over all variables with respect to the background measure) of the DM that occurs when up to N_{\max_it} seed item(s) are excluded at most. It calculates the performance measures for every possible combination, starting from excluding one up to N_{\max_it} seed items at a time.
<code>Checking_Robustness_experts</code> : Calculates the performance measures of the DM that occurs when up to N_{\max_ex} expert(s) are excluded at most, similarly to <code>Checking_Robustness_items</code> .
<code>plotting_itemwise</code> : Produces as many plots as the total number of items (i.e. seed and target items). Every plot presents the assessments (i.e. 5 th , 50 th , 95 th percentiles) of every expert e as well as every DM, for every particular item i .
<code>robustness_plots</code> : Produces three box plots. Each box plot corresponds to one measure of performance in judging uncertainty. Namely statistical accuracy, information score over all items and information score over seed items. Each box plot presents how the values of every measure vary with the number of excluded items (x-axis). In these plots a horizontal line is also plotted, that shows the values of the DM whose robustness is under investigation.

particular, we discuss how limitations regarding the *intrinsic range*, *item weights*, *distributions of DMs* and *robustness* can be overcome with the use of ANDURIL. More information can be found in the provided supplementary material.

Intrinsic range. The bounds of the intrinsic range for every item i (i.e. q_{li} and q_{hi} in the supplement) are calculated by considering the assessments of every expert, even the ones with zero weights. Moreover, the intrinsic range for a calibration item takes into consideration the realization of the seed variable. One could argue that for the calculation of the DM's distribution only the assessments of the experts with non-zero weights could be used. This is not possible to be investigated using EXCALIBUR.

Table 2

Comparison of the four DMs' quantiles regarding seed item 5 using ANDURIL and EXCALIBUR.

Name	EXCALIBUR			ANDURIL		
	q_5	q_{50}	q_{95}	q_5	q_{50}	q_{95}
DM_1	3.02	5.431	8.000	3.0201	5.4311	8.000
DM_2	3.063	5.327	8.000	3.0633	5.3275	8.000
DM_3	2.297	4.684	7.463	2.2971	4.6840	7.4626
DM_4	3.021	5.44	7.999	3.0209	5.4395	7.9994
DM_5	3.098	6.026	7.928	3.0978	6.0263	7.928

For this reason, the `calculate_DM_global` function of ANDURIL was modified in order to investigate the effect of calculating the intrinsic ranges of every item by: (i) taking into account the

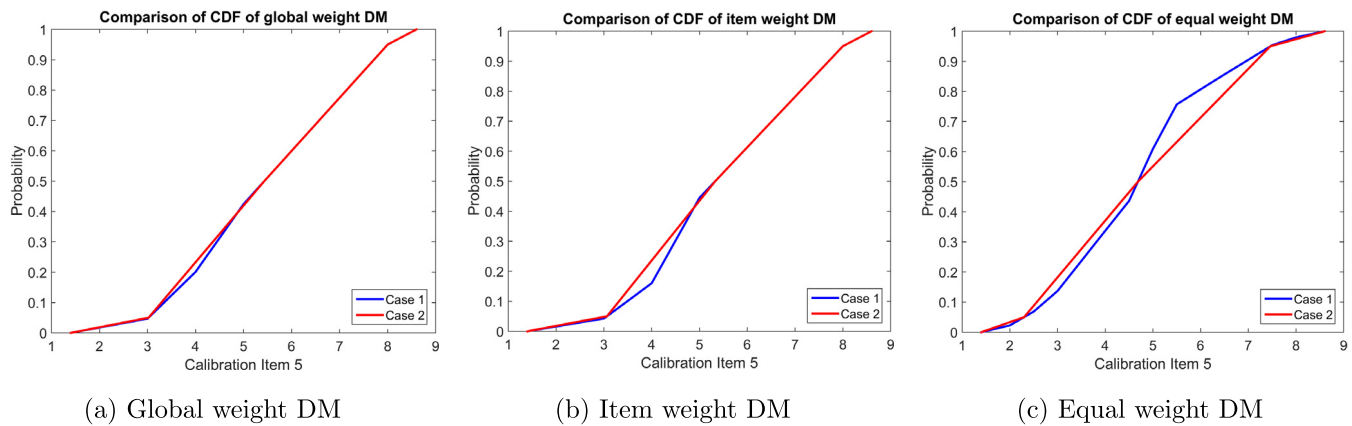


Fig. 2. Comparison of output cumulative distributions obtained by integration (Case 1) and interpolation (Case 2) concerning (a) global weights, (b) item weights and (c) equal weights.

realization and the judgments of only those experts with non-zero weights (that produces DM1_alt1); and (ii) taking into account only the judgments of the experts with non-zero weights (that produces DM1_alt2). This new function was named `alter_calc_DM_global`. It should be noted that in order to investigate the effect of these alternative calculations on the DM_2 , the `calculate_DM_item` should be modified. Significant differences were observed, especially (as expected) in quantiles q_h and q_l of every item. This issue has not been discussed in literature, for example in those related to out of sample performance of Cooke's model [5,6]. This is a subject that could be further explored with the aid of ANDURIL.

Item weights. When the *item weights* weighting scheme is used to combine the expert judgments, the information score of the obtained DM and the weight from EXCALIBUR are calculated using global weights [1]. Therefore, it is not possible for the user to know the exact weights that were used per item. On the other hand, the `item_weights` function of ANDURIL provides the user with tables W_{itm} and W_{itm_tq} which contain the weights of each expert concerning the seed variables and target variables, respectively.

Distributions of DMs. The cumulative distribution of a DM is calculated by integrating the density of the DM. To achieve this, all the values of the quantiles of the experts with non-zero weights are taken into account and the cumulative probability of every unique value is computed. Hence, the $q_{i,5}$, $q_{i,50}$ and $q_{i,95}$ quantiles of the DM are obtained. In EXCALIBUR the output distributions of the DMs are calculated by linear interpolation between these three quantiles (i.e. $q_{i,5}$, $q_{i,50}$ and $q_{i,95}$) of the DM. This may lead to differences between the distributions obtained by integration (Case 1 in Fig. 2) and the distributions that are obtained by interpolating in between quantiles (Case 2 in the same figure). Functions `calculate_DM_global` and `calculate_DM_item` of ANDURIL provide the user with the DM distributions containing the quantiles of experts with non-zero weights.

Fig. 2(a), 2(b) and 2(c) present the two different distributions of DMs concerning seed item 5, combined with global, item and equal weights weighting schemes respectively. From these plots, it can be seen that interpolating linearly between $q_{i,5}$, $q_{i,50}$ and $q_{i,95}$ to obtain a distribution for the DM may cause significant variations in the resulting distributions, especially when the equal weight combination is considered. The integrated cumulative distribution contains more linear components since every percentile provided by every expert is considered in the density.

Robustness itemwise. When investigating the robustness of the obtained DM, EXCALIBUR supports the exclusion of only one item at a time for re-calculation. Hence, it is not possible to investigate

how the performance measures (i.e. statistical accuracy and information scores) vary as more than one item are excluded at a time. For this reason, `Checking_Robustness_items` and `robustness_plots` functions of ANDURIL were developed. The latter produces three box plots. Each plot corresponds to one measure of performance in judging uncertainty. Namely statistical accuracy, information score over all items and information score over seed items. The statistical accuracy score depends on the number of items, and hence a "calibration power" is introduced in order to make the robustness analysis more comparable when items are left out (for details see the SI). Examples concerning the presented case can be found in Fig. 3(a) and 3(b) for statistical accuracy and information score (over seed items), respectively. These measures were calculated while keeping the "calibration power" equal to one.

Each plot presents how the values of every measure vary with the number of excluded items (horizontal axis). In these plots, a green horizontal line that shows the values of the initial DM whose robustness is under investigation. A magenta marker shows the geometric mean for every number of removed items. It should be noted that when the number of excluded seed items increases, then there is the possibility that for some combinations (of excluded seed items) the calibration score of all experts reduces below the significance level α , resulting in zero weights for every expert. Hence, these situations are not considered. As it can be seen in Fig. 3(a) and 3(b), although the interval containing 95% of the recalculated scores increases as more items are removed at a time, the median remains close to the original value (shown by the green horizontal line) for every measure of performance.

5. Conclusions

A MATLAB toolbox named ANDURIL was developed to support decision making under uncertainty, when expert judgments are combined by applying Cooke's classical model for structured expert judgment. The main purpose for developing this toolbox is to create an open source software that can be used by practitioners and researchers who are interested in applying or developing further Cooke's method. The tool was validated with the closed source software EXCALIBUR. For this purpose a recent study concerning green house gases emissions in Mexico was used as a test case. It was shown that ANDURIL can reproduce accurately the results of EXCALIBUR.

The advantages of having a transparent open source software for applying Cooke's method were discussed. The developed toolbox can be used to investigate different ways of calculating the intrinsic range of the aggregated opinions that may result in differences in the performance measures of the obtained DMs.

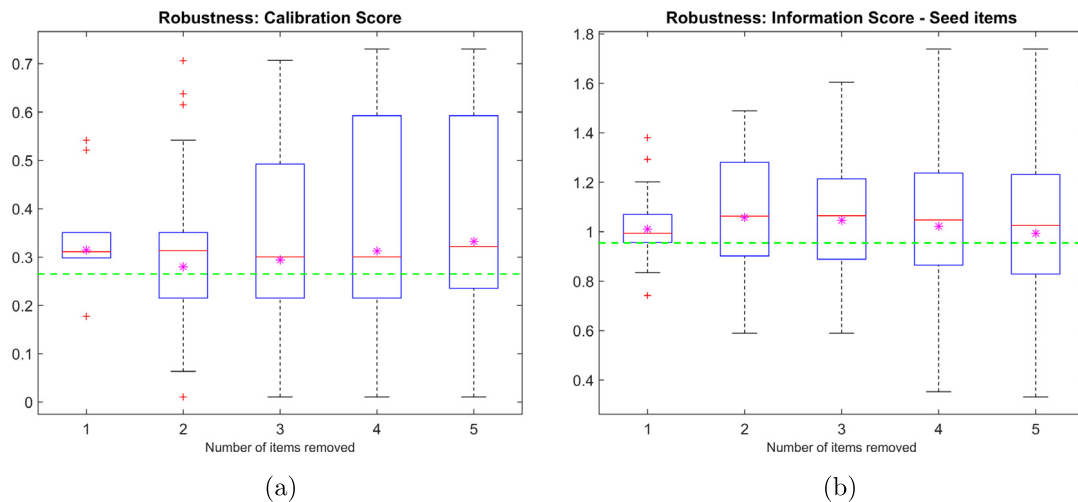


Fig. 3. Robustness plots, obtained using ANDURIL, concerning (a) calibration score and (b) information score (over seed items) with respect to the number of excluded seed items.

Moreover, it is possible to provide the analyst with the weights of each expert per item when the item weights weighting scheme is considered. Also, it gives the opportunity to the user to calculate the integrated cumulative distribution of the DM considering in the density every percentile provided by every expert with non-zero weights, rather than just interpolating in between the 5th, 50th and 95th percentiles of the DM. Finally, the robustness of the obtained DM can be investigated while excluding more than one seed item/expert at a time.

Concluding, the authors want to stress that the developed tool constitutes a first step towards an open source version of Cooke's classical model. Despite the limitations of the current version of ANDURIL, it is to the authors belief that the developed toolbox will be valuable to those who are interested in developing and further applying the method. It is the ambition of the authors to extend ANDURIL with more features that are currently available in EXCALIBUR and with the more recent techniques of elicitation of multivariate dependence.

Acknowledgments

This research is part of the EUROS research programme, which is supported by NWO domain Applied and Engineering Sciences and partly funded by the Dutch Ministry of Economic Affairs. We also acknowledge the support of COST Action IS1304 “Expert Judg-

ment Network: Bridging the Gap Between Scientific Uncertainty and Evidence-Based Decision Making”.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.softx.2018.07.001>.

References

- [1] Cooke RM. Experts in uncertainty: Opinion and subjective probability in science. In: *Environmental ethics and science policy*, Oxford University Press; 1991.
- [2] O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain judgements: Eliciting experts' probabilities. In: *Uncertain judgements: Eliciting experts' probabilities*. John Wiley and Sons, Ltd; 2006.
- [3] Dias LC, Morton A, Quigley J, editors. Elicitation: The science and art of structuring judgement. New York: Springer; 2017.
- [4] Cooke RM, Goossens LL. TU Delft expert judgment data base. *Reliab Eng Syst Saf* 2008;93(5):657–74.
- [5] Colson AR, Cooke RM. Cross validation for the classical model of structured expert judgment. *Reliab Eng Syst Saf* 2017;163:109–20.
- [6] Eggstaff JW, Mazzuchi TA, Sarkani S. The effect of the number of seed variables on the performance of Cooke's classical model. *Reliab Eng Syst Saf* 2014;121:72–82.
- [7] Puig D, Morales-Nápoles O, Bakhtiari F, Landa G. The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. *Climate Policy* 2018;18(6):742–51.