Reliability Engineering and System Safety 163 (2017) 109–120

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Cross validation for the classical model of structured expert judgment

Abigail R. Colson[a,b], Roger M. Cooke[c,d,e,*]

[a] Center for Disease Dynamics, Economics & Policy, Washington, DC, USA
[b] University of Strathclyde, Glasgow, UK
[c] Resources for the Future, Washington, DC, USA
[d] University of Strathclyde, Glasgow, UK
[e] TU Delft (ret), Delft, The Netherlands

ARTICLE INFO

Keywords:
Expert judgment
Calibration
Information
Classical model
Out-of-sample validation

ABSTRACT

We update the 2008 TU Delft structured expert judgment database with data from 33 professionally contracted Classical Model studies conducted between 2006 and March 2015 to evaluate its performance relative to other expert aggregation models. We briefly review alternative mathematical aggregation schemes, including harmonic weighting, before focusing on linear pooling of expert judgments with equal weights and performance-based weights. Performance weighting outperforms equal weighting in all but 1 of the 33 studies in-sample. True out-of-sample validation is rarely possible for Classical Model studies, and cross validation techniques that split calibration questions into a training and test set are used instead. Performance weighting incurs an "out-of-sample penalty" and its statistical accuracy out-of-sample is lower than that of equal weighting. However, as a function of training set size, the statistical accuracy of performance-based combinations reaches 75% of the equal weight value when the training set includes 80% of calibration variables. At this point the training set is sufficiently powerful to resolve differences in individual expert performance. The information of performance-based combinations is double that of equal weighting when the training set is at least 50% of the set of calibration variables. Previous out-of-sample validation work used a Total Out-of-Sample Validity Index based on all splits of the calibration questions into training and test subsets, which is expensive to compute and includes small training sets of dubious value. As an alternative, we propose an Out-of-Sample Validity Index based on averaging the product of statistical accuracy and information over all training sets sized at 80% of the calibration set. Performance weighting outperforms equal weighting on this Out-of-Sample Validity Index in 26 of the 33 post-2006 studies; the probability of 26 or more successes on 33 trials if there were no difference between performance weighting and equal weighting is 0.001.

1. Introduction

Structured expert judgment denotes techniques for using expert judgments as scientific data. A recent overview dates its inception to large scale engineering studies from 1975 [9]. Cooke et al. [13] first proposed the use of calibration (here called "statistical accuracy") and information to score experts' performance, and the use of these scores for defining and validating schemes combining experts' judgments is termed the Classical Model [6]. By 2006, analysts had conducted 45 professionally contracted Classical Model studies. Cooke and Goossens [12] summarized and published the results from these studies, and made the data, called the TU Delft database, available to the research community. The studies in the TU Delft database include those from the dawn of the Classical Model, and their study designs differ wildly. The number of experts in a given study ranged from 4 to 77 and the number of calibration variables (i.e., questions from the field for which realizations are known post hoc; these questions are the basis for creating performance-based combinations of the experts' assessments) ranged from 5 to 55.

The TU Delft database allows researchers to explore how experts and different combinations of experts perform on data from real expert judgment applications. Researchers have used this data to investigate how the performance-weight (*PW*) combinations of the Classical Model compare to equal-weight (*EW*) combinations of experts both in-sample and out-of-sample. Cooke and Goossens [12] demonstrated that *PW* is superior to *EW* on in-sample comparisons, in which the same set of data is used to both initialize and validate the model. Clemen [5] first raised the question of the Classical Model's out-of-sample validity, using the TU Delft database to explore if performance-based combinations predict out-of-sample items better than equally weighted combi-

* Corresponding author.
E-mail address: cooke@rff.org (R.M. Cooke).

http://dx.doi.org/10.1016/j.ress.2017.02.003
Received 28 January 2016; Received in revised form 7 February 2017; Accepted 18 February 2017
Available online 24 February 2017

nations of the experts. In recent years other researchers have proposed various methods for validation of the Classical Model and drawn conflicting conclusions.

Since 2006 use of the Classical Model has continued to expand, thanks in large part to high-profile applications (for example, [1]). Over thirty three independent expert judgment studies were performed between 2006 and March 2015. These studies were contracted by a variety of organizations including: Bristol University (UK), the British government, the European Commission, PrioNet (Canada), Public Health Canada, the Robert Wood Johnson Foundation, Sanguin, the US Department of Homeland Security, and the US Environmental Protection Agency. In these recent studies, panels of 4–21 experts assessed between 7 and 17 calibration variables. These studies are generally better resourced, better executed, and better documented than the very early Classical Model applications.

Updating the 2006 database and establishing a baseline for the in- and out-of-sample validation of performance based weighting is timely and important. The recent report of the National Academy of Sciences on the social cost of carbon lends urgency to this effort, noting "*performance-weighted average of distributions usually outperforms the simple average, where performance is again measured again by calibration and informativeness (and is often evaluated on seed variables not used to define the weights, because the value of the quantity of interest in many expert elicitation studies remains unknown)*" [27, p. 339].

Another recent spur is the 5-year forecasting tournament organized by IARPA of which Philip Tetlock's Good Judgment Project was proclaimed the winner. The tournament concerned current events assessed by "ordinary citizens" as opposed to quantification of scientific/engineering uncertainties. Radically down-selecting from a pool of more than 3000,[1] Tetlock's group distilled a small group of "superforecasters" based on their performance. Although very different in purpose and method to the Classical Model, the Good Judgment Project strongly underscores the value of performance based combinations.

In this study we use data from 33 post-2006 studies (Described in Supplementary Online Material 1) to explore the in-sample and out-of-sample validity of the Classical Model. Based on the post-2006 data, we test the null hypothesis that performance-weight (*PW*) combinations of the experts are no better than equal-weight (*EW*) combinations in terms of statistical accuracy and informativeness. Finally, we develop an Out-of-Sample Validity Index (OoSVI) which can be used to validate future Classical Model studies and related research.

The 33 post-2006 studies considered here excludes two sets of post-2006 applications. One concerns an ongoing expert elicitation program at the Montserrat Volcano Observatory that has produced a wealth of data on expert performance [29,3]. The second is a recently completed large scale study by the World Health Organization involving 72 experts spread over 134 distinct panels [2,20]. Since both sets of studies involve heavily overlapping expert panels, they do not lend themselves to the present analysis where the panels are considered independent.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the Classical Model and reviews alternate pooling schemes, comparing their statistical accuracy across the post-2006 data. Section 3 summarizes the in-sample properties of the post-2006 data. Section 4 reviews previous out-of-sample validation research based on the TU Delft database, and Section 5 summarizes the out-of-sample performance of the newly collected post-2006 data. Section 6 provides two detailed case studies that demonstrate good and poor out-of-sample performance. Section 7 evaluates the hypothesis that *PW* is no better than *EW* out-of-sample. Section 8 compares the present results with those of Eggstaff et al. [16] and a final section gathers conclusions.

The Supplementary Online Material (SOM) provides: (1) references and information on the 33 post-2006 applications analyzed here, (2) a detailed description of the Classical Model, (3) more information on quantile averaging in the post-2006 dataset, (4) improved exposition of proofs of the scoring rule properties (adapted from Cooke [6]), (5) additional details on previous cross validation research, and (6) an expanded list of references for applications of the Classical Model.

2. Aggregating expert judgments

2.1. The Classical Model

In the Classical Model, experts quantify their uncertainty regarding two types of questions. The variables of interest are the target of the elicitation; these questions cannot be adequately answered by existing data or models, so expert judgment is needed as additional evidence. Calibration variables (also termed seed variables) are questions from the experts' field which are unknown to the experts at the time of the elicitation, but whose true values will be known post hoc. Experts are scored and weighted according to their calibration and information, and their assessments are combined into a *PW* decision maker, which can be compared to an *EW* decision maker. The calibration and information scores are briefly discussed below, and more detail is available in SOM 2.

In the context of expert judgment, the term "calibration" gives engineers and scientists the false impression that the judgments of experts are being "adjusted," as they would calibrate instruments by adjusting their scales. This is not the case. Since calibration is only loosely defined in decision theory literature, this confusion is best avoided by replacing "calibration" with "statistical accuracy," defined as the P-value at which one would falsely reject the hypotheses that a set of probability assessments were statistically accurate. Very crudely, it answers questions like "how likely is it that at least 7 out of 10 realizations should fall outside an expert's 90% confidence bands, if each value really had an independent 90% chance of falling inside the bands?"

Information is measured as Shannon relative information with respect to a user supplied background measure. Shannon relative information is used because it is scale invariant, tail insensitive, slow, and familiar. The combined score, the product of statistical accuracy and informativeness, satisfies a long run proper scoring rule constraint and involves choosing an optimal statistical accuracy threshold beneath which experts are unweighted. Weights for the *PW* decision maker are based on this combined score, as described in SOM 2.

The Classical Model's performance measures of statistical accuracy and information do not map neatly onto the terms "accuracy" and "precision", which are familiar to social scientists. Accuracy denotes the distance between a true value and a mean or median estimate, and precision denotes a standard deviation. While appropriate for repeated measurements of similar variables, these notions are scale dependent and therefore not useful in aggregating performance across variables on vastly different physical scales. For example, how should one add an error of 10^9 colony forming units of campylobacter infection to an error of 25 micrograms per liter of nitrogen concentration? Expert judgments frequently involve different scales, both within one study and between studies. For this reason, the performance measures in the Classical Model are scale invariant. That said, the exhaustive out-of-sample analysis of Eggstaff et al. [16] (described in Section 4) found that the realizations were closer to the *PW* combination's median than the *EW* combination's median in 74% of the 75 million out-of-sample predictions based on the TU Delft data. Such non-parametric ordinal proximity measures, proposed by Clemen [5] are not used to score expert performance, as the scores strongly depend on the size of the expert panels. Thus, the present study focuses on the standard Classical

[1] Full documentation is not available at this writing and the information here is based on http://www.npr.org/sections/parallels/2014/04/02/297839429/-so-you-think-youre-smarter-than-a-cia-agent accessed 1/12/2017 and [31].