Reliability Engineering and System Safety 121 (2014) 72–82

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

The effect of the number of seed variables on the performance of Cooke's classical model

Justin W. Eggstaff*, Thomas A. Mazzuchi, Shahram Sarkani

School of Engineering & Applied Science, The George Washington University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form
14 May 2013
Accepted 27 July 2013
Available online 7 August 2013

Keywords:
Expert judgment
Cooke's classical model
Scoring rule
Expert aggregation
Risk analysis
Seed variables

ABSTRACT

In risk analysis, Cooke's classical model for aggregating expert judgment has been widely used for over 20 years. However, the validity of this model has been the subject of much debate. Critics assert that this model's scoring rule may unintentionally reward experts who manipulate their quantile estimates in order to receive a greater weight. In addition, the question of the number of seed variables required to ensure adequate performance of Cooke's classical model remains unanswered. In this study, we conduct a comprehensive examination of the model through an iterative, cross validation test to perform an out-of-sample comparison between Cooke's classical model and the equal-weight linear opinion pool method on almost all of the expert judgment studies compiled by Cooke and colleagues to date. Our results indicate that Cooke's classical model significantly outperforms equally weighting expert judgment, regardless of the number of seed variables used; however, there may, in fact, be a maximum number of seed variables beyond which Cooke's model cannot outperform an equally-weighted panel.

Published by Elsevier Ltd.

1. Introduction

Obtaining advice from experts is a universal heuristic that is used to gain insight from those who are considered to have superior knowledge within their field in an attempt to examine or assess some unknown variable or variables. Codifying and combining expert knowledge to represent the uncertainty or risk in decision analysis is of particular interest when data, with which a model can be established or sufficient statistical inference made, are not available [1]. Relatively easy to understand and implement, Cooke's *classical model* for aggregating expert judgment has been widely used in the field of risk analysis for over 20 years [2–5]. Cooke's model, firmly rooted in sound mathematical principles, evaluates expert responses to training questions (called *seed* variables) in order to determine the accuracy and variability in their performance. Weights are then assigned to each expert and a combined distribution is established and applied to, as yet, unknown variables of interest (called *target* variables).

Undoubtedly, Cooke's classical method has been subjected to much academic rigor. Clemen [2] opened the most recent assessments of the classical model by suggesting the use of out-of-sample evaluations between Cooke's *performance weighted decision maker* (*PWDM*) versus an *equally weighted decision maker* (*EWDM*). He asserted that because the same set of data used to calculate the consolidated decision maker weighting are used again to measure the method's performance, the assessment may be skewed in the classical method's favor [2]. Looking at 14 studies from Cooke and Goossen's [4] database, Clemen determined that the performance-based method did not appear to be better than simply equally weighting expert judgments. Because Clemen's assessment had only been performed on a subset of Cooke's database, Lin and Cheng [6] extended this analysis by conducting leave-one-out cross-validation tests on almost all of the studies acquired by Cooke and Goossen at the time. Their results showed that the classical model did maintain a slight advantage over the equal weight method [6]. However, they opined that this slight advantage may not justify its use when faced with the added effort (and cost) required for its implementation.

Conversely, Cooke suggested that a leave-one-out or remove-one-at-a-time (ROAT) cross validation may be too narrow in scope and inadvertently reward or penalize the pool of experts [7,8]. By excluding only one seed variable, there may be a tendency to favor the experts who assessed that particular seed poorly while punishing the experts who assessed that particular seed well. Instead, Cooke presented a two-fold cross validation whereby each data set was equally bifurcated and analysis performed. In 20 of 26 validation runs, the PWDM out-performed the EWDM. Flandoli et al. [9] performed a similar analysis on data from five studies by conducting a two-fold cross validation after splitting the datasets into combinations of 70% seed variables and 30% target variables. Their results showed that Cooke's model provided the best indication of uncertainty when

* Correspondence to: 43636 Wild Iris Street, California, Maryland 20619, USA. Tel.: +1 703 655 1540.
E-mail address: eggstaff@gwmail.gwu.edu (J.W. Eggstaff).

0951-8320/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.ress.2013.07.015

averaged across the over 500 possible combinations. These results upheld Cooke′s call for a two-fold cross validation. However, a more complete analysis seems necessary.

In all the studies cited, the problem of the number of seed variables required to adequately employ Cooke′s classical model remains an open question. While no definitive answer has been provided in current literature, Clemen [2] suggested that a "safe" threshold of seed variables is most likely greater than 10. Recently, Lin and Huang [8] suggested that the number of seed questions used did have an effect on the calibration of aggregated expert judgment distributions. Thus, in this study, we conduct a comprehensive analysis on almost all of the expert judgment studies compiled by Cooke and colleagues to date and introduce a novel iterative, cross validation test to explore the seed variable requirement. This test is then used to compare the performance between Cooke's classical model, the PWDM, and a simple average of expert judgment, the EWDM. The results of this study will be important to the field of risk analysis by expanding the utility of expert judgment aggregation methods. We purposely omit the performance of the best expert in this study due to the immense computational workload and the seemingly overwhelming evidence suggesting that method's inferiority [2,4–6,8,9].

In Section 2, the methods used to conduct this study are discussed, to include Cooke's classical model, the iterative cross validation technique, and the methods of comparison. In Section 3, the results of our analysis are presented. In Section 4, we discuss our conclusions and the utility to the field of risk analysis.

2. Materials and method

2.1. Combining expert judgment

The use of expert judgment to aid in forecasting and decision analysis became a focus of research after World War II [3]. Expert judgment is of particular interest when data are not available to adequately perform statistical analysis and/or simulate [1,10]. Several techniques to quantify and aggregate expert judgment have been proposed throughout the years and are generally categorized as either *mathematical* or *behavioral* [11]. Behavioral techniques, such as the *Delphi method* [12], are dubious because they can become infected by psychological biases that may affect the validity of the technique's results [3]. Mathematical techniques, on the other hand, seek to eliminate these biases while providing some statistical inference that can explicitly quantify the level of uncertainty of some future outcome [13].

Mathematical methods of combination can be classified as either *axiomatic* or *Bayesian*. Several Bayesian approaches that associate a likelihood function to the expert's judgment are called out in literature, but they are considered difficult to apply [1]. By contrast, axiomatic approaches are considered relatively straightforward and can be easily calculated. The simplest example is the *linear opinion pool* which is simply a weighted sum of individual distributions. If we let $p_i(\theta)$ represent each expert i's assessment or probability distribution, then $p(\theta)$ is the aggregated distribution where:

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta) \tag{1}$$

where n is the total number of experts, and w_i represents each expert′s normalized weight [14]. Another approach is the *logarithmic opinion pool* which uses multiplicative averaging. The difficulty when exercising these approaches lay in the appropriate assignment of weights to each expert [15].

If we were to ask an expert to assess the probability of occurrence for the value of a variable, a well-calibrated expert′s assessment would closely match the value′s true probability distribution. Ideally, we would seek out a pool of experts considered equally knowledgeable about the variables to be assessed and, therefore, naturally apportion equal weights. However, Hora [16] has shown both theoretically and empirically that combining well-calibrated experts by equal weighting may, in fact, decrease calibration. In addition, it seems that we must rely on a less than 'perfect' panel of experts and place confidence in the experts who are perceived as superior [17]. As a result, *scoring rules* were developed to assess the quality of judgments. Several scoring rules have been developed throughout the years and are the subject of current literature [8,13,18–21]. Scoring rules are typically designed to reward or incentivize the submission of accurate opinions while penalizing inaccurate ones. A *proper scoring rule* is one which maximizes a respondent's score or reward only when the elicited judgment equals the actual value of the variable of interest. The *classical model*, developed by Cooke [3], is probably the most widely used method for combining expert judgment and has been applied in a variety of applications, including its extensive use in risk analysis [2,4,10]. The classical model is based upon statistical hypothesis testing and conforms to the principle of proper scoring rules [3].

2.2. Cooke's classical model

In Cooke's classical model, an expert's weight is determined through the combination of that expert's relative *calibration* and *information* scores obtained from a set of *seed* variables embedded within a population of unknown quantities being elicited. The true values of the seed variables are known only to the *decision maker*. The performance measures are then used to generate each expert's relative weight. The classical model is calculated again using the quantities elicited for the *target* variables (this time without the benefit of knowing the true values) and combined in a linear opinion pool. The result is a probability distribution, the *decision maker's assessment*, for each target variable. Although summarized below, a complete explanation of Cooke's classical model can be found in [3].

Experts are asked to provide $Q=\{q_1, q_2,\ldots,q_Q\}$ percentile values which bound $R=\{r_1, r_2,\ldots,r_R\}$ probability bins. Over the set of seed variables N, an aggregated quantile distribution judgment, $S(e)=\{s_1, s_2,\ldots,s_R\}$, is produced and compared to the theoretical probability distribution, $P=\{p_1, p_2,\ldots,p_R\}$. The number and distribution of elicited quantiles can be tailored to the decision maker's requirements. An expert e is considered well-calibrated when the elements of his or her sample distribution s_i (the number of realizations that fall in the ith probability bin) resemble P. As an illustrative example, an expert may be asked to provide the 5th, 50th, and 95th percentiles. These percentiles produce four probability bins with a theoretical probability distribution $P=\{0.05, 0.45, 0.45, 0.05\}$. An expert is considered well-calibrated when approximately 5% of the known values for the seed variables are lower than the 5th percentile judgment given by the expert; 45% of the known values for the seed variables lay between the 5th and 50th percentile interval; 45% of the known values for the seed variables lay between the 50th and 95th percentile interval; and 5% of the known values for the seed variables are above the 95th percentile.

Given R quantiles, the *relative information*, $I(S(e),P)$, is calculated by:

$$I(S(e),P) = \sum_{i=1}^{R} s_i \ln\left(\frac{s_i}{p_i}\right) \tag{2}$$

Given the number of seed variables N, it can be shown that the distribution of $2NI(S(e),P)$ can be approximated by a chi-squared distribution with Q degrees of freedom, $\chi^2_{(Q)}$ [22]. A minimum acceptable calibration score or significance level, α, below which an expert is given zero weight in the analysis, is now introduced. The value for α is chosen such that if the decision maker's assessment were inserted into the expert panel and scored