ARTICLE IN PRESS
International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

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

An investigation of dependence in expert judgement studies with multiple experts

Kevin J. Wilson *

Department of Management Science, University of Strathclyde, UK

ARTICLE INFO

Keywords:
Expert judgement
Dependency
Mathematical aggregation
Bayesian methods
Finance

ABSTRACT

Expert judgement plays an important role in forecasting and elsewhere, as it can be used to quantify models when no data are available and to improve predictions from models when combined with data. In order to provide defensible estimates of unknowns in an analysis, the judgements of multiple experts can be elicited. Mathematical aggregation methods can then be used to combine these individual judgements into a single judgement for the decision maker. However, most mathematical aggregation methods assume that such judgements come from experts who are independent, which is unlikely to be the case in practice. This paper investigates dependence in expert judgement studies, both within and between experts. It provides the most comprehensive analysis to date by considering all studies in the TU Delft database. It then assesses the practical significance of the dependencies identified in the studies by comparing the performances of several mathematical aggregation methods with varying dependence assumptions. Between-expert correlations were more prevalent than within-expert correlations. For studies that contained between-expert correlations, models which include these produced better forecasts. The implications of this for the use of expert judgement in forecasting are discussed.

1. Introduction

Expert judgement has been used in forecasting both informally, when no data are available, and formally, to bound problems, qualitatively structure models and quantify unknowns within models. One example by DeWispelare, Herren, and Clemen (1995) used expert judgements to estimate the probabilities within high-level nuclear waste regulation. Probabilistic forecasts were elicited from five climatologists on parameters of complex climatic models. In a very different application, Barndt, Freeman, and Schrodt (2014) considered the evaluation of forecasts in political conflict dynamics. They considered density forecasts from analysts and the importance of taking the uncertainty in forecasts into account when evaluating models. Jochmann, Koop, and Strachan (2010) considered the use of expert judgement VAR forecasting in combination with data. They identified model parameters for which prior uncertainty distributions could be elicited from experts and indicated that this approach offers improvements in accuracy over solely data-driven models. Thus, expert judgement can be used either on its own when there are no data available or in combination with data in a Bayesian analysis. In both cases, we require a procedure for obtaining the expert judgements that we require.

An important question in any expert judgement study is whose judgements should be elicited. That is, what constitutes an expert? There are various views on this. In a pure subjectivist Bayesian analysis, an expert could simply be the person from whom the unknowns are being elicited. However, if we are considering the expert

* Correspondence to: School of Mathematics and Statistics, Newcastle University, UK.

E-mail address: kevin.wilson@ncl.ac.uk.

http://dx.doi.org/10.1016/j.ijforecast.2015.11.014
0169-2070/

Please cite this article in press as: Wilson, K. J., An investigation of dependence in expert judgement studies with multiple experts. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.014

problem (French, 2011), in which experts are being asked for advice by a specific decision-maker, then the choice of an expert will require more justification. In this case, we could use the definition of Garthwaite, Kadane, and O'Hagan (2005) that experts are "persons to whom society and/or his peers attribute special knowledge about the matters being elicited". Crucially, it is also the ability to use this knowledge that defines a good expert (O'Hagan et al., 2006). For a discussion of the selection of experts, see Sections 1.3 and 5 of Garthwaite et al. (2005). In the expert problem, multiple experts are typically used to improve the information given to the decision maker.

When performing elicitation, questions must always be asked about quantities that are relevant to the expert, rather than abstract model parameters. Typically, questions are put to experts about probabilities or quantiles, and these are then converted to the parameters of probability distributions by the analyst. However, there is considerable evidence in the psychological literature that humans are susceptible to heuristics and biases when providing such quantitative assessments, and therefore in any elicitation, efforts must be made to minimise the influence of these biases. Three common heuristics that can lead to such biases are (1) judgement by representativeness: evaluating the probabilities of events based on how similar two things are, while ignoring base rates; (2) judgement by availability: basing the probabilities of events on how easily the events can be recalled; and (3) anchoring: the expert is given an irrelevant value for the probability of an event and then adjusts this up or down inadequately, based on how likely they think this event is. In particular, availability and representativeness can shift the elicited probabilities, and anchoring can result in quantiles that are too narrow, resulting in incorrect probability distributions. For more information on heuristics and biases, see Kahneman and Tversky (1971) and Slovic (1972).

When multiple experts give judgements in a study, it may be necessary, or at least preferable, to combine their judgements into a single coherent judgement to be reported back to the decision maker. There are two main ways to do this: behavioural aggregation and mathematical aggregation. In behavioural aggregation, the experts are typically brought together in a single place, with the objective of arriving at a consensus about each quantity in the analysis. When it is not possible to bring experts together, or in order to avoid the biases that result from freely interacting groups, methods such as the Delphi technique have been developed, which involve interactions among experts under the control of the analyst (Rowe & Wright, 1999). For further information on behavioural aggregation, see DeGroot (1974), Kerr and Tindale (2011), and for a recent discussion of behavioural versus mathematical approaches, see Bolger and Rowe (2015a). In mathematical aggregation, a mathematical rule is used to combine the judgements of the experts. There are two main ways of doing this: opinion pools and Bayesian aggregation. In opinion pools, a weight is given to each expert, and their judgements are combined linearly or logarithmically using these weights. The weights can be specified based on the performance of the experts on questions to which the analyst knows the answer, the judgement of the decision maker, self-weighting by the experts, or equal weights. In Bayesian aggregation, the expert judgements are regarded as data and are combined using Bayes' theorem.

The majority of the mathematical aggregation methods proposed in the literature assume that experts' judgements are independent, both of other judgements made by that expert and of judgements made by other experts (French, 2011). In practice, this may not be the case, as individual experts may be subject to the same biases consistently, different experts may be subject to the same biases, and different experts may have similar backgrounds and levels of experience. Thus, it seems likely that dependencies will exist within expert judgement studies, and therefore their impact on the model accuracy should be assessed.

There are several models in the literature that consider correlations, or whose modelling could include correlations, among multiple experts. Most of these are in the Bayesian aggregation literature and require the decision maker to specify the dependencies among experts and their biases. Examples include Jouini & Clemen, 1996; Lindley, Tversky, & Brown, 1979; Winkler, 1981; see also French (2011) for more examples. An alternative to the decision maker having to specify these values is to provide the decision maker with the empirical values of the correlations from seed questions to which the analyst knows the answer to but the experts do not.

In this paper, we consider data from 45 expert studies in which the judgements of multiple experts in various fields, from the nuclear sector to health and banking, were elicited. The studies were conducted by TU Delft and released as part of the TU Delft expert judgement database (Cooke & Goossens, 2007). A fuller description is provided in Section 4.2. For each study in the data set, we investigate whether within-expert and between-expert correlations are present for all of the seed variables in that study. This provides the most comprehensive analysis to date of the extent and type of the dependence found in expert judgement studies. Further, we also fit several of the most commonly used mathematical aggregation methods from the literature to each of these studies, and evaluate the accuracy of each using a number of metrics. This allows us to make some general comments about both the types of correlations that are present in expert judgement studies and also whether these correlations have a practically significant effect on the accuracy of the predictions that result from the models.

The remainder of the paper is structured as follows. In Section 2, we review the two main approaches to mathematical aggregation, namely Bayesian aggregation and opinion pooling. In Section 4.1, we consider the different possible sources of correlation within expert judgement studies and ways in which they could be measured, and in Section 4.2, we provide details of the TU Delft expert judgement database. In Section 5, we evaluate the dependence present in the TU Delft studies, and in Section 6, we fit a number of mathematical aggregation models to the case studies and evaluate the accuracy of each model for each study. We conclude the paper in Section 7 with a summary and discussion.

Please cite this article in press as: Wilson, K. J., An investigation of dependence in expert judgement studies with multiple experts. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.014