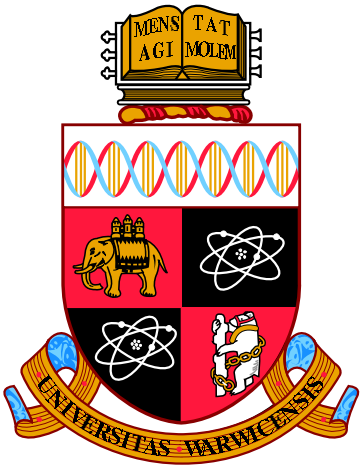


Bayesian hierarchical modelling for structured expert judgement

by

David Stephen Hartley

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

University of Warwick, Department of Statistics

September 2020

Contents

List of Tables

List of Figures

Acknowledgments

Undertaking a PhD is a path never traversed alone and this thesis is no exception. This research would not have been possible without the support and guidance of many people.

First and foremost, I am forever indebted to Professor Simon French for all of the support he has given to me as my supervisor throughout my studies. I have been Simon's student for seven years, a time period so long he retired shortly before my final submission. Although, I will give myself the benefit of the doubt and assume this is entirely coincidental.

Simon has been the consummate supervisor. All of the research conducted within this thesis has been substantially enhanced by his experience, guidance and wisdom. Simon and I spent many hours discussing my research, but also spent hours conversing about other topics of which we have shared interests. Simon's relaxed and collaborative style mean I always look forward to our meetings and I am incredibly grateful he has been so generous with his time.

I would like to thank my internal reviewers, Professor Vicky Henderson, Professor Saul Jacka and Dr David Rossell, each of whom has provided valuable feedback throughout my years of study. Professor Jim Smith, Dr Martine Barons and I often attended the same expert judgement events and I found the conversations we had over lunches and coffee breaks invaluable.

As a part-time student, remaining connected to my fellow PhD students has been critical for motivation and inspiration. I am extremely grateful to Dr Rodrigo Collazo, Dr Sophia Wright, Rachel Wilkerson and James Griffin who all made me feel welcome.

Finally, I must acknowledge the incredible support of my family throughout my research. I am very grateful to have parents who nurtured my passion for mathematics from an early age. My father, Dr Stephen Hartley, was an inspiration for me embarking on a PhD and it has often been my mother Anne's words of encouragement that have kept me plodding on when things have not progressed as expected.

Balancing a PhD alongside a full-time job is not a task that can be conducted without sacrifice or without a love for the work that you are doing. I started the seven years as a single man and finish as a married father. My wife Eliza has been amazingly supportive of the time, in the evenings and weekends, I have spent away from our growing family in order to do the things that I love and keep my research on track. This final thesis was only completed thanks to Eliza's willingness to shoulder the majority of a household burden that should have been equally shared. Completion of my research is a commitment to redress this balance, and I look forward to supporting Eliza in whatever seven year hobby she chooses to undertake now.

This work is dedicated to Eliza and my little daughter Annabel.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Parts of this thesis have been published by the author in the following published and pending articles:

- Hartley, D. and French, S. (2018). **"Elicitation and calibration: a Bayesian perspective."** *Elicitation*, Springer, Cham, 119–140. - Material from this publication is included in chapters 2,3 and 8.
- Hartley, D. and French, S. (2021). **"Bayesian modelling of dependence between experts: some comparisons with Cooke's Classical model."** *Expert Judgement in Risk and Decision Analysis*, Springer, Cham, *In Press* - Material from this publication is included in chapters 5 and 7.
- Hartley, D. and French, S. (tbc). **"A Bayesian method for calibration and aggregation of expert judgement."** *Journal of Approximate Reasoning*, *In Submission* - Material from this publication is included in chapters 5 and 7.

Further potential publications from the work in this thesis, include:

- **"BEAM - A (B)ayesian (E)xpert (A)ggregation (M)odelling tool"** *In Draft* - Utilising material from chapter 9.
- **"Integrating Bayesian modelling into the IDEA framework"** *In Draft* - Utilising material from chapter 8.

Abstract

Decision makers will often approach experts to help understand uncertainty when their problems cannot be analysed through empirical data alone. When formalised, this process is known as Structured Expert Judgement (SEJ).

Despite the fundamental premise of SEJ being about updating belief, which is the core of Bayesian statistics, SEJ studies often do not consider the Bayesian paradigm. Most SEJ studies utilise techniques which essentially take a pragmatic view of probability (e.g. Cooke's Classical model). Bayesian models have been proposed historically but are used rarely in practice.

This thesis outlines a Bayesian framework for SEJ. The research details the structure of an SEJ study and notes the benefits and limitations of traditional expert aggregation techniques. A collection of recently proposed Bayesian models are highlighted, before presenting a new model which aims to combine and enhance the best of these existing frameworks. In particular, clustering, calibrating and aggregating experts' judgements utilising a Supra-Bayesian parameter updating approach combined with either an agglomerative hierarchical clustering or an embedded Dirichlet process mixture model.

The new approach is assessed by analysing data from existing studies in a variety of domains including healthcare, climatology, volcanology and environmental management. These studies highlight significant overconfidence in expert assessments and consequently a wider range of uncertainty when considering the Bayesian approach. Cross-validation of over twenty studies demonstrates that the Bayesian approach generates higher statistical accuracy than performance weighting but at the cost of lost information.

Key process considerations when implementing a Bayesian model within a broader study facilitation protocol are outlined. A mechanism to embed the new Bayesian model into the popular IDEA protocol is proposed. A new tool, BEAM - (B)ayesian (E)xpert (A)ggregation (M)odel, to allow easy deployment of Bayesian thinking into IDEA is presented. Finally, some areas for further research are recommended.

Chapter 1

Introduction

> On an important decision one rarely has one hundred percent of the information needed for a good decision no matter how much one spends or how long one waits. And, if one waits too long, he has a different problem and has to start all over. This is the terrible dilemma of the hesitant decision maker.
>
> *Robert K. Greenleaf, The Servant as Leader*

1.1 Motivation

Over the past decade and a half I have worked in the healthcare and consumer-goods industries deploying analytical solutions. Terminology for the work that I have done has evolved over the years (statistics, analytics, data science, information/decision solutions), but the fundamentals have always been the same. All roles have been about supporting senior leaders to make data led and evidence based decisions.

I have worked almost exclusively in large multinational corporations with big budgets for data and analytics programmes. Even with significant resources at my disposal, and in very mature disciplines, I have seen consistently that the modelling components of analytical projects are typically the easiest element. Surprisingly, the much harder components are often getting the required data in the first place and then, post analysis, ultimately driving organisational change and adoption.

Countless times I have been required to perform analysis with sub-optimal data. Getting the necessary granularity and frequency of information can be extremely expensive. I managed data budgets for marketing analysis in the tens of millions of pounds per annum and still only managed to procure a fraction of the data I needed. Other times, data has simply not existed. Often organisational systems were not setup historically to capture required data and so critical knowledge was lost. At other times, data collection was unethical or unfeasible.

In many cases the lack of data led to other mechanisms being created in order to answer the desired problems. Typically, some form of expert judgement was used in lieu of data. In the worst cases this unstructured judgement represented simply the intuition of the decision makers, even when this was in direct contradiction to the little data available. Data limitations became a rationale to discredit evidence based decision making altogether.

Conversely, even in times of ubiquitous data, I have been dismayed to find that sound advice has been ignored in favour of "gut-feel". Understanding decision makers' prior beliefs became a key passion. I felt that personal and organisational effort was wasted in times where decision makers held beliefs so strongly that any data presented was not going to impact the final decision. Even when beliefs were not so strongly held, at times implicit decision rules also meant, regardless of the analysis, course of action was predetermined. Formally eliciting decision rules has thus become a standard component in my project planning.

These two competing dimensions led to my engagement in structured expert judgement. In addition to the obvious benefits of improving decision quality with better inputs, I see this discipline creating a bridge between technical analysis and cultural engagement. Formally eliciting experts' opinions, particularly in continuous enterprise settings, can significantly increase engagement as people feel heard. Individuals also feel accountable for project outcomes which can have major impact to adoption.

The link between adoption and decision makers' prior beliefs also led to a specific interest in Bayesian approaches within structured expert judgement. I became focussed on robust methods which allowed these priors to be formally considered.

Early research showed that expert judgement had historically been dominated by non-Bayesian methods including the Classical model and there was plenty of scope for future research. Bayesian models had challenged the status quo, but for a variety of reasons had not been broadly adopted. In order for a Bayesian model to challenge this paradigm, it needs to be simple enough for a decision maker to use in practice whilst remaining robust. Four core areas which I believed require

particular attention for a model to be considered practical include:

- **Robustness** – There cannot be too many constraints on the model that must be adhered to in order for the model to produce meaningful results, i.e. many restrictions on the number of experts, extreme sensitivity to prior distributions etc.
- **Application to multiple contexts** – The model must be applicable to different decision making structures and organisational paradigms. The framework should be useable in both commercial and policy decision making.
- **Data Minimisation** – The model must minimise the amount of data required from experts. Each data point requested requires effort from the experts and therefore time and cost for study organisers.
- **Documentable Process** – The expert judgement framework, including protocols for eliciting and modelling, must have a well documented process to allow decision makers and analysts to apply the methodology rigorously and effectively.

No existing Bayesian model appeared to deliver against these goals.

This thesis, and the work considered within, has been conducted to address the above. I aim to use its findings to deliver on my passion for improving outcomes by ensuring a people centric approach to data based decision making.

1.2 Objectives

There are six key objectives that this work will aim to deliver:

- **Objective 1.** Review the literature on expert judgement, in particular Cooke's model, and highlight areas whereby a Bayesian model could add further value.
- **Objective 2.** Build a new Bayesian model for mathematically calibrating and aggregating experts' judgements.
- **Objective 3.** Provide evidence that Bayesian (re)calibration can enhance decision makers' perception of uncertainty.
- **Objective 4.** Apply the new Bayesian model to a comprehensive data set to demonstrate that it can behave comparably to current best in class mathematical aggregation approaches.

- **Objective 5.** Provide a critique of the current protocols and procedures for judgement elicitation from a Bayesian perspective and demonstrate how the designed model can be integrated into these protocols.
- **Objective 6.** Create a tool, aligned to the protocol design, to support in the application of a Bayesian model to an expert judgement study.

1.3 Thesis outline

This thesis has been structured in accordance with the named objectives.

Chapters 2, 3 and 4 formally define structured expert judgement before describing different approaches for combining judgements from multiple experts. Cooke's Classical method is described in detail as it is an exemplar in this space and is used in subsequent sections as a comparator. Methods for validating the output of expert judgement methodologies are presented. Material within these chapters delivers on *Objective 1*.

Chapter 5 outlines my Bayesian model. The hierarchical Supra-Bayesian methodology outlined builds on the work of primarily two existing Bayesian approaches, Clemen and Lichtendahl [2002], Albert et al. [2012], but then extends this further with a unique homogeneity group clustering approach. This chapter delivers against *Objective 2*.

Chapter 6 applies a cross-validation of the calibration component of this model, in isolation, on a comprehensive dataset from Roger Cooke. This demonstrates increased accuracy over raw elicited judgements, delivering against *Objective 3*.

Chapter 7 demonstrates the output of my full Bayesian approach on a number of historical studies before conducting a cross-validation exercise to highlight its efficacy relative to Cooke's Classical model, meeting *Objective 4*.

Chapter 8 highlights how this Bayesian framework can be practically deployed by integrating it into the IDEA protocol: a leading set of elicitation procedures. Chapter 9 provides visibility to the BEAM tool, which I have built to embed the model within IDEA. These chapters deliver on *Objective 5* and *Objective 6*.

Finally, the thesis concludes by discussing the findings and the impact thereof before highlighting some recommended areas for further research. To aid the reader a glossary of commonly used abbreviations and mathematical symbols is included after the Bibliography.

Chapter 2

Introduction to structured expert judgement

2.1 Definition of structured expert judgement

In deciding a course of action decision makers (DMs) are required to make judgements in the face of uncertainty. In a surprisingly large number of cases it is not feasible to infer the bounds of this uncertainty from comprehensive empirical data and modelling. Problems DMs face are often too complex to model holistically e.g. impact of global warming. Or, do not have sufficient data available in order to model robustly.

Data may not be available to a DM for many reasons. It is often *too expensive* to collect at the granularity required. For a marketing company to get visibility to data on the sales of its products through consumer retail channels at the granularity required to perform price elasticity modelling can cost millions of pounds. At other times it is *unethical to collect* requisite data as the collection process itself would endanger someone. In deciding a course of action from the threat of a future nuclear meltdown, a decision maker may wish to understand effects of different radiation levels on health in ways which have not been documented before. Exposing individuals in order to gain this understanding would be immoral and so decisions need to be made with incomplete data. In other cases events may be so *rare*, that statistically meaningful samples of data do not exist.

Human judgement thus needs to be made as to the likelihoods of different scenarios. Typically, a DM will not hold all the necessary information personally and will consult an expert to provide a viewpoint on the probability of potential outcomes. Structured expert judgement, SEJ, then defines:

- The formal processes a DM would utilise to frame the problem appropriately and then extract the desired knowledge from an expert (formally known as *elicitation*).
- the methods they should employ to update their belief given this additional information.

Should more than one expert be consulted, a key component of the methods will pertain to how the judgements from multiple experts should be considered relative to one another and, if necessary, combined together.

One of the major assumptions regarding SEJ studies is that experts are giving an opinion solely on probability of potential outcomes; they are not offering value judgements on the decision (French [2011]). Experts are also only being utilised for their knowledge, they are not being consulted for their skills. Here, skills are determined to be the ability to execute specific tasks efficiently and effectively as a result of significant training, repetition and feedback (Burgman [2016]). Experts can also provide valuable insight to DM's in these regards, and indeed in many other ways, however, considerations surrounding this sort of expert consultation sit outside of SEJ. SEJ is solely focussed on the use of experts to understand uncertainty. One of the key areas that SEJ is consequently deployed is risk analysis.

The reader will note, that to this point, no assertion has been made relative to the type of variable for which the expert is consulted. SEJ outlines methods to help DM's utilise experts for both single/multiple events or for discrete and continuous variables. Expert judgements can be utilised for observable variables or to gain perspective on latent variables in bigger models. Indeed the viewpoint from an expert may come in the form of a point estimate or as a probability distribution over potential future states of nature.

For the purpose of clarity within this thesis it is helpful to outline a more formal and mathematical definition of SEJ, including some common nomenclature that we shall use throughout. To aid the reader, a glossary of terms is included, preceding the Appendix.

Statistical decision theory outlines how decision makers are regularly required to make judgements in the face of uncertainty. Suppose a decision maker (DM) needed to assess the consequences of their decision based on the outcome of a set of impactful variables, $\mathbf{X}$, known as *target variables*. Let us assume he or she starts by specifically considering the expected outcome of a single random variable $X \in \mathbf{X}$. Furthermore, let us assume that observations of X, denoted x_X, follow an unknown probability distribution $p(x_X)$. In many cases the DM may not have

the data to assess $p(x_X)$ robustly and so must reach out to experts to give their assessment of their uncertainty. Let $\mathbf{E}$ define the set of experts, $|\mathbf{E}|$ the number of experts and $e \in \mathbf{E}$ an individual expert. Let us assume the DM held the belief, $\pi_{DM}(x_X)$, prior to talking to the experts and the experts' beliefs are given by $p_E(x_X) = \{p_e(x_X) : e \in \mathbf{E}\}$. The goal of a structured expert judgement model will thus be to build a distribution for the DM's perspective of the uncertainty given the experts' statements, $p_{DM}(x_X)$. We are looking for a function ϕ such that $p_{DM}(x_X) \propto \phi(p_E(x_X), \pi_{DM}(x_X))$ (Genest and Zidek [1986]).

This definition of SEJ assumes that there is a single decision maker. Such problems are termed '*expert problems*' as outlined in French [1985] and French [2011]. In reality the structure of the problem and decision making body may mean that the situation is significantly more complicated. We will outline later the impacts of other decision contexts. Regardless of the context, SEJ defines the set of robust processes employed to support the elicitation and subsequent use of judgement from experts in lieu of available empirical data.

There is an important decision to be made about whether experts are making an assertion regarding the uncertainty of an input variable to a decision/model, (i.e. helping with the construction of an informative prior for a larger more complex model) or to the resulting output (supporting the development of the DM's full posterior for the model). This has implications regarding whether experts should provide judgements on parameters or only upon observables. For now we will leave aside this question as in many cases this subtlety does not impact the final result and therefore, unless explicitly stated, we will assume it does not matter.

2.2 Why expert opinions differ

In many decisions a DM may go to more than one expert to elicit judgements on uncertainty surrounding key variables of interest. Indeed, probability knowledge emerging from a group will often be higher quality than from an individual (Burgman [2016], Galton [1907], Gordon [1924], Surowiecki [2005], O'Leary [1998]). Eliciting different viewpoints is thus often recommended. Perhaps surprisingly, consulting multiple experts on the same topic can lead to substantially different evaluations of uncertainty. In the cases of harmful effects of low level radiation and potential harm from water fluoridation (Mazur [1973]), differences in viewpoints between well respected and rational scientists would lead to radically different policy decisions. It appears contradictory to have substantially varying judgements arise from experts (assuming the same unbiased elicitation process) and may call the validity of SEJ

into question if not well addressed. It is important therefore to understand the drivers of these discrepancies.

There are many potential reasons experts' opinions can differ (Camerer and Johnson [1997], Shanteau [1992]). Mumpower and Stewart [1996], building on the work of Kenneth Hammond, identified three common explanations potentially driving discrepancies:

- *Incompetence* – this assesses the validity of an expert to be making judgement on the basis of qualification, experience or intelligence.
- *Venality* – this outlines that differing personal interest in the output of a process can drive differences in assessment. *i.e. an expert's opinion on a stock they hold is likely to be biased by their own position.*
- *Ideology* – Political, religious or ethical values or beliefs can impact the position experts are willing and able to take on policy decisions.

The authors argue, in addition to the above rationale, the process of human analysis can itself lead to disagreement and often it is possible to identify reasons for disparity without resorting to one of the above. With a specific focus on scientific and technical policy decisions, Mumpower and Stewart go on to claim it is possible to cluster additional reasons for disagreement into a number of groups:

Different problem definitions *i.e. the specific question the expert perceives they have been asked to comment on.* This can manifest itself as:

1) Fact value confusion
2) Different environmental criteria

A key driver of different problem definitions can often be linguistic uncertainty including ambiguity, vagueness, context dependence and under-specificity. This can also be driven by poor decision/problem framing.

Different information *i.e. the choice of key variables (and the values these take) that are critical to make a judgement against the perceived problem.* Often perceived as:

3) Differential availability and use of information

Different organising principles *i.e. the data modelling and analysis approaches used to produce the relevant distribution given the problem definition and the information available.* This breaks down into:

4) Qualitatively different mental models
5) Different modes of cognition
6) Different organising and integrative models.

Please see Mumpower and Stewart [1996] for further details.

Whilst these are the main structural reasons that may lead experts to have substantive differences in assessments, there are also a number of psychological phenomena related to heuristics and biases which people are subject to. The seminal work in this area, Kahneman and Tversky [1979], Kahneman et al. [1982], highlighted that elicitation itself can introduce biases based on the way that information is shared and questions are posed. This topic will be discussed further in Section 2.5 which will cover formal elicitation processes. For now, it suffices to highlight that discrepancies can emerge within experts' judgements for both justifiable and unjustifiable reasons at almost any point within the conception, cognition and extraction processes.

There are a number of reasons experts may perceive both the problem and resultant probability distributions differently. Thus, a critical challenge for any SEJ process is to assess which drivers of discrepancy between judgements are legitimate and which are those that need to be removed/minimised. This minimisation can take place through the choice of experts within the study, the method of elicitation used or through the subsequent analysis of the outputs.

2.3 The context of SEJ

The impact of disagreement among experts can vary depending on the decision making structure used (French [2011], French [1985]). Three such contexts are:

- *The expert problem* – In this situation a group of experts are asked for opinions by a single DM who has a specific decision problem. Here the DM is considered separate to the group of experts and holds sole accountability for the decision.
- *The group decision problem* – In the group decision problem it is the expert group itself responsible for the decision and the resulting consequences. For this reason, the group will not only wish to be able to demonstrate the decision is rational but also potentially there was a fair and democratic method for aggregating opinions across the experts involved.
- *The textbook problem* – There is no defined decision to be made, the expert group may simply be required to give judgements for future analysis and decision making for an as yet undefined problem.

In most practical applications there is not a single decision, more a chain of decisions, thus the process is often a composite of more than one of these contexts (French [2011]). Each context brings with it a unique set of challenges and appropriate actions for the decision maker(s). Taking the suitability of recalibration as an

example. In the *expert problem*, a DM rightly would be very keen to understand the accuracy of the experts she/he has enlisted. It would then be reasonable for her/him to adjust the relative credence she/he applies to each expert's judgements accordingly. In the *group decision problem* however, it would be inappropriate to calibrate in this way. If each member of the group were asked to assess the accuracy of others they may use different models for this, they would also have a biased opinion on their own performance. Experts are also empirically found to be very poor judges of one another's ability to estimate (Burgman et al. [2011b]). Performing recalibration adjustments would appear undemocratic. Here as French [2011] remarks, the appearance of complete democracy is an illusion. Indeed, as given by Arrow's impossibility theorem and related results, the sought after "rational democracy" is an impossibility.

The *textbook problem* comes with a number of unique challenges; mostly stemming from the lack of a defined problem. This can occur because an expert is just asked to give an opinion on the probabilities of a number of variables without knowing the context this will be used, or as a result of a previously published piece of work being used as an input into a completely separate problem. The latter here is more common, although as French [2011] points out, with the increase in the public availability of data and the broader nature of conversations regarding policy, the former is likely to be more and more important. One of the particular challenges of the textbook problem occurs when decision mechanisms are complex, i.e. multi stage decisions with multiple contexts. Here, by definition, the inherently fixed nature of documentation vs. the potential flexibility of a "live" expert can create further issues.

Whilst most SEJ studies tackle the *expert problem* and significant work has been undertaken to consider the democratic nature of the *group decision* problem, little work to date has investigated the *textbook problem*. For an expert judgement method to be truly practical, it should be as context agnostic as possible and thus, given the additional complexities the *textbook problem* puts on analysis, there is significant scope for research in this space. Throughout this thesis the *textbook problem* shall be used as a lens through which to assess various expert judgement models in addition to the context they were originally designed.

2.4 Practical procedures for elicitation - the EFSA model

Given a specific context, a set of processes that surround any given SEJ study must be selected. Of the major texts covering the processes and procedures surrounding

structured expert judgement, the European Food Safety Authority's (EFSA [2014]) guidance is possibly the most complete with regards to expressing the roles of the different players involved. Many of the topics and models are also outlined in other texts such as Meyer [2001], Cooke and Goossens [2000], Hemming et al. [2018], Dias et al. [2018] and O'Hagan [2006]; but we shall take EFSA [2014] as an exemplar to introduce the topic. Further work in Chapter 8 will assess methods for embedding a Bayesian model into one of the other major protocols, the IDEA framework (Hemming et al. [2018]).

It is important at this stage to note the role of EFSA[1], the European agency. EFSA operates independently of both the European legislative, executive institutions and EU Member States, and is responsible for risk assessment in the area of European food safety. This is completely separate from risk management or policy making, and was legally established under the General Food Law – Regulation 178/2002. EFSA plays an important role in collecting and analysing data to ensure that risk assessment is supported by scientific information, including expert judgement, and then appropriately communicated to both stakeholders, such as policy makers, and the public at large. Acting in this way, EFSA will be most interested in evidence and analysis for societal decisions, impacting the approaches and contexts in which EFSA operates SEJ. EFSA is a regulator and so deals with expert problems or, occasionally, textbook problems. In these contexts, EFSA would seem to be developing probability distributions that represent the views of a *rational scientist*.

The EFSA SEJ process starts with the formation of a Working Group. The Working Group comprises individuals accountable for the overall program of work. They are tasked with problem definition and the development of a risk assessment model. In undertaking these they will identify when limited evidence is available for some of the critical variables, deciding that there is a requirement to consult experts to fill such knowledge gaps.

At this stage the Working Group will typically hand the program over to a second group, the Steering Group. The role of the Steering Group, *inter alia*, is to refine the parameters to be elicited and to identify the precise expert knowledge that is needed. Once these elements have been finalised, it is critical to select the experts themselves, which may have implications for the elicitation method used and thus the final aggregation model. The final decision on each of these elements lies with the Steering Group.

In practice, the selection of the experts is not a trivial matter and can be impacted by a number of variables quite outside considerations of their expertise on

[1]https://www.efsa.europa.eu/en/aboutefsa

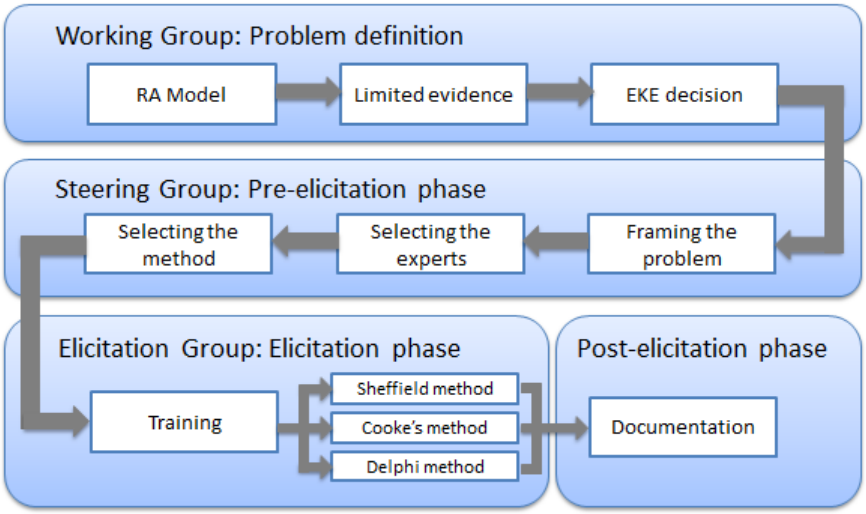


Figure 2.1: The EFSA expert knowledge elicitation process, reproduced from EFSA [2014].

the parameters. Availability is obviously a critical factor; and for EFSA there are potentially political constraints factored into the decision. There may, for example, be quotas on attendance from EU member-states or other issues of representation. This constraint may impact the analysis as it potentially introduces a further risk of expert bias which may need to be controlled as part of the elicitation/modelling process. This issue is not just limited to EFSA nor to similar contexts, it is a common phenomenon in SEJ that experts may be assigned rather than selected. From a Bayesian perspective, were this assignation to happen, having a clear understanding of these affiliations may be critical for understanding dependence and the ultimate reduction of bias within the model.

In selecting the model, the Steering Group can choose from three approaches ratified by EFSA, these approaches are the SHEFFIELD method (O'Hagan [2006]), Cooke's method (Cooke [1991]) or a version of the Delphi method (Dalkey and Helmer [1963]). Each of these versions have different requirements and the ultimate selection of the model will depend on factors such as geographical split of the experts, diversity of backgrounds, or simply time or skill requirements.

Following the selection of the elicitation method and the model to be used, the Steering Group will hand over to an Elicitation Group. The Elicitation Group will typically be more familiar with facilitation and elicitation, and will be accountable for training the experts to help them understand the process, the requirements necessary and to also to help them be aware of their own biases and how to mitigate

these. Following this the Elicitation Group will perform the elicitation and any subsequent modelling necessary. The information from here is then either handed back to the Steering Group or directly translated into a set of Post Elicitation documentation. Please see Fig. 2.1 for a simple visual of the process.

In a typical EFSA study, there are a wide variety of individuals involved throughout the process. It is imperative to ensure that common understanding of context and models flow through the groups. Documentation is a critical component to this. In other SEJ contexts, it is feasible that the process is much leaner and the Working Group and the Steering Group are compressed. Or in certain decisions, it is feasible that a decision maker will come directly to an analyst for insight that will ultimately lead the analyst to play the role of all three of the groups outlined here. In these circumstances the common understanding is much simpler to achieve, however the training and documentation requirements may be harder to implement with limited resource.

2.5 Eliciting expert opinions

People do not naturally think in subjective probability distributions and as such, it is often necessary for the DM to work through a process to help an expert formalise their personal beliefs and to share these in a coherent way. Before embarking on this, it can be helpful for both the DM and the expert to understand a little of the psychology of judgement. The seminal work on the psychology of decision making, Kahneman et al. [1982] demonstrated a number of heuristics (representativeness, availability, anchor adjustment) that can drive judgement bias, and the authors also developed the much lauded prospect theory (Kahneman and Tversky [1979]). Building methods for mitigating these biases in expert elicitation can lead to a more refined assessment of probabilities. In recent years Kahneman furthered the conversation and bought this topic into the mainstream by adopting the concepts of System 1 and System 2 thinking in his popular book "Thinking Fast and Slow" (Kahneman [2011]). Methods the DM can employ to ensure he/she does not bias the expert include (French et al. [2009]):

- Be very specific with the expert on defining the problem statement and thus make sure any ambiguity the expert may have is removed. This reduces the chance of the representativeness bias adjusting his beliefs.
- Asking experts to explicitly consider alternative conceptual models of the problem they are addressing, counterfactuals and extreme tail probabilities.
- Never prompting the expert with explicit values as this may anchor the experts

judgements.

- Asking the expert to explain qualitatively the theory of the systems and phenomena involved and his/her understanding of how they relate. This encourages the expert to think through the problem holistically and potentially counter the availability bias.
- Training experts themselves on the key biases to look out for within the process and themselves. For example, highlighting the well documented issue of probability weighting.
- Considering multiple elicitation methods to ensure that individual expert responses are consistent when elicited in multiple ways and to challenge where they are not.
- Ensuring an appropriate decision framing to minimise the influence that language and question structure has on the experts' responses. This will reduce the framing effect.

Once this has been established, there are several methods for eliciting the actual quantitative probabilities. Many of the techniques used to draw out these subjective probabilities are identical to those used in the elicitation of utility functions and the following examples draw on literature from both.

The most common form of elicitation is simply to ask experts their perspectives post a training exercise. However, common examples of other conceptual devices used to help structure expert thinking are the betting wheel or the lottery (Smith [2010]). In the example of the betting wheel (the lottery is similar), the DM is required to interview the expert and ask them to decide whether they believe their subjective probability of an event is more or less likely than the probability represented by the wheel. The DM can then recursively iterate through refinements of this value, by making slight adjustments to the wheel, before finalising on an agreed number. As outlined before, the DM needs to be very careful he/she does not bias the expert with the initial choice. In this well-studied area, a number of techniques have arisen, Table.2.1 outlines some of the most common.

Typically, eliciting over a continuous distribution is technically harder than over a finite set of events. A number of the techniques above do support this, however, are generally harder to realise. Usually, by either choosing one from one of the conjugate families, (or a combination there of) or by eliciting a finite subset and then extending this to the whole event space.

In practice it can be difficult for experts to think in terms of distributions and therefore it is often wise to simplify the problem by not directly eliciting the distribution or parameters involved. Instead we can extract each expert's perspec-

Direct Methods
– Magnitude or Direct Estimation
– Ratio or Odds Estimation
– Equating Sense Distances
– Graphical
– Probability Density Function Estimation
– Hypothetical Future Samples
– Equivalent Prior Sample
– Distribution Parameter Estimation
Infer Methods
– Betting
Hybrid Methods
– Lottery
– Bid or Bet Choice

Table 2.1: Numeric elicitation methods. Reproduced from Chesley [1975].

tive on some intuitive points within the broader distribution and then construct a function g_e which represents our best approximation to the expert's beliefs given the elicited data. Cooke [1991] outlines the benefits of eliciting in this way, given the challenges experts have in mentally formulating parametric distributions. Typically we would elicit three quantiles[2] from each expert $e \in \mathbf{E}$, L_e, M_e, U_e, associated with three probabilities P_L, P_M, P_U (often the 0.05, 0.50 and 0.95 quantiles) and the full distribution for the expert is approximated by $g_e(\cdot|L_e, M_e, U_e)$. In certain studies, five quantiles may be elicited (often these represent the 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles). In this case, the full distribution g_e is thus conditional accordingly. g_e will often be from a parametric family and as such we encode our model further by utilising L_e, M_e, U_e to infer the parameters γ_e of g_e, that represents the expert e's conceptual model of X. Thus, g_e and γ_e should be chosen in such a way that it closely approximates the expert's beliefs at the elicited quantiles (i.e. $g_e(x_X|\gamma_e) \equiv p_e(x_X) + \epsilon$, with minimal error term ϵ.). This process should not be done in isolation from the experts and a feedback process is often employed to playback g_e and give an opportunity for refinement. Other methods exist for obtaining g_e more directly from experts and many authors have considered the best elicitation methods for expert judgement models (Aspinall [2006], Chesley [1975], Gosling et al. [2007], Gosling et al. [2012], Garthwaite et al. [2005]). The decision maker's prior will often be elicited and parameterised in a similar way.

[2]The elicited quantiles should be equivalent to the expert's true beliefs +/- any elicitation error Smith [2010].

The manner in which questions are posed and the subsequent impact on the type of data elicited can also be a critical tool in ensuring that expert's underlying beliefs are accurately reflected. Natural frequencies, as highlighted in Gigerenzer [2011], are an alternative manner to conditional probabilities of describing the joint frequency of multiple events. Natural frequencies consider breaking down large samples in easily interpretable ways. For example, the conditional probabilities denoted by the statements: the probability that a woman has breast cancer is 1% and if a woman has breast cancer the probability that she tests positive is 90%, can be reimagined in natural frequencies as 10 in 1000 women have breast cancer and of these 10 women, nine will test positive. Formatting the information in this way makes the data more digestible but it also has empirically been shown to improve the ability of individuals to conduct inference. Most notably when considering Bayes rule and base rate neglect in medical settings, (Gigerenzer and Edwards [2003]). It does not matter mathematically how this information is presented, logically they are the same, but to a human it has a significant impact and therefore the format of elicitation is critical.

Visual formats for eliciting subjective probability distributions can be particularly powerful when the subject is remote, a study by Delavande and Rohwedder [2008], conducted to assess a ball and bin method for assessing Social Security estimations, showed this generated more usable results and higher probability in the central bin, not generated by an anchoring effect, than the more traditional percent chance method and in turn highlighted a potential bias in the traditional model. Ball and bin methods ask respondents to distribute a fixed number of balls across a fixed number of bins in line with their beliefs.

Further work, specifically on the biases involved when considering heuristics and biases in eliciting expert judgement have led to numerous recommendations for methods to optimise the process before during and after elicitation has occurred, (Kynn [2008]).

Chapter 3

Combining experts' judgements

3.1 Introduction to judgement combination

In multi-expert problems, once the problem is defined, the decision making context is decided and all of the experts are chosen, it is then necessary to combine the judgements of the group in a reasonable way. There are many different models for doing this.

Let us, for a moment, retract the assumption in the first section that the target of the probability distribution is not relevant and focus on the case that the distribution is for an input variable to be passed into a model through which the uncertainty flows, the output of which is a further distribution. In this case, either, each expert can run the model against their personal subjective distribution and the resultant output distributions can be reviewed and aggregated or the distributions can be combined into a single distribution before parsing into the model. In both cases the aggregation can be performed either mathematically or behaviourally. In the first case aggregation may not be necessary at all; the output distributions could be presented to the DM simply as a range of outputs. Please note, unless the structure of the model is linear, the outputs of the two approaches are likely to be different.

Behavioural techniques for aggregating data involve gathering experts together in order for them to discuss and to form a consensus on the distributions to be used. This can be done in either an informal or a more structured manner, such as the SHEFFIELD method (O'Hagan [2006]).

A simple mathematical approach which recognises that the DM may explicitly rely solely on the experts, is for the DM to simply average over the experts' beliefs, i.e. $p_{DM} = \phi(p_E) = \sum_{e \in \mathbf{E}} (1/|\mathbf{E}|) p_e$. Here DM prior beliefs are either ignored or

added into the sum by stating that the DM is an additional expert in **E**. This method can then be easily extended to allow the DM to vary the effect of each expert on the resulting distribution.

$$p_{DM} = \sum_{e \in \mathbf{E}} \omega_e p_e \tag{3.1}$$

Where ω_e is a weighting factor for expert e. These methods are termed linear pooling methods (Hammitt and Zhang [2013], Genest and McConway [1990], Clemen and Winkler [1999]). Different pooling models then propose different ways to determine the relative weightings ω_e. The key linear methods used are (Hammitt and Zhang, 2013):

- *Cooke's Classical model* – Here the weights depend on the quality of experts' judgements on variables for which the DM knows true realisations *a priori*.
- *Simple Average* – each $\omega_e = 1/|\mathbf{E}|$, i.e. this is a straight arthmetic average.
- *Best-expert Method* – In this case, the expert with the highest value in a performance evaluation is given weight 1 and all others are given weight zero.

Whilst linear methods are more common, it is also possible to combine judgements with a weighted geometric mean or something more general all together (French, 2011), here:

$$p_{\text{DM}} = \prod_{e \in \mathbf{E}} \omega_e p_e \quad \text{or} \quad p_{\text{DM}} = \phi(p_1, p_2, ..., p_{|E|}) \tag{3.2}$$

A Bayesian expert model flowing directly from Bayes theorem assumes that the posterior the decision maker has on x_X, conditional on the experts' beliefs, $p_{DM}(x_X|p_E)$ is proportional to the likelihood they ascribe to hearing the experts' elicited values of p_E given x_X, $p_{DM}(p_E|x_X)$, multiplied by the prior belief the DM has, π_{DM}, i.e.

$$p_{DM}(x_X|p_E) \propto p(p_E|x_X) * \pi_{DM}(x_X) \tag{3.3}$$

Thus the DM treats the elicited information as data. The Bayesian methods are much more complex in nature than the opinion pool methods and can often be impractical despite being very strong in theory. One of the key issues which can make the Bayesian method particularly complex is the potential for correlation between experts and DM's judgements based on shared experiences (Booker and Meyer [1988]). French [1980], first highlighted this issue but since this point little research has been done to identify methods for counteracting this. Cooke highlighted in "Experts in Uncertainty" (Cooke [1991]) that Bayesian models can outperform

the Classical model in certain circumstances but are often complex to perform in practice.

By retracting the assumption regarding the target of the probability distribution, we have assumed the expert group have a single model through which to pass their probability distributions. Here the issue of different organising principles outlined earlier is reduced. However, in each of the three decision contexts discussed before, it is feasible that output probability distributions are desired rather than the formulation of priors. This adds a further layer of complexity to the problem. In the case of the textbook problem this can be particularly troublesome as the expert is often not there to defend the decision and highlight some of the underlying assumptions in the models used.

There are those who would argue this form of aggregation should not occur at all (Kaplan [2000]). Building on the work of Clemen and Winkler, Kaplan suggests that we bypass the issue of combining probability distributions by instead getting experts to share the information they have relevant to the issue at hand and then having the group of experts discuss, align and agree on a "single body of evidence" which is then passed through Bayes theorem item by item to arrive at a posterior probability curve. This is arguably a mixture of a behavioural and algorithmic method for aggregation. A potential significant issue with this approach is that it relies on experts aligning on a single perspective of what will go into the process and the salient data underpinning this. All of the data is given the same credence within the model. There maybe entirely justifiable differences in belief between experts on these elements. Getting alignment on all the inputs could be even more challenging than simply trying to elicit a consensus distribution for the target variable utilising a purely behavioural method.

3.2 Overview of behavioural approaches to SEJ

The focus of this thesis is on a new mathematical approach to expert judgement aggregation. To this extent we do not review behavioural approaches in detail. It is however important to understand some of the key concepts which underline these approaches to see how and where mathematical approaches have had issues in the past. In Chapter 8, we will also outline how some of the approaches taken by behavioural methods can be integrated with our mathematical model to build more holistic processes.

Behavioural methods of expert elicitation utilise a variety of techniques. They are, however characterised by a similar flow. A facilitated group discussion is

conducted to ensure common understanding of the issues and data available. Experts' opinions are then elicited either independently or within the group setting. Following which, the group debates the rationale for the diversity of opinion demonstrated by the elicited values. Through discussion of these differences, typically, experts' opinions evolve, and their position adjusts. The process concludes when the group reaches a consensus on what the final judgement should be for a given target variable. Some methods rely on multiple rounds of elicitation and debate to drive to this consensus. Effective facilitation is an extremely important component of behavioural methods in order to minimise potential biases and ensure a consensus is reached. Whilst there is the potential for groups never to converge to a single perspective (if no expert is willing to adjust their position), this is not common, and most methods have a process by which they deal with such cases.

The key philosophical difference between behavioural methods and mathematical methods lies in the aggregation process and therefore, ultimately what the final elicited figure represents. The original group debate regarding the issues and data is common to both approaches (or at least it can be, not all mathematical approaches require this). Convergence to a single consensus in behavioural methods means that the final number represents the expert groups' perspective. The expert group as an entity is ultimately owning the value. Whilst there are some similarities with the group problem outlined by French [1985], this is not implying that the expert group is the ultimate decision maker. Behavioural techniques can be used in the expert problem where there is still a single DM. Although in these cases, it is recommended the DM is not part of the expert group itself, as this can lead to a hybrid expert/group problem with some inherent bias.

Within mathematical techniques the aggregation process is conducted by applying some agreed algorithm to combine the judgements elicited from the individual experts. Group consensus over the result is not mandated (although often the results are played back to the group for review). In this context the expert group does not exist practically as an entity and the final elicited figure is owned by the individual (often the DM) who decided on the algorithmic method utilised. In both behavioural and mathematical techniques, how the DM's own perspectives are played into the elicitation is an important and nuanced topic.

Three major behavioural expert elicitation techniques are: the DELPHI method, decision conferencing and the SHEFFIELD method (also known as SHELF). Initially developed by RAND in the 1950s, the DELPHI method was the first popular expert elicitation process (Helmer-Hirschberg [1967], Dalkey and Helmer [1963]). Indeed, by the time that Harold Linstone and Murray Turroff wrote The DELPHI Method in

1974, they noted that:

'In 1969 the number of DELPHI studies that had been done could be counted in three digits; today, in 1974, the figure may have already reached four digits' (Linstone and Turoff [1975]).

In 2020, this figure is at least 1 if not 2 orders of magnitude bigger again. The DELPHI method, often characterised by multiple elicitation rounds, has become a staple judgement elicitation process. It has been used in many fields and has evolved in many directions (Rowe and Wright [2001], Taylor [1989], Skulmoski et al. [2007], Hsu et al. [2010], Steurer [2011]). As a result of this DELPHI can now be thought of as a collection of behavioural (and in some case mixed behavioural/mathematical) techniques, (Wang et al. [2012], Gordon and Pease [2006], Hsu et al. [2010]).

Decision conferencing was developed in the late 70s and early 80s, initially by Cam Peterson and then built upon considerably by Larry Phillips (Phillips [1991], Parnell et al. [2013]). The focus of the decision conference is to provide a structure to the behavioural elicitation process. The model is not used to try and discover a 'right' answer, it is there to serve as a 'tool to enable thinking'. The ethos of the decision conference is that by exploring issues as a group, creative thinking is enhanced and thus improvements to the decision model (and judgements held by the parties involved) emerge. Convergence of opinions then lead to a common sense of commitment to action. The decision conference is a holistic process, not just focussed on the aggregation component of the expert's beliefs. Often the most significant value of the decision conference can be driven by the process of arriving at common understandings of the issues and the decision model to utilise, (French [1993], Morton et. al [2011]). The role and importance of independent facilitation evolved considerably within the decision conferencing literature. The structure of a decision conference workflow is outlined in Fig. 3.1.

The SHELF method is a relatively more recent expert judgement process (O'Hagan [2019]). Developed by Tony O'Hagan and Jeremy Oakley in the first decade of the twenty first century (O'Hagan [2006]). The SHELF method is characterised by a clear delineation between initial elicitation from experts and then a subsequent group discussion to find the consensus. The consensus aimed for by the SHELF method is that of an independent impartial observer (Gosling [2018]). There are a comprehensive set of processes[1] and tools[2] to support in the elicitation and aggregation of the experts' opinions within the SHELF method. SHELF also has an R package and online training readily available which makes the method broadly

[1] http://www.tonyohagan.co.uk/shelf/

[2] http://optics.eee.nottingham.ac.uk/match/uncertainty.php

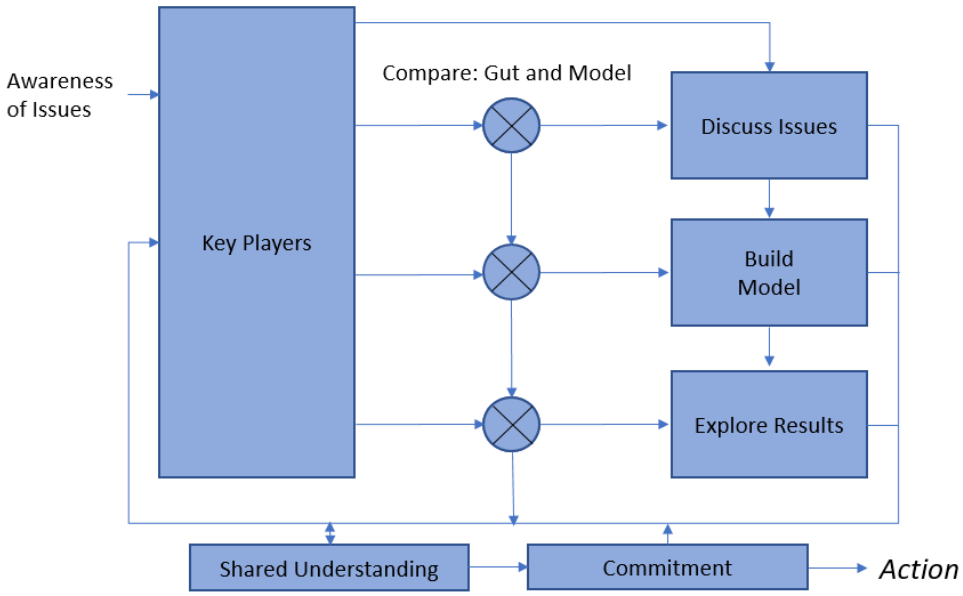


Figure 3.1: Overview of the decision conferencing workflow. Replicated from Phillips [1991].

accessible (Morris et. al [2014]). Model choice can often be very context dependent and the SHELF method, or a variant thereof, is often used for prior elicitation within Bayesian models for the pharmaceutical industry (Best et al. [2020], Dallow et al. [2018]).

There are a myriad of different advantages and disadvantages of individual behavioural methods, however, there are a number of benefits and limitations which are consistent to all of these approaches when compared to mathematical aggregation techniques. The most powerful advantage of the behavioural methods is that mental models of the experts can be discussed and challenged through the process resulting in clearer more rounded judgements when elicited. The other benefit is that the outcome produced is recognised, and in the case of a true consensus, representative of the adjusted beliefs of the experts given all the information. This brings in a strong sense of ownership from the experts. This can be a powerful force to enable commitment to action. Mathematical techniques can suffer from the fact that there is no perceived ownership (and at times recognition and agreement), to the final numbers produced. Of course, behavioural methods need to ensure that the outcome produced does represent "consensus" and not simply "compromise". Behavioural methods are also conceptually simpler for experts to understand and can be approached in a "mathless" way. Mathematical methods can often be extremely complex and thus obtuse to the experts which can further damage expert sense of

ownership.

The most significant challenges facing behavioural methods are that they are susceptible to several psychological challenges. Behavioural methods risk issues such as group-think and deference to authority. All of these group behaviours can lead to consensus forming that is either overconfident or simply misinformed. Whilst some of these issues can be mitigated through strong facilitation it is not possible to eradicate them completely.

The ability to replicate is another constraint for behavioural methods. The result is essentially fixed to a particular date and cannot be repeated (without completely redoing your elicitation from scratch). One of the core tenets of mathematical models of aggregation is that they follow scientific principles and can be repeated and tested. In this way, it is much more common to see cross-validation of mathematical models, which leads to enhanced confidence and allows earlier elicitations to be re-utilised and re-examined. The final limitation that is pertinent to our stated aims at the start of the thesis is that inherent in behavioural methods is the philosophy that experts discuss and align on the issues at hand. This means that text book problems cannot have a behavioural aggregation solution.

In recent years, mixed models have emerged which combine behavioural approaches for elicitation, mathematical methods for aggregation and then bring results back for group review and discussion. These models aim to benefit from the advantages from each approach whilst mitigating some of the risk of the disadvantages. Whilst mixed models are still a fast evolving field (Petrolia et al. [2020]), of these approaches, the IDEA framework has gained significant traction. The IDEA framework was used as part of the IARPA good judgement project, which also bought the Superforecasting approach outlined by Tetlock and Gardner [2016] to the fore. Mixed models are the future of expert judgement elicitation. In Chapter 8 we outline how the model outlined in this thesis can be embedded into the IDEA framework.

One other expert judgement approach which is neither strictly a behavioural approach nor a mathematical approach is prediction markets. Prediction markets were very popular in the late 19th century, however they fell out of popularity in the early 20th century before seeing some resurgence at the end of the first decade of the 21st century (Green et al. [2007]). Prediction markets, based on classical economic theory, converts experts' knowledge and therefore probabilistic belief in an event into a willingness to pay. By creating a practical market in which options/futures can be traded, market belief can then ultimately be inferred based on asset price.

Prediction markets have a number of benefits, however, one of the major limitations is that often the topics over which expert elicitation is conducted are

sensitive (e.g. terrorism, cyber security, health, environmental policy). Prediction markets ask participants to essentially profit from outcomes on these sensitive topics. This can be seen un-favourably as allowing people to profit from harm coming to others. In the worst case potentially these markets could incite people to act unethically to ensure profit. A DARPA prediction market for intelligence had to be recalled the day after it was announced (Looney [2004]). Many methods for expert elicitation ask experts to think in terms of their willingness to place a bet, it is rare for them to actually ask them to place it. There are also practical issues to conducting a prediction market. Prediction markets require a very significant effort to setup and require a very large number of active participants. In order to be effective these participants must be continuously engaged. These practical considerations limit their deployment in many situations. The experience of prediction markets go to show that both the technical and ethical considerations regarding expert judgement are wide ranging and must be considered carefully before final methods are decided.

3.3 Review of common non-Bayesian mathematical approaches to SEJ

Earlier we outlined that in mathematical aggregation settings, common SEJ methods are typically either some form of weighted, linear (3.1) or geometric combination (3.2) of the experts' judgements. The key decision impacting the final assessment of the risk in this context becomes the choice of weighting criteria, ω_e, to use.

The first common and simplistic method to consider is just defining all ω_e to be equal (by definition they would thus be set to $1/|\mathbf{E}|$). In this way, under a linear combination the proposed aggregate of the judgements would thus just be the straight average.

It is important to note here that at each point in the elicited plausible values for X, this defines the average of the *probability distributions* from each expert. This is not the same as simply averaging the elicited quantile values. To demonstrate why, let us consider the simple example of two experts who have been asked to provide their judgement on the median, the 0.25 and 0.75 quantiles and the extreme values 0 and 1, (i.e. the maximum possible value below which there is no probability that X will occur and the minimum possible value below which there is absolute certainty that X will occur) for some variable $X \in \mathbf{X}$. Please note, in practice, extremes like this are very rarely elicited, however, it will make our calculation easier for this example. Suppose that the experts provide the perspectives outlined in Table.3.1.

If a DM were to simply average over the quantile values themselves (known as

	Quantile				
Expert	0	0.25	0.5	0.75	1
Expert 1	0	1.25	2.5	3.75	5
Expert 2	5	7.5	10	12.5	15
Quantile Aggregation	2.5	4.375	6.25	8.125	10
Linear Pooling	0	2.5	5	10	15

Table 3.1: Toy example of experts' elicited values.

quantile aggregation), they would arrive at their final values by doing a columnwise mean of the data. i.e. their 25th-centile would be $(1.25 + 7.5)/2 = 4.375$ and their 75th-centile value would be $(3.75 + 12)/2 = 8.125$ with a median expectation of $(2.5 + 10)/2 = 6.25$. If we calculate this by averaging the probability distributions, as per the linear pooling approach, rather than quantile values themselves, then we get a median of 5, and upper and lower quartile values of 2.5 and 10 respectively. The reason for the linear pooling values is not immediately as intuitive as the quantile aggregation values, but can be visualised with a simple plot of the aggregate probability density function, Fig.3.2.

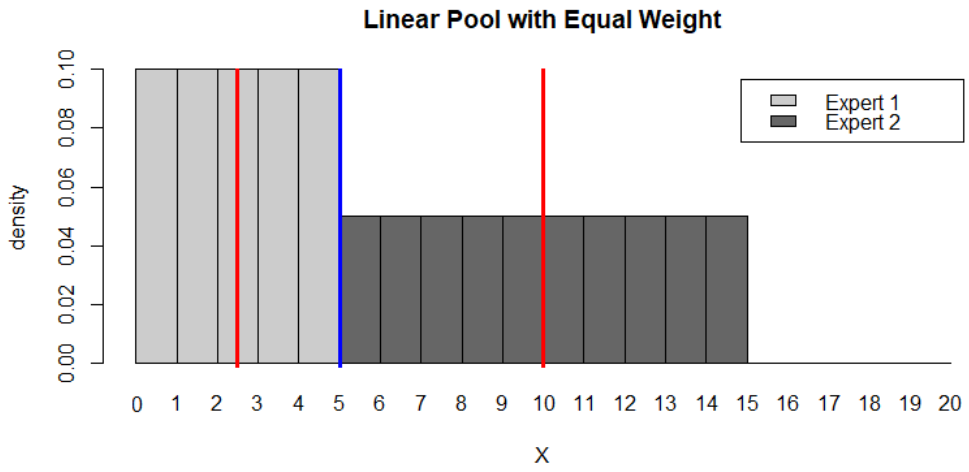


Figure 3.2: The probability density function of the equal weighted linear opinion pool in our toy example. Both the dark grey and light grey shaded areas are 0.5. The blue line denotes the median, 50% of the mass sits either side of this line. The red lines denote the outer quartile lines.

This example highlights that even in a very simple case with only two experts, these two seemingly logical methods lead to significant differences in the final

understanding of the uncertainty around X. The difference between these two methods can be thought of as either averaging horizontally (quantile aggregation) or vertically (linear opinion pool) across the cumulative distribution functions of the experts.

There is ongoing discussion over the validity of the quantile aggregation method vs. the linear opinion pool. Quantile aggregation has compared favourably in some cases (Lichtendahl et al. [2013]), although there is argument that the results cannot be extrapolated broadly. Large scale cross-validation has typically shown the quantile aggregation approach performs poorly relative to other aggregation methods, (Eggstaff et al. [2014], Colson and Cooke [2017], Cooke et al. [2020]). The quantile aggregation method does not require any distribution to be fitted so it is the most straightforward approach to implement. As a result, despite the questions remaining over the validity of the method it is one that remains in common use (Hemming et al. [2018]).

When only point estimates for X have been elicited from experts this is equivalent to them stating their prediction as a single value with probability 1. In these circumstances both the quantile aggregation method and equal weighted linear opinion pooling will lead to identical outcomes. This will simply be the mean of the experts' elicited values. With point estimates, the median can be as a robust measure as the mean which arises from either of these two methods. The seminal text, *Vox Populi* (Galton [1907]), kick started the use of multiple experts to improve predictive power. Within this text Galton described a weight-judging competition held at a stock and poultry exhibition. Competitors were asked to estimate the weight of an ox. In total there were 787 valid estimates the median of which was 1207pounds. This had an error of only 0.8 per cent vs. the true weight of the beast which was 1198pounds.

The simple averaging methods described demonstrate the advantage of multiple experts over single opinions. They do not necessarily however represent the best aggregation methods for these judgements. There has been significant research into methods of averaging opinions which aim to enhance the predictive accuracy of these combinations (Winkler et al. [2019], Winkler [2018], Cooke [1991]).

Other approaches to improving the aggregation typically look at shifting the weighting, ω_e, applied to each expert. One approach to enhancing the aggregation, that appears conceptually appealing, is to apply weights in some way according to the seniority of the experts (Ronen and Wahrmann [2005]). Here a distinguished professor with many publications in the area may be weighted higher than a junior researcher as it would be perceived their considerable experience would make them

better judges of potential uncertainty.

Unfortunately, as appealing as weighting in this way is conceptually, empirically it does not appear to be a strong method of aggregation. In the much lauded book, *Superforecasting: The art and science of prediction.* - Tetlock and Gardner [2016], the authors document how the approach to making predictions an expert utilises significantly outweighs experience or training when considering forecast accuracy. Those whose cognitive approach considered multiple sources of information, different explanations of what they were seeing and made judgements on the basis of more than one idea or theory were better predictors. The common attributes one might have anticipated to denote expertise, such as seniority, were a weak indicator of their ability to predict. To create aggregations according to such mechanisms therefore would risk artificially up-weighting poor experts and should be utilised with extreme caution.

Another potential method for identifying relative weightings for experts would be to consider whether experts can judge their own abilities compared to their peers (self assessment), or, the abilities of the other questioned experts (peer assessment). A review in 2011 (Burgman et al. [2011b]), was conducted to test this theory and to see how either self or peer assessment related to standard measures of expertise such as, age, experience, and qualifications. This study highlighted that, inline with the social expectation hypothesis, both self assessment and peer assessment was highly correlated with the standard measures of expertise. Furthermore, the study also identified that there was a strong correlation (0.85) between peer and self assessments of expected performance across the group of circa 110 life scientists considered. It was interesting to note, that there was a minor bias that peer assessment was stronger than self assessment, suggesting that individuals display a certain level of modesty, although those at the highest expected performance levels rated themselves on average higher than their peers, Fig.3.3.

Whilst perceived performance was highly correlated with these measures of expertise, when prediction accuracy was assessed, no such correlations arose. If these standard measures of expertise really did drive better prediction performance there should have been strong evidence of correlation here. Evidence for the lack of a relationship between perceived and realised expertise replicates findings in further work conducted in other domains (Cooke et al. [2008],Grove et al. [2000]).

The Burgman et al., 2011b study did highlight however, that discussions between experts following an initial elicitation did improve judgement accuracy when a second round was conducted. This aligns with the findings of Tetlock and Gardner [2016], that process is critical to the successful use of expert judgement. This concept

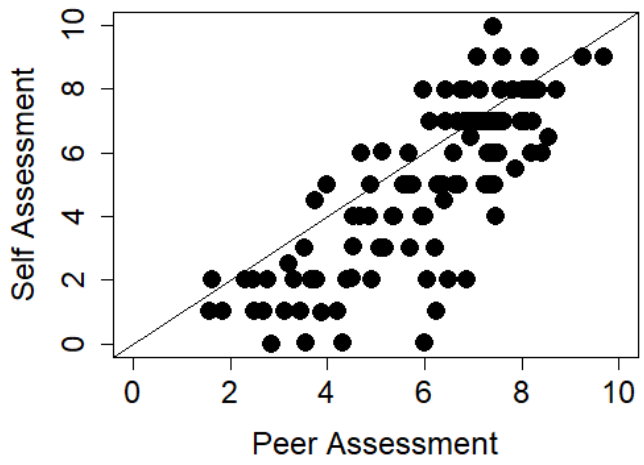


Figure 3.3: 'Self assessment' a score of how experts thought they would perform on technical questions vs. a 'Peer assessment' of how other experts within the workshops in Burgman et al. [2011b] expected them to perform on the same questions. A strong correlation (0.85) is evident between the two assessments. Replicated from Burgman [2016].

also forms the basis of the IDEA framework discussed more thoroughly in Chapter 8 (Hemming et al. [2018]).

If *perceived* expertise is not a good predictor of performance for experts in SEJ studies a DM must look to other methods in order to justify differentiated weightings. If no other methods were available then it would be logical to simply use unweighted opinion pools (or quantile aggregation) and put all of the SEJ research effort into processes and procedures rather than aggregation methods. Fortunately, for the more mathematically inclined SEJ researcher, there are other potential approaches to discriminate between different expertise levels in the studied group, **E**. These methods rely on *realised* expertise rather than *perceived* expertise.

'Realised' here denotes that experts make judgements on secondary variables, known as *seed variables*, in addition to the target variables of the study, **X**, for which there is known realisations *a priori*. These secondary variables could have been elicited historically and longitudinally. Imagine a weather forecaster who makes predictions every day, this builds a considerable database of historical judgements that could be tested. Or, they can be elicited at the same time as the target variables. Regardless of the mechanism of elicitation, all methods which consider these variables have a similar ethos. The judgements of each expert for these sec-

ondary variables are tested for accuracy using some performance metric, aggregated in some way to give a total expert score and then the relative scores for each expert are translated into the weightings ω_e.

An example of a seed variable can be found in a 2005 study considering the risks of campylobacter transmission during broiler-chicken processing, (Van der Fels-Klerx et al. [2005]). This study had target questions that needed assessing such as "What is the likely number of campylobacters (cfupercarcass),per processing stage,that will be found on the first carcass of the flock after it has passed the particular stage?" In order to arrive at seed variables for this question the study facilitators asked experts to consider a specific experiment performed on a flock of broiler chickens just before they are prepared for transportation into the processing plant in 1995, (many years before the study was conducted). An example of the seed variable questions then asked of the experts was what is "the number of campylobacters (cfu per gram) that is found in the caecal content of the broiler chicken just before it would have been transferred to a transport crate." Note here the similar domain and question structure between the target of the study and the seed variable questions posed to the experts.

Some researchers argue that finding these secondary variables is a fruitless task, (Bolger and Rowe [2015]). If the variables do not relate closely to the target variable of interest then they are not likely to be good predictors of each expert's performance. If however, a significant number of variables which meet the requirements of seed variables can be found, with known realisations, then this should simply be data and treated accordingly, thereby bypassing expert judgement altogether. The definition of suitably similar seed variables for different domains remains an open debate.

Linear pooling methods which test for this *realised* expertise in some way are denoted performance weighted opinion pools. The most commonly applied version of this is Cooke's Classical model. Given the importance of this model to the remainder of the thesis, we will outline this model in further detail before preceding further.

3.4 Review of Cooke's Classical model

As described in the introduction, the Classical model is the most common practically used form of performance weighted mathematical aggregation of expert judgement, and has "stood the test of time" (French [2011]). It is important to review this model before embarking on a detailed analysis of the Bayesian methods, but it also critical to highlight the issues with this model in the context of the motivated questions.

Based on the linear opinion pool, the format of the classical model is:

$$p_{\text{DM}} = \sum_{e \in \mathbf{E}} \omega_e p_e \tag{3.4}$$

Where the weights, ω_e are determined by the relative statistical accuracy and information provided by the expert.

Statistical accuracy considers how likely it is, given their performance on the seed variables, that an experts' predictions will reflect reality. Information is a measure of how broad/narrow the perspective of the uncertainty is that the expert provides. More informative predictions, i.e. those with narrower distributions, are more valuable to DMs than broader distributions (assuming they are accurate). Information is a concept closely aligned with 'sharpness' of forecasts, outlined by Gneiting et al. [2007]. Sharpness, in this context, is defined by the concentration of predictive distributions. Similar to information, sharpness is a property of the forecasts only.

Cooke's model attempts to balance these two concepts by creating a combined accuracy and information score for each expert. These scores are generated based on their performance over the seed variables. The final DM proposed distribution is an optimised weighting of experts given these combined scores.

3.4.1 Distribution fitting process

The Classical model, as with most well defined SEJ approaches starts with a detailed preparation and training exercise (Cooke and Goossens [2000]). This thesis will not comment on these elements given our focus on the statistical aspects of the approach, however Cooke [1991] is recommended for a thorough outline.

The elicitation process itself starts, as per our standard notation, with a set of experts, $\mathbf{E}$ asked to assess their judgement for a set of quantities of interest, the target variables, $\mathbf{X}$. In the Classical model however, the experts are also asked simultaneously to provide judgements on a set,$\mathbf{Y}$, of seed variables $(Y_1, ..., Y_{|\boldsymbol{Y}|} \in \boldsymbol{Y})$. Seed variables are additional quantities, elicited at the same time as the target variables, the true realisations of which are known to the DM, but not to the expert, *a priori*. The core tenet of the Classical model is to infer the calibration and information the expert is likely to provide to the unknown $\mathbf{X}$, given their performance assessing the known seed variables $\mathbf{Y}$. As is common in other methods (Albert et al. [2012], Wiper and French [1995],Clemen and Lichtendahl [2002]) distributions are elicited as a set of quantiles $(q_i : 1 \leq i \leq I$, for some $I \in \mathbb{N})$ against a predetermined set of probabilities (P_i) and then interpolated. A commonly elicited set of values

are for the 5%, 50% and 95%iles.

The first step in the modelling process, before considering any form of aggregation, is to extrapolate each experts elicited values on to the whole domain for each variable. This is done by fitting piecewise uniform components between the elicited quantiles.

Let us first define a seed matrix:

$$\mathbf{Q_Y} := (q_{Yei} : Y \in \mathbf{Y}, e \in \mathbf{E}, 1 \leq i \leq I : i \in \mathbb{N}) \tag{3.5}$$

And a quantile matrix of the target variables:

$$\mathbf{Q_X} := (q_{Xei} : X \in \mathbf{X}, e \in \mathbf{E}, 1 \leq i \leq I : i \in \mathbb{N}) \tag{3.6}$$

Here the matrix element $q_{Yie} \in \mathbf{Q_Y}$ can be interpreted as the response from the eth expert to the ith probability value for the seed variable Y. Similarly element $q_{Xei} \in \mathbf{Q_X}$ represents the eth expert's response to the ith probability value for the variable of interest X.

Given these two data sets, the decision maker first calculates a set of intrinsic ranges or bounds for the seed variable distributions. This is typically done through the use of an overshoot percentage ($t \in [0, 1]$). Let the true realisation of seed variable $Y \in \mathbf{Y}$ be y_Y then the modelled upper and lower bounds for $\mathbf{Y}$ are given by:

$$\begin{aligned} y_{Yupper} &:= max\{q_{YeI}, y_Y\}; e \in \mathbf{E} \\ y_{Ylower} &:= min\{q_{Ye1}, y_Y\}; e \in \mathbf{E} \end{aligned} \tag{3.7}$$

Then given overshoot percentage t (typically set at 10%), we define:

$$\begin{aligned} q_{Y0} &= y_{Ylower} - t(y_{Yupper} - y_{Ylower}) \\ q_{Y(I+1)} &= y_{Yupper} + t(y_{Yupper} - y_{Ylower}) \end{aligned} \tag{3.8}$$

The use of the overshoot percentage here ensures that each expert can have a probability distribution fitted for the whole interval [0,1], despite the extremes not being elicited. It also forces all experts to have non-zero probability on identical domains. q_{Y0} is a single value and does not vary with $e \in \mathbf{E}$ and similarly for $q_{Y(I+1)}$, hence the lack of an e in the subscript. For consistency, where necessary, we can set $q_{Ye0} := q_{Y0}$ and $q_{Ye(I+1)} := q_{Y(I+1)}$ $\forall e \in \mathbf{E}$.

The inferred probability density function, g_e, the Classical model uses for expert $e \in \mathbf{E}$ for $Y \in \mathbf{Y}$ is thus given by:

$$g_e(y|\mathbf{Q_Y}, e, Y) = \begin{cases} 0 & if \quad y \leqslant q_Y 0 \\ \frac{P_1}{q_{Ye1} - q_{Ye0}} & if \quad q_{Ye0} \leqslant y < q_{Ye1} \\ \frac{P_2 - P_1}{q_{Ye2} - q_{Ye1}} & if \quad q_{Ye1} \leqslant y < q_{Ye2} \\ ..... & \\ \frac{1 - P_I}{q_{Ye(I+1)} - q_{YeI}} & if \quad q_{YeI} \leqslant y < q_{Y(I+1)} \\ 0 & if \quad y \geqslant q_Y(I+1) \end{cases} \tag{3.9}$$

Reminder, g_e is an approximation to each expert's true belief, p_e, which is created by fitting a function to the elicited quantiles.

Expert e's approximated cumulative density function, G_e, for seed variable Y can then be considered a simple linear interpolation between the points:

$$(q_{Y0}, 0), (q_{Ye1}, P_1), (q_{Ye2}, P_2), \ldots\ldots (q_{YeI}, P_I), (q_{Y(I+1)}, 1) \tag{3.10}$$

The approximated distributions for the target variables are then calculated in an identical manner.

The fitted distributions are then utilised to calculate the information and calibration scores used to determine the weights, ω_e in 3.4.

3.4.2 Information

The Information (*Info()*) provided by an expert $e \in \mathbf{E}$ measures the additional value the expert provides to the DM relative to some background function $F()$. $F()$ is commonly the uniform or log uniform distribution. Simply put, *Info()* is a measure of the spread of uncertainty experts have given in their responses. An expert who is very confident in their judgement will have very tight distributions and an expert who is unsure will provide wide plausible ranges for the outcome. Clearly an expert who is sure of their judgement is giving more information to the DM than an expert who is less confident and so has more value. The weighting ω_e thus increases as the information e provides increases.

Info() is captured for each seed variable Y by:

$$Info(Y, e) = \sum_{i=1}^{I+1} (P_i - P_{i-1}) Ln(\frac{P_i - P_{i-1}}{F(q_{Yei}) - F(q_{Ye(i-1)})}) \tag{3.11}$$

This is effectively a Shannon entropy measure, a common tool in information theory (Lin [1991]).

The total expert information score is then the average over all seed variables:

$$Info(e) = \frac{1}{|\mathbf{Y}|} \sum_{Y \in \mathbf{Y}} Info(Y, e) \tag{3.12}$$

If an expert's distributions are very close to the uniform distribution the relative information score will be low, (it will be zero if they are exactly the uniform distribution), If an expert expresses strong certainty about the seed variables and provides a probability distribution accordingly then the information score will be high. Please see 3.4 for an example.

The final information score is determined by the choice of the overshoot percentage and thus q_{Y0} and $q_{Y(I+1)}$. Consequently, the information score cannot be calculated for an individual expert, ignoring the other experts in the group. It is a measure which is relative to both the background function chosen *and* the other experts in the study. This has important implications when discussing validation techniques and reproducibility.

Information alone however, is not a sufficient metric to support a DM in deciding how to weight judgements of multiple experts. Information only provides a perspective on how certain/confident an expert is, it does not support in the critical assessment of whether they are likely to be accurate or not. An expert who is incredibly certain in their beliefs, but who is wholly misinformed will be significantly less valuable, indeed potentially more dangerous to a DM, than an expert who is less informative but more accurate. To this extent, Cooke's Classical model pairs the *information* metric with a second metric, *statistical accuracy.*

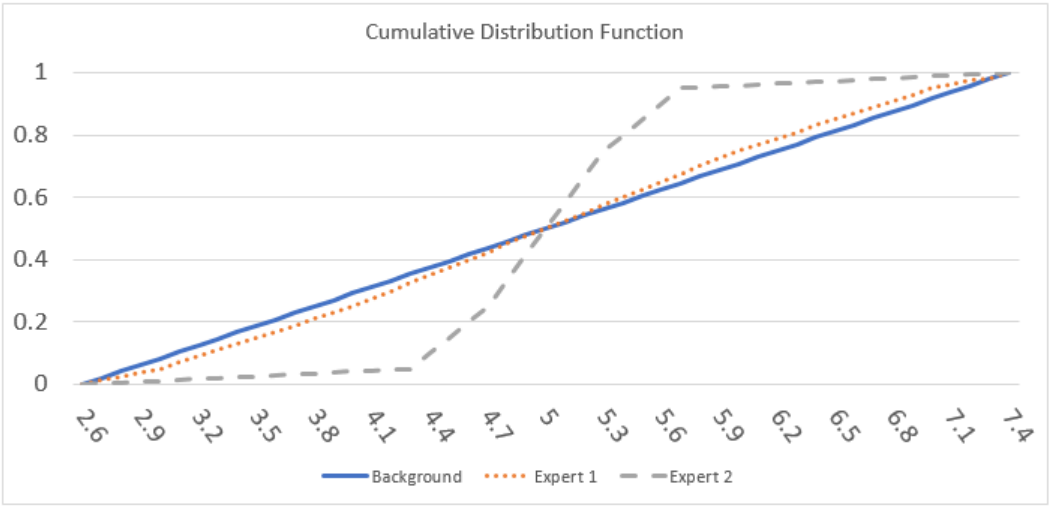
3.4.3 Statistical accuracy

The statistical accuracy score is used to test the degree to which the expert's fitted probability distributions map to true realisations over the seed variables (Wiper et al. [1994], Cooke [1991]). By specifying the quantiles associated with the probabilities P_i the experts are specifying an I bin multinomial distribution which has the probabilities p_i for each seed variable response. For each expert e, each seed variable realisation, y_Y, is the result of a multinomial experiment with:

$$Probability(y_Y) \in [q_{Ye(i-1)}, q_{Yei}] = p_i \quad \forall i \in I \tag{3.13}$$

Given that there are $|\mathbf{Y}|$ seed variables, for a reasonable sized $|\mathbf{Y}|$:

$$s_i = \frac{[\#seedvariables : y \in [q_{Ye(i-1)}, q_{Yei}]]}{|\mathbf{Y}|} \tag{3.14}$$



Q_Y

	P						
I	1	2	3	4	5		
P_I	5%	25%	50%	75%	95%	Realisation	
Expert	3	4	5	6	7	5	Lower Information
Expert	4.3	4.7	5	5.3	5.7		Higher Information

t	10%
y_{upper}	max(7,5.7,5) = 7
y_{lower}	min(3,4.3,5) = 3
q_{Y0}	3 - 10%*(7-3) = 2.6
q_{Y6}	7 + 10%*(7-3) = 7.4

Expert 1: Info(1) = 0.024

Expert 2: Info(2) = 0.848

Figure 3.4: Expert 1 is barely distinguishable from the background distribution and provides little information to the DM. This is reflected in the corresponding Information Score Info(1). Expert 2 conversely provides more value through a very narrow probability distribution reflected in their score Info(2).

Is a reasonable empirical estimate of p_i. e.g. if you had elicited two quantiles from an expert either side of a median, for a large enough set of seed variables $\mathbf{Y}$, for a statistically accurate expert you would expect that 50% of the values would fall above the median and 50% below the median.

We can now test the extent of the expert's true statistical accuracy through hypothesis testing. Set:

$$\mathbf{H_0}: \quad s_i = p_i \quad \forall i \in I \quad \mathbf{H_1}: \quad \exists i : s_i \neq p_i \tag{3.15}$$

Then returning to the information/entropy definition we utilised before and repurposing slightly, we can set:

$$Info(s,p) = \sum_{i \in I} s_i Ln(\frac{s_i}{p_i}) \tag{3.16}$$

Now if $|\mathbf{Y}|$ is large enough:

$$2|\mathbf{Y}|Info(s,p) \sim \chi^2_{(I-1)} \tag{3.17}$$

Thus the statistical accuracy score can be calculated as:

$$StatisticalAccuracy(e) = 1 - Pr\{\chi^2_{(I-1)} < 2|\mathbf{Y}|Info(s,p)\} \tag{3.18}$$

3.4.4 Weighting

The final weighting for expert e is then determined by:

$$\omega_e \propto Info(e) * StatisticalAccuracy(e) \tag{3.19}$$

The experts' forecasts can then be aggregated as per equation 3.4, where experts are removed from the sum if their weights do not meet a cut-off threshold and the remainder renormalised.

A multiplication and the removal of experts here is utilised as this ensures that the final weighting represents an asymptotic proper scoring rule in which experts are rewarded for proposing their true beliefs (Cooke [1991]). The statistical accuracy measure is also a faster function than the information score and so Cooke's model is more sensitive to improvements in accuracy than information.

Please note, this represents the basic elements of the Classical model, and there are a number of extensions of this such as the use of global/item based weighting or the inclusion of an alpha value, DM optimisation and the DM within the pool. However, we shall not go through these here, for more information on these please refer to the original material (Cooke, 1991).

3.4.5 Benefits and limitations

The number of practical applications of the Classical model in decision problems is testament to the clear benefits this framework has. One of the primary advantages is simplicity for the DM. It is important in many contexts aggregation approaches do not appear to be a black box to the DM. Intractability can result in the final outputs

not being believed and ultimately utilised. The weighted linear opinion pool will be familiar, even if not in terminology, to all most all DM's. For the Classical model weightings it is easy to see how an analyst could explain to the DM the principal components, if not the technical details, of both the statistical accuracy and the information elements and thus why a DM should adjust their beliefs accordingly.

The most important benefit of the Classical model however is that it has proven to be robust relative to other opinion pooling techniques when empirically validated. Cooke [2014] has maintained a growing database of over 80 studies conducted using this methodology and made these publicly available. Significant energy has been invested into validating the results of the model both by Cooke and other researchers. Very few other SEJ methods have been validated in such a way and this form of validation sets a benchmark for all future algorithmic methods. The Classical model has been compared to the SHELF behavioural method more explicitly recently (Williams et al. [2020]). This analysis demonstrated that SHELF outperformed the Classical model on the small data set considered. This is exactly the type of validation exercise required, but with only a small case study it is difficult to extrapolate more generally. To date there have been no comprehensive validation attempts conducted using Bayesian methodologies. Significantly more effort is required to empirically validate different SEJ approaches.

Both the simplicity and proven robustness have led the Classical model to be a common SEJ tool.

There are however, a number of limitations to the Classical model approach. Whilst the method does consider diversity of opinion between experts it does not explicitly model consensus nor capture expert differences from that consensus (Albert et al. [2012]).

Whilst clearly beneficial to the expert problem, the applicability to both the group decision problem and the textbook problem is not so clear. In practice in a number of the studies within Cooke's database a large number of the experts are set with a final weighting of zero. This means that, with the exception of the most extreme bounds q_{Y0} and $q_{Y(I+1)}$, which will still have taken all experts into account, only a few impact the final result and the bulk of the probability density (typically where the mass of circa 90% sits). In the case of the group decision problem, this could pose a significant issue. Removing experts may appear undemocratic as the final decision does not truly represent the opinion of the entire group of experts (French [2011]).

The statistical accuracy calculation of the Classical Model relies on a significant number of seed variables, however, ultimately only uses these as a scoring

mechanism. If an expert were to display consistent bias, their calibration score would be low and their impact on the final decision would be minimal. If it were possible to adjust for this bias in a rigorous way, however, their full judgement could be utilised to help inform the DM. Gathering judgement from experts is expensive and thus it is critical to both look at modelling options which limit the number of data points to elicit and to get the maximum amount of information from each elicited data point. In the case of the textbook problem there is not readily available access to the experts and as such the amount of available data is inherently limited to what was historically captured. Obtaining the maximal information per data point is even more critical in this context.

One other potential challenge of the statistical accuracy calculation in the Classical model is that there is no distance measure implicit in the mechanics of the calculation. Two experts with the same number of seed variable realisations in each of the probability bins (defined by the elicited quantiles), will have the same p-value and thus the same statistical accuracy score. This could be despite the fact that one experts median judgement is consistently closer to the true realisations. Again, this means that potentially valuable data is being lost. In at least one instance this phenomenon has been identified by an expert in a study and challenged whether this could be right or fair (Aspinall [2021]).

The final challenge of the Classical model is expert overconfidence. It is widely recognised that experts are overconfident with regards to the plausible bounds they put around events. This phenomenon is particularly true for very rare events often of most interest in SEJ studies. The Classical model utilises "overshoot percentages" in order to ascertain the domain over which each experts fitted distribution will apply. This overshoot percentage therefore has an implication on the total feasible domain the model predicts. There are no robust methods for defining the correct overshoot percentages to use and often standard heuristics (such as 10%) are utilised. A DM, aware of the comprehensive literature on expert overconfidence, may well challenge that any result coming back from an expert elicitation process does not reflect the true uncertainty, unless this overconfidence is in some way considered. Whilst the Classical model therefore potentially produces an empirically robust aggregation of experts' judgements in isolation, it does not identify how a DM should update their belief in light of this, given DM knowledge of overconfidence. As overconfidence is expert specific, the final aggregated output of the Classical model will mean it is unfeasible for a DM to account for overconfidence *a posteriori*.

The above challenges to the Classical model are not there to indicate that the approach has major deficits, they are more nuances that have arisen due to

substantial use of the approach. They do suggest however that the Classical model is not a panacea for SEJ, particularly in the context of the motivated topics. It is consequently meaningful to identify other frameworks which demonstrate the advantages of the Classical model whilst, reducing the elicitation burden on experts, utilising all of the available data, considering overconfidence and taking into account both diversity and consensus between experts. In order to do this, we shall consider Bayesian approaches to expert judgement.

3.5 Review of Bayesian approaches to SEJ

Bayesian models have evolved considerably since their inception. The first models used conjugate prior methods, which simplify the calculations by utilising a restricted set of distributions in the Bayesian model to make the mathematics easier. Examples of such models may be found in Winkler [1981], Lindley et al. [1979], Mosleh and Apostolakis [1986] and Wiper and French [1995]. These models conceptually demonstrated the power of the Bayesian approach, often producing favourable results on small datasets, however, in practice were not broadly adopted. The reasons for this relate to the restrictions the conjugate assumption put on the model, the complexity in modelling approach in comparison with intuitively simpler opinion pooling methods, and the sensitivity to inputs that was apparent in some of these approaches.

Following the conjugate prior models there was some further investigation into other Bayesian approaches that could be fruitful. Some progress was made on Bayesian Nonparametrics (Lichtendahl [2005]) and Copulas (Jouini and Clemen [1996]). Working with multivariate distributions – those with more than one dimension, a phenomenon prevalent throughout structured expert judgement (SEJ) – can increase complexity significantly; however, copulas simplify the process by separating what are known as the marginal distributions, which are distributions for the individual dimensions, from the dependence structure which demonstrates how they are linked together. There were some positive signs from these models, however, empirically there were questions over the method. Copulas can be very peaked at the extremes and as a result have numerical stability issues. In the extreme, if a single experts' PDF goes to zero at any given point then the total result will forced to zero (Kallen and Cooke [2002])[3]

[3]It is important to note that the models themselves encode expert judgement and therefore cannot be thought of as ad-hoc. Indeed it would be reasonable to question whether it is appropriate to separate the elicitation exercise from the modelling process, however in doing this some of this encoded knowledge would be lost. Therefore we need to be very careful with the treatment of

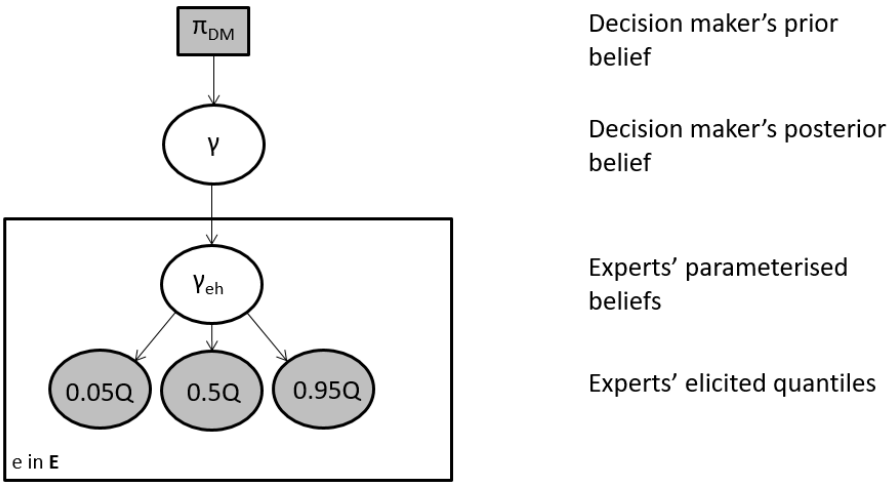


Figure 3.5: Bayesian network for aggregating expert judgement.

In more recent years there has been some resurgence in the use of Bayesian analysis for expert judgement studies; however the focus has now shifted to Markov Chain Monte Carlo (MCMC) approaches. These methods are a class of algorithms for stochastically approximating Bayesian posterior multi-dimensional distributions. Despite MCMC approaches having existed for many years, the use of these techniques in expert judgement studies is a recent development. One of their attractions to Bayesian expert judgement studies is the ability to describe the model through the use of a Bayesian Network. This seemingly simple box and arrow diagram, Fig. 3.5, can allow the analyst to easily demonstrate to the decision maker what the different variables or parameters are and how they are linked together, thereby hiding much of the modelling complexity (Charniak [1991]).

Each graph outlined in this thesis has three components; nodes, edges and plates. The nodes represent the quantities in the statistical model. Rectangular nodes denote constants and elliptical nodes are stochastic variables. Deterministic nodes are logical functions of other nodes and will also be elliptical. An edge defined by a solid arrow indicates stochastic dependence between the variables. An edge with a dotted arrow indicates a logical function. Grey shapes are known values. In order to keep the size of the graph small, repeated parts of the graph are represented using a plate (a large rectangular box). The plate will contain an index in the bottom left corner which will denote the element that is repeated.

In this toy example we see a simple expert judgement aggregation model[4]

elicited data outside of the modelling paradigm, a particular challenge in the textbook context outlined earlier.

[4]It is interesting to consider whether it would be possible to create a Bayesian network (BN)

described. Here, experts have their beliefs elicited in the form of a set of quantiles, these quantiles are utilised as representatives of the experts "true" belief which is in the form of a parameterised distribution with parameters γ_e (e.g. in the standard Gaussian $\gamma_e = (\mu_e, \sigma_e)$). Quantiles are often elicited rather than parameters due to the relative complexity of asking experts to think in terms of distributions. The decision maker's beliefs γ, again parameterised, are as in the standard Bayesian model outlined, updated based on the underlying prior he/she had π_{DM}, combined with the probability assessments from each expert. The actual inference over this network would be produced by running an MCMC algorithm which would infer an output distribution for each of the unknown variables. Another advantage of MCMC algorithms is that they maintain uncertainty throughout and therefore it is possible to visualise the underlying distributions for each of the points of interest within the model, such as the experts' true beliefs or the decision maker's full posterior distribution. With the appropriate treatment, this also allows the decision maker to understand more complex items such as the correlation between experts or their relative calibration, without explicitly trying to elicit or model these separately.

One of the early MCMC models for SEJ was from Clemen and Lichtendahl [2002]. In this paper the authors tackled a specific portion of the expert judgement problem, the issue of expert calibration. Building on ideas from Cox [1958] and Morris [1974]; Clemen and Lichtendahl [2002] developed a model of expert overconfidence using past data to estimate, what they term, "inflation factors" for assessed distributions post hoc. While some common models treat all experts as exchangeable, Clemen and Lichtendahl use hierarchical MCMC models which allow experts to be calibrated individually. Here, for simplicity we can imagine, the model has a parameter, α, for each expert which describes whether that expert displays consistent bias (i.e. continuously over/under estimating) on their best guess, (or 0.5 quantile) across forecasts. The MCMC algorithm then sequentially reviews experts' previous performance at forecasting, over the seed data, and infers the value for α (it will ultimately be a distribution rather than a point estimate). From this the decision maker can decide how to consider each experts judgements for future forecasts. The authors' then extended this model to consider the other elicited quantiles

for the group problem or the textbook problem in addition to the expert problem. Unfortunately this is significantly less trivial. For the textbook problem, by definition, the problem statement is not known at the time of elicitation and therefore it is impossible to generate a corresponding BN. It would be feasible to generate specific networks as individual problems are solved but a generic version does not exist. For the group problem, there is significant work looking at complex decisions involving both groups and Bayesian networks in the field of adversarial risk analysis. Due to the focus of the EFSA guidance on the expert problem, a detailed review of group decision problems is not given in this Thesis. One paper covering this topic is French [2011].

and expert to expert correlation by creating further parameters, which represent the non-independence of experts.

Clemen and Lichtendahl did not explicitly consider the choice of variables used for calibration, though this is clearly important. As mentioned, the underlying assumption of all calibration techniques is that the behaviour experts' display on the seed variables is indicative of their final behaviour on the variables of interest. In particular, there are systematic biases, of a similar nature to those outlined by Kahneman et al. [1982], in consistent evidence which should be removed from the decision maker's analysis. The case for this is compelling but critically, only when the variables used for calibration are representative of the target variables of interest. A decision maker should not expect an expert's performance with relation to a weather forecast to be indicative of their ability to accurately assess the likelihood of a bolt breaking in a suspension bridge. Similarly the data must be on a similar scale. Experts are notoriously inaccurate when assessing probabilities for extremely rare events, and one would expect that behaviour seen here would not correlate with behaviour seen for more commonly occurring variables (Slovic and Weber [2002], French [2011], Kahneman and Tversky [1979], Koehler et al. [2002], Goodwin and Wright. [2010]). Assessing the right variables to use for a calibration model remains a question, and something that should be researched further.

Two critical elements in assessing appropriate calibration questions and variables should be considered. The first is the domain of the question. Currently the decision on whether the domain of a set of calibration variables is similar enough to be useful to infer information about the target variables is very challenging and guidance here is important. It will clearly vary from context to context but some specific guidance on how to search for and develop these questions within a domain would be beneficial. The second is the structure of the questions themselves. Even minor fluctuations in language and framing could have impacts on the inference that can be drawn. Ideally all calibration questions would have identical structure to the target questions but in practice this is not always feasible. Boundaries on the extent to which variation here is acceptable would be valuable.

Although Clemen and Lichtendahl tackle the question of how we calibrate multiple experts whilst assessing the expert to expert correlation, they do not consider the issue of what a decision maker should do once he/she has received this data. Utilising the authors' methods a decision maker would be able to translate multiple experts elicited quantities into their unbiased counterparts, however, how the decision maker would actively use these is not apparent. It would seem a shame to precisely calibrate experts but then for the decision maker to update his/her be-

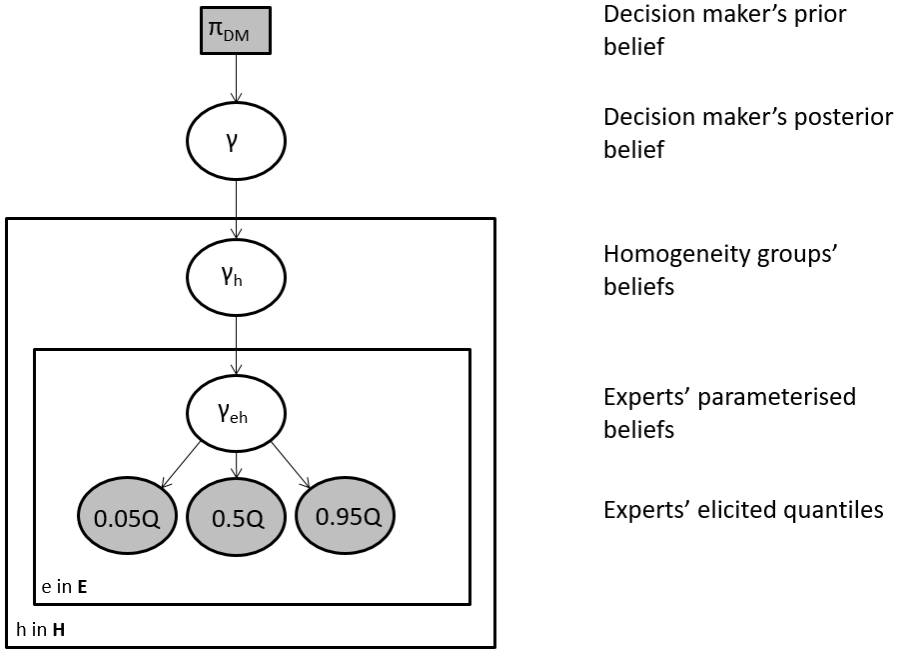


Figure 3.6: A Bayesian Network for aggregation of expert judgement with homogeneity groups.

lief in a method that does not use this richness of information. To this extent, it is important to examine Bayesian methods of *aggregating* the data also.

More recently, Albert et al. [2012] proposed such a model. Their model is known as a Supra-Bayesian parameter updating approach. This outlines a class of models which assume that the aggregation represents the belief of an overarching rational, but hypothetical, decision maker, the Supra-Bayesian, who has beliefs that can be represented by particular parameters. For example, they may believe that the output is Gaussian with an unknown mean and variance. The model then updates these parameters based on inference over the experts judgements (here, as usual, utilised as data). Similar to the calibration model, this method also considers expert judgements from indirect elicitation, i.e. rather than trying to elicit a mean and a variance for each experts beliefs, expert's knowledge is elicited on more intuitive observables, such as quantiles, and the parameters then inferred. Here the inference is made by mapping the elicited quantiles (or similar) to a selected parameterised distribution, using distribution fitting. Different models often use different parameterisations for this. The parameters of the fitted distribution are assumed to represent the expert's underlying belief that often cannot be directly elicited due to the complexity of mentally processing these concepts.

The model that Albert et al. [2012] use is hierarchical in nature and captures correlation in an interesting way. One of the drivers of inter-expert correlation is that experts may have had similar education or historic frames of references. The proposal from the authors of this paper therefore was to group experts together into homogeneity groups, where each group is defined by like-minded experts. Here the authors take like-minded to mean "similar background or schools of thought", although do not go into specific detail on how this may be assessed. For now we will assume that experts can be appropriately grouped in some way, however we will return to this shortly. The aggregation model will then assume that each expert's beliefs are linked to that of the other experts in their group and the groups likewise are linked to each other. Each group, h, will have a parametrised distribution γ_h defined by the beliefs of its expert members. The final combined posterior distribution represents the updated decision maker judgement and is calculated through MCMC. A simple diagram of this model is shown in Fig. 3.6. This is an extension of the model outlined earlier.

The motivation for this expert partition is that rather than explicitly calculating the correlation for each expert, the grouping approach is used to appropriately weight the impact of each expert in the final model, offsetting overconfidence effects driven by correlation. The theory here is that past experience and knowledge is one of the underlying key drivers of this correlation. One of the advantages of this approach is that the hierarchical model can capture both the consensus and diversity between experts, and this is very compelling.

As mentioned above, one of the areas not overtly tackled by the authors was how to support a decision marker or analyst in assigning experts to groups. More recently, in Billari et al. [2014] the authors proposed a relatively parsimonious Bayesian aggregation method considering mixture models and in Perälä et al. [2019] the authors proposed an interesting model for calibration utilising Gaussian hierarchical processes. In none of the models outlined however, has calibration, aggregation and homogeneity group definition been attempted within a single framework which is the aim of the model outlined later within this thesis.

Overall, with these recent developments, it appears that the goal of finding a practical Bayesian framework for SEJ is not an impossibility and MCMC may be a critical design element. The more recent models outlined also show the potential for the Bayesian approach to take a significant step forward in being more context agnostic. There is, however, a number of complex idiosyncrasies that make these techniques challenging in a practical environment compared to the Classical model, in particular:

- Technical details being intractable to non-analytical decision makers.
- Complexity (and model overdependence) in setting the correct priors.
- Reliance on very large volumes of calibration data which can be expensive to elicit.

Some of these issues can be resolved by research into the modelling techniques utilised, however, others can be better addressed by considering the processes and procedures that may need to be different for a Bayesian model of SEJ.

Chapter 4

Validation of SEJ models

4.1 The challenge of validation for SEJ

Like any good scientific discipline, empirical validation of different aggregation methodologies should be available to DMs to support in the choice of SEJ methodology. Unfortunately, with a few notable exceptions, a comprehensive body of empirical evidence for different methodologies is not widely available.

One reason for the lack of comprehensive evidence is that the applications of SEJ often occur in very sensitive domains. I work for a company which applies SEJ heavily to understand the probability of technical and registration success of medicines in a pharmaceutical R&D pipeline. The judgements collected here are categorised as confidential and sensitive information as they relate directly to the likelihood of different drugs entering the market, and therefore the potential impact on patients and company value. Consequently, whilst internal validation for this type of data is often conducted by comparing observed frequencies of project successes to predicted probabilities, this data will never be public domain and therefore not widely accessible for the process of methodology validation.

Even in public sector projects the data may not be widely made available. A recent SEJ study by the department for Business Energy and Industrial Strategy assessed the likelihood of mass interruption to the UK's electricity supply. Whilst it is in the public interest for such risk assessments to be conducted for contingency planning, making the actual results public would potentially lead to public anxiety if misunderstood or worse, direct intervention from external forces, if there were perceived vulnerabilities.

Even when any data involved is in the public domain, validation is still often not performed. This is driven by a number of factors. One primary reason is

that there is little professional motivation. The risk that judgements may be found wanting acts as a disincentive for the exercise overall. Avoidance of validation of expert judgment leads to *argumentum ad verecundiam* (argument from authority) reasoning. Argument from authority is well studied in philosophy and highlights the logical fallacy of appealing to an expert's testimony outside of that individual's specialised field of expertise, Walton [2010]. Any expert's judgments are fallible even within their specialist domain. Further risk arises, as without submitting to any form of appropriate testing, the bounds of an individual's specialised expertise are not known.

Another challenge is that events often unfold over time scales that preclude useful feedback. The inherent rarity of events assessed, or the lack of ethical means of data collection, which make validation difficult are the same drivers which lead to the studies in the first instance. Consequently, standard models of assessing forecast accuracy of data driven prediction modelling are rarely feasible. There are two primary concerns when designing a validation framework for expert judgement studies. Firstly, how to generate a significant sample of testing data for which there are both modelled aggregate judgements and realisations. Secondly, which testing methodology to use to assess the validity of the final predictions on this testing set.

4.2 Building a testable data set

Significant research on validation within SEJ models is relatively new and there is much work to do in order to formally define an agreed upon approach. Several different methods to building the testable set have been proposed.

The first and arguably most simplistic approach of building out a testable set is to simply wait for a significant enough period of time for a body of evidence with true realisations to be generated. This can occur either by conducting a study at a single point and then waiting until the target variable outcomes have been realised or, by conducting longitudinal studies, for which judgements over events are realised through the course of the study. An example of this was the IARPA tournament outlined in Tetlock and Gardner [2016]. It is important when considering methodology validation to carefully manage these two contexts. In the case of longitudinal studies, experts have the opportunity to learn, both about events and their own prediction capability. As a result, the corpus of knowledge upon which predictions are built is not static. This could lead to misleading comparisons if not appropriately addressed.

The second method of generating testable data to compare methodologies

is to conduct studies explicitly for this purpose (Williams et al. [2020]). Experts are asked to only predict variables for which the investigators know the answers already. This is equivalent to generating an SEJ study with only seed variable questions, $\mathbf{Y}$, and no target variables $\mathbf{X}$. This method is advantageous as it does not require significant time to wait in order to generate results, multiple methods can be used simultaneously so both process and mathematical method can be tested and it is largely agnostic to the approach considered. Behavioural and mathematical aggregation methods can be tested in this way. The downside of these approaches are that they are extremely expensive to conduct. Utilising the time, whether as part of a face to face or remote elicitation exercise, of the experts required to conduct such a test without making predictions on variables for which there is direct application may be perceived as wasted opportunity. Generating a significant enough set of data to create statistically meaningful tests in this way is also challenging. It is also important to consider how the ordering of the application of the tested methods may impact the findings and so they need to be very carefully managed. These methods have been rarely used.

4.3 Cross-validation

When time scales are too long and events are too rare to justify waiting for realisations and it is too costly to conduct tests with enough data specifically, methodologies for SEJ validation typically involve cross-validation. Cross-validation also only considers judgements made about seed variables. Seed variables by construct have both judgements and known true realisations. As seed variables are already collected at the same time as target variables, there is no additional overhead in utilising these for methodology validation purposes as well.

Cross-validation involves generating some partition of the seed variables. This partition divides the seed variables into two components, the first a *training set* and the other a *testing set*. The proposed methodology is then run, utilising the *training set* as data within the analysis, with the aim of predicting the *testing set* outcomes. These predictions are then measured against the true realisations of the *testing set*. Cross-validation of this form is widely used outside of SEJ to assess how the results of statistical analysis will generalise to independent data sets, (Shao [1993], Browne [2000], Picard and Cook [1984], Arlot and Celisse [2010], Kohavi [1995]).

Clemen [2008] utilised such a technique using a method known as ROAT (Remove One At A Time). Here, each seed variable is removed from the training

set one at a time and all remaining variables are used to train the model; i.e. if there are $|\mathbf{Y}|$ seed variables in the data set, each training set will be of size, $|\mathbf{Y}|-1$, each testing set of size 1, and there will be $|\mathbf{Y}|$ final forecasts. ROAT is a fast method of cross validation as relatively few judgements need to be made, however, it was demonstrated that this method could have an inherent bias against a performance weighted decision maker (Cooke [2008]). In other domains, ROAT is known as a jack-knife test (Miller [1974]).

Other methods of cross validation considered have utilised bigger training subsets, (Colson and Cooke [2017], Lin and Cheng [2009], Flandoli [2011] and Eggstaff et al. [2014]). Rather than removing a single variable for prediction, these methods remove multiple variables at a time. Each of the cross-validational studies conducted in Colson and Cooke [2017], Lin and Cheng [2009], Flandoli [2011] and Eggstaff et al. [2014] considered different sized training sets determined by the method of testing chosen. Extrapolating from one to many variables in the testing set is known as either K-fold analysis or Leave-p-out cross validation depending on how exhaustive it is.

Please note, as detailed in Cooke [2016], for the Flandoli [2011] and Lin and Cheng [2009] studies, questions arose over the implementation as the numbers quoted did not align with those generated in the original studies and from the original study software EXCALIBUR (Cooke and Solomatine [1992]). In the case of Lin and Cheng [2009] there were discrepancies which could not be explained, even when analysed, relative to a study conducted by Cooke directly (Cooke and Goossens [2008]). The code used by Flandoli [2011] was assessed after the analysis was published and found to both optimise to conflate background measures and, for the application of the Classical model, optimise incorrectly. This highlights some of the challenges, given the complexity, when conducting cross-validation that needs to be very carefully managed to ensure that the validation technique accurately reflects the true behaviour of the model being tested.

Leave-p-out cross-validation works by selecting p observations from the original set of seed variables as testing data and then utilising the remaining items as training data. This approach is then repeated for all potential subsets of size p. This is an exhaustive cross-validation method as every possible combination is calculated. This relies on running the model $C_p^{|\mathbf{Y}|}$ times as the number of potential subsets of size p is determined by the binomial coefficient (Celisse [2008]).

K-fold analysis is a non exhaustive method which aims to reduce the calculation burden by shrinking the number of subsets which are predicted. Here, the total set of seed variables is randomised and then split into K subsets of equal size.

One subset is taken as the testing set and the remaining $K-1$ subsets are used as the training set. The selection process is repeated K times so that each testing set is predicted once. This will result in a total of $|Y|$ individual item predictions generated from K modelling runs (Fushiki [2011]). If p is selected to be $|Y|/K$ then the K K-fold modelling runs will be a small subset of the $C_p^{|\mathbf{Y}|}$ modelling runs from Leave-p-out cross-validation.

Clearly, when K is set to $|\mathbf{Y}|$, or p is set to one, then both Leave-p-out and K-fold cross validation is identical to ROAT analysis.

Arguably the most comprehensive SEJ cross validation was outlined in Eggstaff et al. [2014]. Within this cross-validation model, for each of the studies under consideration, the authors calculated every permutation of the seed variable partitions from 1 to $|\mathbf{Y}|$. This is a leave-p-out cross-validation for every possible value of p and represents a completely exhaustive cross-validation method.

Whilst this approach is the ideal mechanism to use for cross-validation it relies on an extremely large number of forecasts. This would be a significant struggle to replicate at scale for other modelling approaches. The number of subsets of a set of size n $= 2^n$ and the total number of individually forecasted items is calculated by:

$$No\ Forecasted\ Items = \sum_{p=1}^{p=|Y|-1} p * C_p^{|\mathbf{Y}|} \tag{4.1}$$

For a single study of 10 seed variables, this would create 1022 forecasted subsets (both the empty set and the complete set are removed) and 5110 individual forecasts.

Repeating this scale of validation for other aggregation approaches may not be feasible and as studies potentially get bigger could end up intractable. Colson and Cooke [2017], whilst building on the work of Eggstaff et al. [2014], recently recommended considering all permutations of training subsets 80% of the size of the original set of seed variables. This value was generated by extending the analysis of Eggstaff et al. [2014] and attempting to balance calculation burden and potential bias generation. Training subsets of this size create a manageable number of forecasts to perform whilst overcoming some of the biases particularly evident in the ROAT methodology. For all cross-validation analysis outlined later we have utilised this 80% methodology. For a study of size 10 this creates 45 training subsets of size 8 and 90 resultant forecasts (two for each model run). If 80% was non integer we have shrunk the training set size to the nearest integer, and where necessary the minimum number of variables removed was set to 2 to ensure the methodology was

not applying a ROAT process.

4.4 Measuring predictive accuracy

Methodologies for assessing the accuracy of the given testable sets, known as scoring rules, also vary, and there is further opportunity for research and consolidation on an agreed approach here. One simple method that is considered in many studies is to ignore the uncertainty bounds within the predictions and simply assess the accuracy of the median within the distribution, assuming that this represents the most likely value a DM would use in practice.

The test metrics considered when looking at the individual forecasts are drawn from the extensive literature outside of SEJ. There are simple measures for assessing accuracy, such as Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE) or the Root Mean Square Percentage Error (RMAPE). These methods are broadly applied but can be strongly impacted by a small number of divergent responses.

Other approaches, when comparing multiple aggregation methodologies, try and standardise forecasts across variables in some way in order to allow each to contribute equally. One example of this is the average log-ratio error (ALRE). This method starts by range-coding the estimates for each model type (and the true value) for each question. Each forecast is rescaled as follows:

$$\overline{M}_Y^i = (M_Y^i - M_Y^{min})/(M_Y^{max} - M_Y^{min}) \tag{4.2}$$

here M_Y^i represents the median estimate modelling methodology i makes for $Y \in \mathbf{Y}$. M_Y^{max} and M_Y^{min} represent the maximum and minimum median values of the tested methodologies predicted for Y if the true value sits within this range or the true value itself if it is outside this range. The ALRE is then calculated as:

$$ALRE_i = \frac{1}{|\mathbf{Y}|}\sum_{1}^{|\mathbf{Y}|} |log_{10}(\frac{\overline{y}_Y + 1}{\overline{M}_Y^i + 1})| \tag{4.3}$$

where $\overline{M}$ is calculated as above and $\overline{y}_Y$ denotes the range-coded realised value for Y. Under this method smaller scores denote more accurate prediction with the minimum possible score for a perfect forecaster being 0. This method is as per Burgman et al. [2011b], but replacing the different experts with different modelling approaches.

When predictions within the testable set are probabilities then a commonly

used metric is the Brier score (Tetlock and Gardner [2016], Walker et al. [2003], Poses et al. [1990]). The Brier score can be thought of as a cost function. It measures the mean squared difference between predicted probabilities assigned to outcomes and the actual outcome itself. It is appropriate when considering binary outcomes. The basic calculation for $|\mathbf{Y}|$ events would be:

$$BS_i = \frac{1}{|\mathbf{Y}|} \sum_{Y \in \mathbf{Y}} (p_Y^i - y_Y)^2 \tag{4.4}$$

where p_Y^i denotes the probability that modelling approach i assigned to the seed variable Y occuring and all other terms are as before. Again, similar to ALRE the smaller the Brier score the better the prediction performance. The perfect predictors would get a score of 0 and perfectly inaccurate predictors would get 1. Note, in the original Brier definition there was a double summation applied to the Brier score which results in calculations being on the range 0-2 rather than 0-1. This extension allows for multi-category forecasts, although this is now less common.

When applied, the Brier score can be broken down into a number of sub-components one exmple of which is calibration and expertise (Walker et al. [2003]). These breakdowns are used for analysis of the rationale behind the final score given, but the sum of their values equals the base Brier score.

The Brier score has been used extensively in mass expert judgement studies (Tetlock and Gardner [2016]), but its application is not always appropriate (Cooke [2014]). If we took a blanket assumption that the probability was zero for rare events which actually occurred with frequency 10^{-3}, then the difference in Brier score from this blanket assumption and perfect forecasting would be the gap between $2 * 10^{-3}$ and $2 * 10^{-3}(1 - 10^{-3})$ i.e. a shift in score of only 0.1% (Benedetti [2010]), very large sample sizes are accordingly required to see an appreciable difference between forecasters. A similar phenomenon can be seen with extremely common events. Because the Brier score does not always sufficiently discriminate between small forecast changes under common sample sizes, this can create issues for very rare events such as these, a common phenomena in SEJ (Wilks [2010]).

The methods outlined so far consider the modelling predictions as only outputting a point prediction. Whilst this defines the most common number that DM's use directly, they typically setup SEJ studies as they are interested in the uncertainty present in the outcomes of the target variables of interest. As such, it is also important to consider scoring rules that assess the uncertainty that different modelling choices provide.

One such method is a reapplication of the Classical model itself (Eggstaff

et al. [2014], Colson and Cooke [2017]). Here, there is no change to the mathematics behind the approach. Conceptually the Classical model is just applied with each of the different modelling types under consideration considered an "expert". For any given subset within the testable data set, the predictions the models have made are set to be the seed variables. The performance weighted measures (statistical accuracy and information) are calculated for each model across all of the forecasted variables considered. The performance weighting provided for a model is then adopted as the "score" for this model in this subset. This process is then repeated for all of the subsets within the testable set. When there are 10 seed variables this therefore requires an application of the Classical model 1022 times.

Typically, aggregate statistical accuracy and information values are calculated for the total study by taking the mean or median of the values for each subset. Geometric and arithmetic means are often both calculated with the arithmetic mean being the value most commonly utilised (Eggstaff et al. [2014], Colson and Cooke [2017]).

This method of cross-validation is the most comprehensively utilised today although it's application has typically extended only to evangelists of the Classical model itself. This is partly because, based on Roger Cooke's original ethos regarding SEJ (Cooke [1991]), this is the group that has invested the most in championing validation on SEJ aggregation methodologies. It is likely partly also because the sceptics of the Classical method feel that, utilising one of the tested models as your method for validation creates potential bias (Bolger and Rowe [2015]).

Regardless of the rationale behind challenges for any particular validation method, there is need for more wide spread agreement and use of a single validation approach within the SEJ community. It is in DM's and SEJ practitioner's interests for SEJ to be given stronger support. This can only be done if comprehensive validation of methodologies is available.

Notably absent from any of the validation literature are Bayesian models. As has been highlighted already, significant investment has been placed in designing new Bayesian approaches to SEJ, but few have been utilised outside of a small domain or case study.

Bayesian models, given their inherently more complicated nature, typically include more nuance in application. Capturing elements such as DM's priors and inter-expert correlations, priors on model hyperparameters, etc. all improve model performance and give DMs and practitioners more flexibility whilst conducting a study. This flexibility however creates challenges when considering application on the scale required to conduct validation exercises as outlined. Cross-validation of

a single study utilising a Bayesian methodology may require defining hundreds of potential priors. The overhead to set each of these independently makes this infeasible.

The answer here however, is not to bypass cross-validation for Bayesian models but to endeavour to find ways to make it tractable. One option, is to conduct cross-validation for Bayesian models where standard and relatively uninformative priors are used consistently across subsets, and potentially studies, in lieu of uniquely defined priors for each application. This is appealing for two reasons.

Firstly it sets a lowest bar for the relative performance of that Bayesian methodology vs. other techniques. If the methodology can perform well with uninformative priors then logically, well calibrated informative priors will only improve the model performance.

The other benefit of this approach is that one of the historical challenges with Bayesian methods is that complex prior definition made them fiddly to implement (Wiper and French [1995]). If a model can perform well with uninformative priors then this gives a DM or analyst the opportunity to only include further complexity when desired, rather than by necessity.

This thesis aims to justify the application of Bayesian SEJ approaches by creating some empirical data on their application. Chapter 8 outlines the first significant large scale cross-validation of Bayesian methods for SEJ utilising the uninformative prior approach outlined above. This hopefully will be the start of many more research efforts to generate the empirical data the Bayesian SEJ community need.

4.5 The random expert hypothesis

A recent paper, Cooke et al. [2020], has highlighted a new mechanism for providing validation of expert judgement studies. This is known as the random expert hypothesis (REH). The REH proposes an experiment on the elicited data from experts to identify whether discrepancies in expert performance can be explained by random stressors that emerged as part of the elicitation exercise. If the REH is true then random reallocation of assessments between expert members should statistically not affect the groups performance.

The REH experiment is conducted by selecting a random permutation of experts' elicited responses for each seed variable and assigning these item wise to each expert. i.e. the first *random expert* takes the first set of responses in the permuted list for each variable. The second *random expert* takes the next and so

on until the last expert has the remaining permuted responses. The performance of each of these new random experts is tested utilising the previously outlined statistical accuracy, information and combination scores.

The random scrambling is repeated many times in order to generate a distribution for each of the performance metrics. If expert performance were only due to random stressors in the elicitation process then the variation seen should be described by this distribution. In practice this is not the case. In 20 of the 49 tested studies the average statistical accuracy of the original expert group sat above the 95th percentile of the same metric for the random expert distribution. The probability of this happening is $6.6 * 10^{-14}$. This is a powerful statement in a much stronger test than the cross-validation outlined earlier.

It is important to note here that this is a slightly different test to preceding examples. The REH does not test the aggregation methodology but tests the expert scores themselves. It is essentially a validation that seed variables are a useful tool, i.e. they provide tangible information on expert performance that cannot be explained by random fluctuations. It does make an important comment about non performance weighted aggregation methods however.

Any aggregation method which does not consider expert performance in its formulation, e.g. quantile aggregation, equal weighted linear opinion pools etc, will be unaffected by this random permutation. Therefore, for any of these methods to be considered at least as good as a performance adjusted method implies that the random expert hypothesis is true. Given that the random expert hypothesis can be statistically rejected (a full outline for why this is the case across the 49 studies tested is outlined in Cooke et al. [2020]), therefore it is claimed that it follows directly that methods which consider expert performance must outperform these other approaches.

It is important to highlight here that the REH only considers expert performance on seed variables, it does not technically make a comment on the direct link between seed variable and target variable performance. The test is actually specifically highlighting that there is very low probability that on average experts in an SEJ study are all "equally good" or "equally bad" at assessing seed variable questions (in some specific cases this may well be the case but it must be the minority).

This is a important validation on the value of seed variables and the fact that there is identifiable differences in expert performance on SEJ panels. It is not however, critically, a comment on the performance weighting approach of Cooke's Classical model itself. It is a statement about the experts. It demonstrates that a top performing mathematical aggregation method must consider the performance of the

experts within the group, however does not make a case for how this must be done. The results of this study can be seen as a boon for the Bayesian model outlined in this thesis which does consider expert performance as part of the calibration component of the model.

As the REH analyses the experts rather than the aggregation model, when assessing the Bayesian approach compared to Cooke's Classical model the cross-validation approach highlighted in the previous segment shall be utilised.

Chapter 5

An extended Bayesian framework for SEJ combination

5.1 Model outline

The Bayesian framework outlined within this thesis simplifies some of the inherent complexity within expert judgement mathematical aggregation by breaking the post elicitation processing into four distinct steps:

- Expert clustering - *a process to group experts based on consistency in responses.*
- Distribution fitting - *a mechanism to convert from experts' elicited responses into parametrised distributions.*
- Recalibration - *a method to assess potential systemic over or under confidence in expert's elicited bounds and to adjust accordingly.*
- Weighted Aggregation - *a model to bring together the disparate views of the experts into a single* DM *posterior.*

The method as outlined here is applied to judgements in the form used by the Classical model, including a target variable set, $\mathbf{X}$, and a calibration data set (the seed variables $\mathbf{Y}$). All elicitations are made against a standard set of quantiles (typically 3 - 0.05, 0.5, 0.95 or 5 - 0.05, 0.25, 0.5, 0.75, 0.95).

The key behind the model structure is that whilst the method is applied here to data of this form, at the core the model is very generic. As outlined previously, the applicability of some of the above steps may vary by context. Recalibration may be inappropriate in certain settings or data may be elicited through other methods which results in elicited parametrised distributions. In these circumstances

a DM may want to avoid the recalibration process or the distribution fitting step, respectively. The model is structured such that any of the above steps can be added or removed depending on the use case. This is one of the advantages of building a model with this modular design.

Section.5.7 demonstrates how these modules are mathematically linked and therefore how the process of technically adding steps works. We start by outlining the steps individually.

5.2 Expert clustering - the calculation of homogeneity groups

One of the risks leading to overconfidence in a final posterior comes from the shared knowledge or common professional backgrounds that experts may have which drives correlation. As outlined before, finding such an underlying correlation and correcting for it is often a challenge for Bayesian models. One approach to bypass the issue of directly calculating complex correlation matrices would be to identify the sources of the underlying similarity in estimation and with this knowledge cluster experts into homogeneity groups in which all experts with similar historic knowledge are grouped together. As part of the aggregation exercise this knowledge could be utilised to reduce the risk of overconfidence (Albert et al. [2012], Billari et al. [2014]). One approach to forming these groups would be to attempt to elicit information about potential sources of common knowledge, in addition to the quantiles, from the experts. This approach is appealingly simple and would require only a procedural update. In practice, however, this elicitation is likely to be challenging as sources of this correlation may be opaque, even to the experts themselves. Thus algorithmic approaches, which attempt to infer these groupings, could be considered.

The framework employed, similar to Billari et al. [2014], utilises algorithmic clustering techniques to group and re-weight experts. Given the classical model data structure, there is a choice of data set to use for the clustering exercise, the target variables, the seed variables, or a combination thereof. Here seed variables are used. If there is underlying correlation between experts, driven by their shared knowledge, then this correlation should be apparent in their seed variable estimations. If there is no such link on the seed variables, then there is limited risk of overconfidence on the target variables. This is clearly true only if the seed variables are within the same domain as that of the target variables, i.e. shared knowledge of experts in rare genetic conditions within hamsters does not imply shared knowledge in the risk of a bolt breaking within a suspension bridge. Representativeness of seed variables

is similarly a core tenet underlying the use of these variables within the Classical model. This will be left aside for now however and note, that similar to Billari et al. [2014], the target variables could have been used in their place.

Given seed variable estimations, it is easy to apply any number of clustering algorithms to the seed variable space (in which each expert is a point) in order to generate the expert groupings. Either hierarchical clustering is recommended, due to its efficacy over sparse datasets (and easy comprehension by DMs) or mixture models, specifically, Dirichlet process mixture models (DPMMs). DPMMs are valuable due to their limited assumptions about the number of groupings *a priori* and their ability to integrate easily with the broader Bayesian framework (Billari et al. [2014]). Other clustering approaches such as K-means clustering (Bradley and Fayyad [1998], Wagstaff [2001], Kanungo et al. [2002]), are feasible, although have not been tested. Further work to consider impact of clustering choices could be considered.

Suppose that two experts exist within a single homogeneity group, i.e. come from a similar background and school of thought. Given the structure of a hierarchical model such as Fig.3.6, the parameters (γ_{eh}) which represent these experts beliefs about $\boldsymbol{X}$ are drawn from a single distribution $f(\cdot|\gamma_h)$. It is reasonable for the DM to suppose therefore, that the experts' beliefs about the seed variables $\mathbf{Y}$ are also drawn from similar distributions i.e. experts belong in the same homogeneity group for all elicited variables. If we consider a single expert e who belongs to one of the homogeneity groups h then this expert will be similarly grouped in the homogeneity sets for both the target and the seed variables. i.e.

$$e \in \mathbf{E} \;\; s.t. \;\; \{e \in h, h \in \mathbf{H}\}_{\mathbf{X}} \implies \{e \in h, h \in \mathbf{H}\}_{\mathbf{Y}} \tag{5.1}$$

Here $\{\cdot\}_{\mathbf{X}}$ refers to the set of homogeneity groups formed when considering the target variables, $\{\cdot\}_{\mathbf{Y}}$ is similarly defined for the seed variables.

Following from (5.1); given that $\{h \in \mathbf{H}\}_{Y_1} \equiv \{h \in \mathbf{H}\}_{Y_2} \equiv \cdots \equiv \{h \in \mathbf{H}\}_{Y_{|\mathbf{Y}|}}$, let us consider the $|\mathbf{Y}|$ dimensional space $\in \mathbb{R}^{|\mathbf{Y}|}$ formed by responses to the $|\mathbf{Y}|$ seed variables, which are linearly scaled to the unit interval. Scaling variables before the clustering is critical to ensure that clusters are not determined purely by experts' responses to a small number of variables that may have a larger magnitude than others. Performing this scaling using a linear transformation is only one option here, however, as the model is restricted to variables on a uniform domain (as stated earlier the whole model structure would need amendment for the consideration of logged variables) it is a reasonable approach. Other scaling techniques should be considered when deploying this clustering approach in practice, however, the necessity to move away from linear scaling will be context dependent and therefore not

in scope of this thesis. Each expert's response for the mid-quantile defines their position along the relevant axis, and therefore each expert is represented in this space by a single point $\mathbf{Y}_e$. It is reasonable to assume from here that experts from a single homogeneity group have responses clustered in some sense within this space. Thus, assigning experts to homogeneity groups simplifies to identifying clusters in the seed variable space and creating an index for each expert based on the cluster within which their seed variable responses sit.

Clustering is an exploratory data analysis technique and there is no single definition of what constitutes a cluster, nor how rigidly items must be allocated into these clusters. Given that experts cannot be in multiple homogeneity groups, each expert's responses must belong to exactly one cluster and therefore a *hard clustering* is needed. Furthermore, the set of homogeneity groups must be a covering of the experts, therefore we are ultimately looking for a *strict partition clustering* of the space.

One approach to defining the clusters is to do this through visual inspection. In low dimensional data sets, appropriate clustering can often be determined simply by looking at a plot of the elements in either the x,y plane or the x,y,z cube. Visual inspection is often not feasible in SEJ studies as there are often many more than three seed variables. This creates a high dimensional space which cannot be visualised easily.

To overcome these challenges we recommend algorithmic cluster determination followed by targeted visual inspection, for validation. The algorithmic approach ensures that the full dimensionality of the data is considered, and provides a mechanism which removes as much subjectivity as possible in the cluster definitions. Visual inspection provides an opportunity for the rationale behind the clusters proposed by the algorithmic approach to be made clear. This can enhance buy-in and allow for adjustment if there is staunch disagreement or further knowledge to be embedded.

There are a multitude of algorithms that could be considered to estimate the cluster structure (and therefore the underlying homogeneity groupings). Agglomerative hierarchical clustering algorithms are an appealing method to use as they are easy to conduct pre-analysis, easy to understand and provide a nice visual way for a DM to review the clusterings that will ultimately impact the model. They have also been shown to work well over sparse data sets.

The hierarchical clustering process is an iterative algorithm which initially puts each element into its own cluster and then merges clusters together based on their distance apart and a linkage criteria. This process is repeated until all of

the elements are merged into a single group. Each step in the process creates a different potential set of clusters. In order to arrive at our final strict partition a cut of the clusters which exist at one step in the process is taken. This cut can be created either visually by looking at a *dendogram* of the hierarchical clustering and considering the distance shift at each merge or by conducting the clustering using a library which considers many potential metrics and determines the cut for you, e.g. NbClust in R.

It is important to note that dendograms are a summary representation of the underlying data set and the associated dissimilarity/distance matrix (the table that shows the distances between pairs of objects in the raw data). As a result, different dendograms (and by extension clustering approaches) for the same data will preserve the dissimilarity matrix to different extents. Goodness of fit measures are used to assess the degree to which the dendogram faithfully preserves the original dissimilarity matrix. Common metrics utilised are the cophenetic correlation coefficient (Rohlf and Fisher [1968]) and the delta metric (Mather [1976]). The topic is still studied widely and other metrics have also recently been proposed, (Mérigot et al. [2010]).

Note. The maximum number of feasible clusters is simply the number of experts present within the study. If each expert sits within their own cluster, and therefore their own homogeneity group, the middle step in the hierarchical aggregation model becomes redundant. In this instance the model can be thought of as only having two levels, an expert level and a total global level. For utilising the three levels in the hierarchical aggregation process, therefore, we need to consider groupings whereby at least two experts are combined, i.e. where the maximum number of clusters is $|\boldsymbol{E}| - 1$.

The disadvantage of a hierarchical clustering approach is that it is not possible to integrate it fully into the Bayesian model (clustering would need to be processed first and then included). In this way, we will have a two-stage method. This will result in the seed variable data being used twice and completely independently, once for clustering and once for calibration, which is unappealing.

This two-stage method therefore gives results which are an approximation to a fully Bayesian method. Using mixture models it is possible to take a further step closer to a fully Bayesian model by integrating the clustering directly into the MCMC. With sufficient data, the number of clusters can be inferred by extending to Dirichlet process mixture models (DPMM). SEJ studies are often not large enough to necessitate this method and simpler hierarchical clustering will suffice. For completeness, the fully Bayesian approach is outlined in the Results section and, for

one study, compared to the hierarchical approach, to determine how reasonable an approximation the two-stage method is.

As clustering is an exploratory process, when used post an algorithmic determination, visual inspection gives the opportunity for validation (and if necessary tweaking) of the clusters defined. It can help both in ensuring that recommended clusters are appropriate and in getting buy-in from DMs and other stakeholders to the choices made. As the seed variable space is high dimensional some processing of the data is required in order to create visuals which can be analysed easily. We recommend running a principal component analysis (PCA) over the data set to reduce dimensionality.

Principal component analysis essentially solves an eigenvalue and eigenvector problem to change the coordinate system by defining new uncorrelated variables. In doing this the originally high dimensional space can be described as a space with a small number of meaningful dimensions, known as principal components. Each principal component captures a certain percentage of the variance between elements that existed in the original description. When applied to the SEJ data, the first few principal components can be visualised pairwise in two dimensions and the rationale for the clusterings created by the algorithms easily spotted. A scree plot which highlights the cumulative variance of the principal components can help build confidence that by visualising only this small segment of the total space a significant portion of the variance is being explained. In most expert judgment studies there will be a reasonably large number of dimensions relative to the number of experts and therefore significant dimension reduction should be visible. In the rare case that this does not happen and each PCA component only outlines a small portion of the variance, then this should correspond to those occasions where no homogeneity groups are identified by the clustering. In these instances the model should run with two levels an expert level and a global level, the mid homogeneity group section should be removed.

Applying a hierarchical clustering over the experts in the way described is not merely a mathematical exercise, it is implying that expert knowledge is inherently hierarchically structured. These types of clustering technique are effective only really when the data itself is hierarchical. Hierarchies of knowledge are constantly discussed, albeit not always explicitly, in the scientific literature. There are many obvious hierarchies, for example work is often described and published in a hierarchy starting with an overall field, such as Maths and Statistics which is then broken down into smaller and smaller domains, Mathematics $\supset$ Number Theory $\supset$ Algabreic Number Theory $\supset$ Local Class Field Theory etc... Although in practice many of

the bounds of these hierarchical elements have been drawn arbitrarily over time and are often blurry. Similarly there are cross-discipline hierarchies in evidence such as the behavioural vs. mathematical aggregation schools of thought discussed earlier. The theory behind the hierarchical expert clustering approach is that for any given domain, we assume that there exists some theoretical complete set of existing knowledge (in principle there is a further hierarchical step above this which separates existing knowledge from potential knowledge, that we will ignore for simplicity). The hypothesis is that experts are then exposed to elements of this knowledge in hierarchical ways, defined by the school of thought they are from, the specific field of their study, the domain they are in etc. The exact structure of this hierarchy is likely to be very complex, and impossible to infer at a general level, in some fields the school of thought may be dominant in others the location/language knowledge is published in may be more critical. It is this hierarchy that actually determines the important interdependence between the experts that we are trying to identify through the clustering exercise.

Once the proposed clusters are defined, this uniquely determines the homogeneity groups, $\mathbf{H}$, used to determine the index h assigned to each expert in Fig.3.6. The calibration and aggregation model components will now be outlined in their generic versions before highlighting the particular distribution fitting that shall be used for the remainder of the thesis. Whilst in model deployment the choice of distribution to use happens before the calibration and aggregation components, understanding the structure of these is critical to the rationale for the distribution choice.

5.3 A model for calibration

Bayesian models typically consider the topic of recalibration differently to other approaches. In the Bayesian model, as probability is subjective and thus a property of the observer (typically the DM) of the system. It appears reasonable therefore, for any such observer to consider all the information at hand in forming their final posterior distribution. An example of such information may be any bias which the experts have exhibited in historic judgements. Many potential drivers of bias, such as anchoring (Kahneman et al. [1982], Kahneman [2011]), can be minimised through elicitation procedure (Cooke and Goossens [2000]). Others, such as consistent over/under confidence, are often still visible (Burgman [2016]). Thus if expert e, from a pool of experts, has historically been systematically overconfident, a Bayesian DM may choose to broaden the tails in expert e's elicited judgement distributions,

before aggregating with other experts, in order to truly reflect the DM's belief of the uncertainty. One form of resistance to the study of calibration, and thereby recalibration, is that the links between cognitive processes and the mathematical models which underpin the concept of calibration are poorly understood and therefore we risk drawing incorrect conclusions, or making modifications in a way that do not scale (Keren [1991]). Keren [1991] provides an excellent review of many calibration studies, highlighting both conceptuals and methodological issues presented. A potential other challenge to this form of recalibration is that by adjusting the experts forecasts this creates an ownership problem (effectively the forecasts are no longer the experts' once you have adapted them, they belong to the analyst) and this produces an accountability issue accordingly. The Classical model does consider potential over/under confidence through the statistical accuracy score, but only uses these as a form of weighting selection, thereby bypassing the accountability issue. I would argue that the use of recalibration is context dependant. In expert judgement problems with a single decision maker it would potentially be remiss to ignore any such information about potential additional uncertainty. Regardless, the model outlined in this thesis is modular in design and recalibration could be included or excluded as appropriate given the context of the problem at hand. Significant overconfidence is apparent in many studies within the Delft database thus analysis within this text has included recalibration.

Understanding and mitigating overconfidence in experts' judgments is studied across disciplines including psychology, economics, statistics and engineering, (Lin and Bier [2008], Angner [2006], McKenzie et al. [2008], Lambert [2012], Kaustia [2012]). Consistent with ideas from Cox [1958], Morris [1974] and Sasirekha and Baby [2013], Clemen and Lichtendahl [2002] developed a model of overconfidence using past data to estimate, what the authors term, "inflation factors" for assessed distributions post hoc. Bayesian hierarchical models are used, allowing experts to be calibrated individually whilst simultaneously capturing inter-expert calibration effects. Before outlining the more complex hierarchical elements of the model however, it is helpful to outline how the inflation factors for a single expert are calculated.

Let us suppose, as per prior notation, a DM has reached out to a group of experts ($\mathbf{E}$) in order to help assess uncertainty for an unknown quantity X. Let us assume further he or she has reason to believe expert $e \in \mathbf{E}$ may be prone to some form of consistent bias which the DM wishes to remove before updating their own belief accordingly. Finally, we assume the DM has asked e to assess three quantiles, denoted by L_e, M_e and U_e, (e.g. 0.05, 0.50 and 0.95), corresponding to lower, middle

and upper estimates respectively. We will outline later the impact of other choices here. The goal of the DM is to be able to transform e's responses on the tail quantiles into their unbiased counterparts L_e^* and U_e^*.

Remark: For a three quantile model, it is possible to infer inflation factors for the spread of the distribution (i.e. calculate the unbiased values L_e^* and U_e^*) or to create a factor for assessing bias on the location parameter, $M_e^* = \beta_e M_e$, but not both. To attempt to define all three simultaneously, given only three elicited quantiles, would lead to an overspecified model, almost completely defined by the choice of priors. The original model outlined by Clemen and Lichtendahl [2002] attempted to infer all three parameters and so we use a slightly different structure. Empirical evidence (Lichtenstein et al. [1982]) suggests expert overconfidence is ubiquitous throughout elicited expert judgments, there is no evidence to suggest that identifiable bias occurs as commonly as overconfidence. Miscalibration with respect to spread is therefore typically the first calibration metric to assess. In a three parameter model we therefore define the median estimate to be its own unbiased counterpart, i.e. $M_e^* = M_e$ and attempt to infer L_e^* and U_e^*. Later an extension to the parameterisation to a situation with five elicited quantiles whereby β_e can be inferred is demonstrated.

Rather than calculating inflation factors directly on the elicited values the bias in the spread is calculated relative to the distance from M_e^*. The theory here is that there exists multiplicative parameters α_{le} and α_{ue} such that $\alpha_{le}(M_e^* - L_e)$ and $\alpha_{ue}(U_e - M_e^*)$ are unbiased. α_{le} and α_{ue} are therefore scale parameters whereby a value, for either, strictly greater than 1 suggests that the expert is overconfident. If either factor is strictly less than 1, this suggests that the expert is under-confident. L_e^* and U_e^* can thus be calculated by:

$$L_e^* = M_e^* - \alpha_{le}(M_e^* - L_e) = (1 - \alpha_{le})M_e + \alpha_{le}L_e \tag{5.2}$$

$$U_e^* = M_e^* + \alpha_{ue}(U_e - M_e^*) = (1 - \alpha_{ue})M_e + \alpha_{ue}U_e \tag{5.3}$$

Having established the relationship between the elicited values and their unbiased counterparts, we need to fit a model, g_e, to approximate the unbiased expert distribution $p_e(x_X|L_e^*, M_e^*, U_e^*)$. The model Clemen and Lichtendahl propose fits two uniform components on the intervals $[L_e^*,\ M_e^*]$ and $[M_e^*,\ U_e^*]$ respectively with exponential tails above U_e^* and below L_e^*. However, this choice is arbitrary and later we will outline another approach. The assumption Clemen and Lichtendahl make is that final results should be largely invariant to these modelling assumptions.

Remark. The formulation of inflation factors in this way makes the assump-

tion that all of the training variables are on the same scale (i.e. if some variables are logarithmic in nature and others are not this would create a challenge in this approach). Wiper and French [1995] rescaled judgements through the DM *'s prior to avoid assumptions on the common scale.*

Thus, the task now becomes how to assess the unknown parameters α_{le} and α_{ue}. The core premise of calibration is that each of these variables is assumed to be constant for each expert (within the pool of seed and target variables). SEJ studies of this nature are typically one-off activities, and the seed variables are elicited in a single process alongside the target variables. If seed variables were captured longitudinally over time then experts would have the opportunity to learn and adjust and this assumption regarding constant bias will be incorrect.

From standard Bayesian theory, if $Y_1, ..., Y_{|\mathbf{Y}|} \in \mathbf{Y}$ are assumed to be random variables sufficiently similar to X, e has historically made judgements against $\mathbf{Y}$ and the DM holds the observed values $y_1, ..., y_{|\mathbf{Y}|}$, (which can be perceived to be exchangeable). Then the data set compromising $\{y_Y\}$ and $\{L_{Ye}, M_{Ye}, U_{Ye}\}$ for $Y \in \mathbf{Y}$ can be used to discover the posterior distributions for each of the unknown parameters.[1] The set $\mathbf{Y}$ is termed the set of *seed variables*. We can build a model utilising a Markov chain Monte Carlo (MCMC) method such that $\forall e \in \mathbf{E}, Y \in \mathbf{Y}$:

$$y_Y \sim g_e(\cdot|L_{Ye}, M_{Ye}, U_{Ye}, \alpha_{le}, \alpha_{ue}) \tag{5.4}$$

Where L_{Ye}, M_{Ye}, U_{Ye} denotes expert e's elicited quantile for Y and y_Y represents the true realisation of Y, $\forall Y \in \mathbf{Y}$. After a sufficient burn in period the model can outline posterior distributions for the hyperparameters α_{le} and α_{ue} for each expert. Thus, the DM has the scale by which he or she should adjust the expert's elicited opinions on X before updating their own belief.

This calibration approach could be calculated for each expert individually, however given potential common sources for bias across experts, expert to expert correlation should be assessed. Sources of bias common to multiple experts might include mutual experiences or identical literature reviewed. To establish potential correlation here, the model is extended hierachically to capture this behaviour. Let α_{le} be assumed to be a random draw from a gamma distribution:

$$\alpha_{le}|A_l, B_l \sim \Gamma(A_l + 1, B_l) \tag{5.5}$$

[1]The subscript Y here, and in all future formulas, is used to denote that these are quantiles elicited for the seed variable $Y \in \mathbf{Y}$. Similarly, from here onwards, the subscript X, denotes a variable relating to a target variable $X \in \mathbf{X}$ and the superscript $*$ denotes an unbiased value calculated post recalibration.

where hyperparameters A_l and B_l are defined by:

$$A_l \sim Pois(a_l) \quad and \quad B_l \sim Exp(b_l) \tag{5.6}$$

If we set a_l and b_l to 2; this results in a relatively diffuse positive prior, with mean near 1. This is the prior as outlined in the original paper, Clemen and Lichtendahl [2002]. The gamma distribution was chosen due to its strictly positive shape and the Poisson and exponential forms were selected in order to govern the behaviour of the gamma and ensure that it was diffuse and with a suitable mean. There is no empirical interpretation of the forms evident here, they were selected for their shapes. Nonetheless, this is a compelling prior to use as clearly the scale parameters must be greater than zero and we are starting from the premise that experts are likely to be calibrated.

An identical model can then be applied to α_{ue}. The complete parameterisation of this model will result in a set of hyperparameters (A_l, B_l, A_u, B_u) which capture the similarities in behaviour across experts. When there is internal structure within the set of experts, i.e. a subset of experts come from similar backgrounds or schools of thought, experts can be grouped together into what are known as homogeneity groups. Each group can then have its own set of these hyperparameters which infer group behaviour. Implementing this through a hierarchical model will result in the posterior distribution for these hyperparameters in addition to those of each experts' characteristics. The calculation of expert to expert correlations it could be argued is a significant advantage of the Bayesian approach over some of the classical SEJ models.

If experts were also subject to cognitive biases which drive grouping, in addition to their school of thought/background this would potentially create challenge for the approach outlined. Whilst there is no empirical evidence tackling this explicit question, it is important to ensure that if it were the case it would not invalidate a model of this kind. In the primary definition of this new model, only overconfidence bias is addressed. This mitigates some of the potential issues, but is not sufficient for bias management overall within an SEJ study. The modelling exercise is not the only mechanism to manage bias. Other biases should be addressed during the elicitation process and be embedded into the facilitation protocol. A graph of the full calibration model is outlined in Fig. 5.1.

If experts operate as coherent subjective Bayesians, certain forms of recalibration drive philosophical mathematical inconsistencies (Kadane and Fischhoff [2013]). The exact form of calibration we are employing is explicitly excluded from

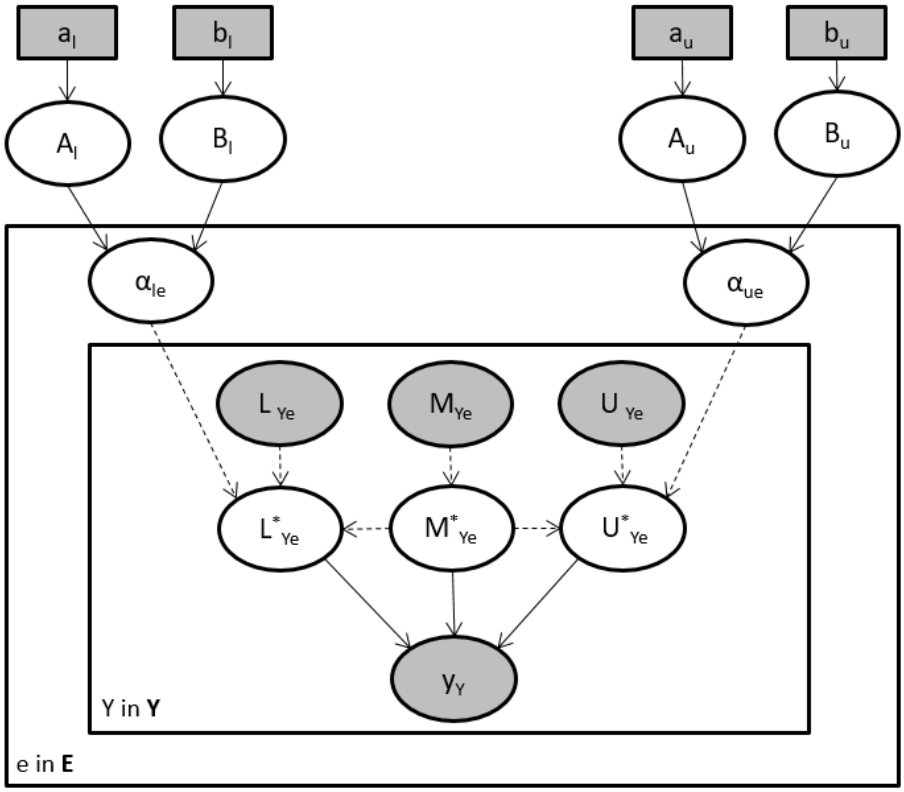


Figure 5.1: Hierarchical model for expert calibration utilising standard plate notation, grey ellipses represent known values, white ellipses unknown variables and smaller squares indicate fixed model parameters. True seed variable realisations y_Y are assumed to be random draws from a distribution determined by expert e's unbiased quantiles $L^*_{Ye}, M^*_{Ye}, U^*_{Ye}$. These unbiased quantiles are a logical function of the elicited quantiles, L_{Ye}, M_{Ye}, U_{Ye} and the inflation factors α_{le}, α_{ue} calculating the expert's overconfidence.

the mathematical analysis in Kadane and Fischhoff [2013]. In large data sets, such as Cooke's Delft database there is also evidence of incoherence among expert judgements, even on small numbers of elicited quantiles.

It is important to note that this formulation of expert recalibration is reliant on a certain level of mathematical coherence in expert's responses. Whilst it manages for lack of statistical accuracy it does not adjust for incoherence. If, for example, an expert had judgements in which the lower quantile is greater than the mid quantile or in the extreme case an upper quantile less than the lower quantile then this approach will not work. These forms of incoherence should be challenged and managed as part of the elicitation process rather than the modelling process.

It would be possible, if desired, to create a simpler parameterisation here. Rather than having different hyper-parameters for the upper and lower inflation factors A_l, B_l, A_u, B_u, we assume that α_{ue} and α_{le} are random draws from a single distribution determined by just two hyper-parameters, A and B. This would minimise the number of elements that need to be specified, but would put constraints on the relationship between the two inflation factors. Leaving these separate, allows for freedom in the model for these to be unrelated and potentially in opposing directions, i.e. over-confident in the lower quantile and under-confident in the upper quantile.

5.4 A model for aggregation

The traditional Bayesian approach treats the elicited information from experts as data and updates the DM's prior via Bayes formula (French [1985]). The aggregation model, taken from Albert et al. [2012], utilises a Supra-Bayesian parameter updating approach for combining indirect elicitation across multiple experts. Here we use the term indirect elicitation as rather than eliciting parameters from experts directly, experts' knowledge is elicited on more intuitive observables and the (hyper)parameters then inferred. Similar to Clemen and Lichtendahl's model, this method is generic and can be utilised with a multitude of parameterisations.

The aggregation model starts with the clustering of experts into homogeneity groups. Let us assume the experts are broken into a set of homogeneity groups $\mathbf{H}$, comprising of groups $h \in \mathbf{H}$, each of size $|h|$ s.t. $|\cup h| = |\mathbf{E}|, h \in \mathbf{H}$. The aim of the model will be to assess the variation both between and within these homogeneity classes. Homogeneity classes effectively use weighting to adjust for dependence between experts rather than by attempting to elicit some form of correlation structure. The ability to account for inter-expert dependence is important to ensure

uncertainty is not understated and is one of the advantages of Bayesian approaches (Wilson [2017], Wilson and Farrow [2018], Hartley and French [2018]). Selecting the right homogeneity classes is imperative. The guidance from Albert et al. is for experts within a class to be selected "corresponding to similar backgrounds or schools of thought." With no pre-determined acceptable threshold for correlation, the number of homogeneity groups is essentially arbitrary and it is recommended that you should have as many as possible. Creating too few groups will artificially reduce diversity. When assignment is not trivial, or there are multiple potential grouping choices, a protocol such as the clustering outlined previously for defining the groups can be useful.

Let γ_{eh} be a parameterisation, such that $g_e(\cdot|\gamma_{eh})$ represents the conceptual model about X held by expert e who is a member of homogeneity group h. Subscript eh is used from here on to denote this membership. The authors suggest the following hierarchical model to group experts:

$$\begin{aligned} \gamma_{eh} &\sim f(\cdot|\gamma_h, \rho_h) \quad \forall e \in \mathbf{E} \\ \gamma_h &\sim f(\cdot|\gamma, \rho) \quad \forall h \in \mathbf{H} \\ \gamma &\sim \pi_{DM} \end{aligned} \tag{5.7}$$

Here, experts within a single homogeneity group have parameters drawn from a consistent distribution $f(\cdot|\gamma_h, \rho_h)$. Each of the homogeneity groups has their parameters drawn from a single distribution $f(\cdot|\gamma, \rho)$. The ρ terms here represent dispersion parameters and the γ terms represent location parameters. *Note. the term γ, represents the output of the model, or more explicitly, the agreement of the experts given the decision maker's prior.* Fig. 5.2 outlines a graphical view for the aggregation model in standard plate notation. The functional form of f will be dependent on the choice of parameterisation g_e.

A simple parameterisation of (5.7) would be the two parameter model s.t. $\gamma = (\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\tau = 1/\sigma^2 > 0$. The complete model would thus be:

$$\begin{aligned} \mu_{eh}|\mu_h, \rho_h &\sim \mathcal{N}(\mu_h, \rho_h) & \frac{\tau_h}{\tau_{eh}}|\tau_h, \xi_h &\sim \Gamma(\xi_h, \xi_h) & & \\ \mu_h|\mu, \rho &\sim \mathcal{N}(\mu, \rho) & \frac{\tau}{\tau_h}|\tau, \xi &\sim \Gamma(\xi, \xi) & & \forall e \in \mathbf{E} \\ \mu &\sim \mathcal{N}(\mu_{DM}, \rho_0) & \tau^{-1} &\sim \tau_0^{-1}\Gamma(a, a) & & \forall h \in \mathbf{H} \end{aligned} \tag{5.8}$$

In this model, (μ, τ) represent the target consensus values; (μ_h, τ_h) are the homogeneity classes values and the (μ_{eh}, τ_{eh}) represent the parameterisation of the views of the individual experts, i.e. $g_e() = \mathcal{N}(\mu_{eh}, \tau_{eh})$. The expert level location

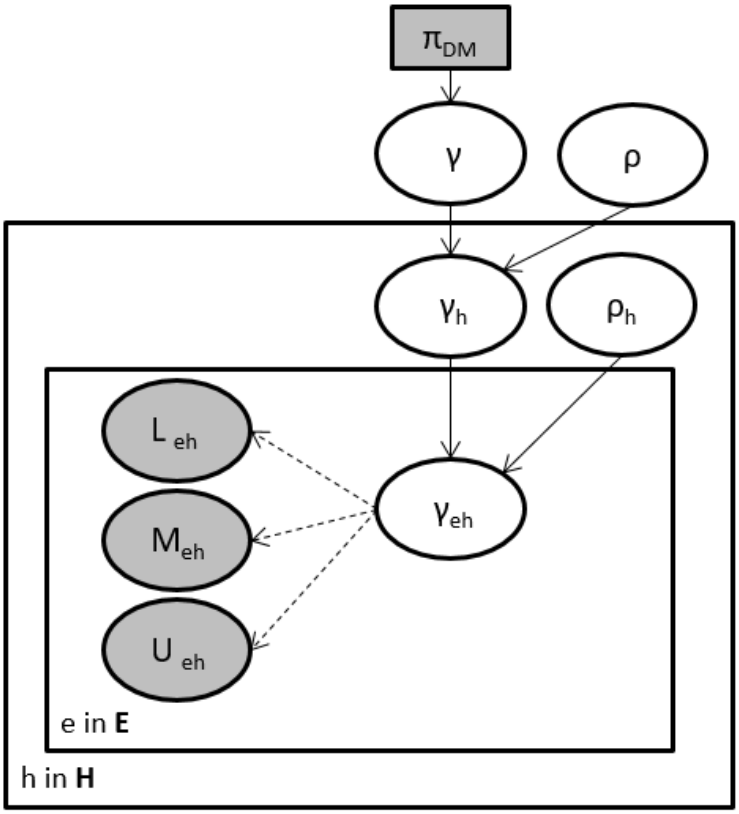


Figure 5.2: Hierarchical model for expert aggregation. Experts' opinions, which are elicited as quantiles, L_{eh}, M_{eh}, U_{eh} are related, via a logical determination, to a distribution parameterised by γ_{eh}. The γ_{eh} are considered samples from homogeneity groups γ_h. Each γ_h are in turn, drawn from an overarching distribution, with parameters γ, representing the DM's aggregate belief.

parameters, μ_{eh} are assumed to be random draws from a normal distribution with a location parameter defined by the homogeneity group within which the expert sits. The ρ values represent the dispersion of these distributions. The ratio between the expert level dispersion parameter and the homogeneity group dispersion parameter, will be strictly positive and is determined by a gamma distribution, with parameters ξ_h. The homogeneity group parameters are linked to the global parameters in a similar way. ξ and a represent the parameters of the distribution of the global to homogeneity group dispersion ratio and the dispersion prior respectively. When we build this out further, we will use a split normal and thus γ will be extended to include a third parameter.

With suitably diffuse priors selected by the decision maker (specifics outlined later) the full posterior of this model can be calculated utilising Gibbs sampling. The above has demonstrated the appropriate model for aggregation of a single target variable. Fig 5.2. (and equation 5.8) could trivially be extended to the whole set $\mathbf{X}$ (with a plate around the whole diagram and the appropriate subscripts), as each aggregation is independent.

5.5 Distribution fitting and model parameterisation

To parameterise our model correctly, we need first to define the generic distributions g_e and the corresponding unbiased parameters γ^*_{Xeh} $\forall X \in \mathbf{X}$ and $\forall e \in \mathbf{E}$. These choices, at an expert level, define the form of the random draw in the calibration model given in equation (5.4) and the first line in the aggregation step given in equation (5.7). Suppose, as proposed in the calibration section, that the experts have provided 3 quantiles (0.05,0.50 and 0.95) for each $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$; the parameterisation we choose should preserve all of the information that the experts have provided. The natural first choice would often be a Gaussian model whereby a distribution $\mathcal{N}(\mu^*_{Xeh}, \sigma^{*2}_{Xeh})$ is selected $\forall X \in \mathbf{X}$,(or similarly for $Y \in \mathbf{Y}$) where $\mu^*_{Xeh}, \sigma^{*2}_{Xeh}$ are chosen s.t. $\mu^*_{Xeh} = M^*_{Xeh} = M_{Xeh}$ and σ^{*2}_{Xeh} is defined to minimise the error between the probability density function (p.d.f) at 5% and 95% and the elicited values from the expert. In this context the Gaussian model would certainly simplify the computation required, however it would make a very strong assumption that experts' true beliefs are symmetric around the mid quantile (even when their elicited values are not). Often this is not the case and therefore utilising this parameterisation would immediately distort the elicited data.

A second choice for the parameterisation; as initially proposed by Clemen and Lichtendahl [2002] is to model the experts' beliefs with two uniform components

and exponential tails. We define the subscript Xeh to denote values for the variable $X \in \mathbf{X}$, from expert $e \in \mathbf{E}$ who is a member of homogeneity group $h \in \mathbf{H}$ and P_L, P_M, P_U, as before, denote the probabilities which were originally elicited against. In this way:

$$g_e(x|L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}) = \begin{cases} P_L \lambda_L e^{-\lambda_L (L^*_{Xeh} - x)} & if \quad x < L^*_{Xeh} \\ \frac{P_M - P_L}{M^*_{Xeh} - L^*_{Xeh}} & if \quad L^*_{Xeh} \leqslant x < M^*_{Xeh} \\ \frac{P_U - P_M}{U^*_{Xeh} - M^*_{Xeh}} & if \quad M^*_{Xeh} \leqslant x < U^*_{Xeh} \\ (1 - P_U) \lambda_U e^{-\lambda_U (x - U^*_{Xeh})} & if \quad x > U^*_{Xeh} \end{cases} \tag{5.9}$$

Here parameters λ_L and λ_U are given by:

$$\lambda_L = (\frac{P_M - P_L}{M^*_{Xeh} - L^*_{Xeh}}) \frac{1}{P_L} \tag{5.10}$$

and

$$\lambda_U = (\frac{P_U - P_M}{U^*_{Xeh} - M^*_{Xeh}}) \frac{1}{1 - P_U} \tag{5.11}$$

This approach has a distinct advantage over the basic Gaussian parameterisation as the distribution will exactly fit the quantiles given by the expert and thus there is no loss of data. However, the uniform component puts very little mass near the central quantile suggesting that the expert gives us very little information except the range of probable outcomes (0.05 - 0.95 quantiles).

The approach that we have taken is to utilise the natural shape of the Gaussian, in which we suggest that in practice experts have a strong belief about the mid-quantile with diminishing probabilities from here, without the associated loss of information. In this way we will model utilising a split normal:

$$g_e(x|L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}) \sim \begin{cases} \frac{1}{\sigma^*_{Xleh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x - M^*_{Xeh}}{\sigma^*_{Xleh}})^2} & if x < M^*_{Xeh} \\ \frac{1}{\sigma^*_{Xueh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x - M^*_{Xeh}}{\sigma^*_{Xueh}})^2} & if x \geqslant M^*_{Xeh} \end{cases} \tag{5.12}$$

where, the unbiased standard deviations σ^*_{Xleh} and σ^*_{Xueh} are calculated by:

$$\sigma^*_{Xleh} = \frac{M^*_{Xeh} - L^*_{Xeh}}{\delta_1} \quad and \quad \sigma^*_{Xueh} = \frac{U^*_{Xeh} - M^*_{Xeh}}{\delta_2} \tag{5.13}$$

Here δ_i represents the number of standard deviations between the elicited quantiles. Clearly if the probabilities for L and U are symmetric around M then $\delta_1 = \delta_2$. In the case 0.05, 0.5 and 0.95 then $\delta_1 = \delta_2 \cong 1.64$. τ^*_{Xleh} and τ^*_{Xueh} follow

directly from these assignments and are then hierarchically calculated as per (5.8). The final parameterisations are given by $\gamma^*_{Xeh} = (M^*_{Xeh}, \tau^*_{Xleh}, \tau^*_{Xueh})$. With this model the decision maker location prior will be M_{DM} rather than μ_{DM}.

The formulation of the split normal in this way will not result in a fully continuous distribution as at the mid-quantile there is a step. It would be trivial to adjust for this simply by factoring each half of the distribution, however, the result of this would be a shift in the median point, which is an unappealing result given the way we have defined calibration. Given that it does not largely affect the complexity of the modelling we leave this point of slight discontinuity.

Remark. this model is only one of many approaches which could be taken, it would be interesting, although not covered within this thesis, to assess the impact of non-Gaussian parameterisations on the final output.

5.6 Parameterisation challenges within BUGS/JAGS

The programming language JAGS (an R implementation of the commonly used Bayesian Gibbs sampling software BUGS) supports a broad set of sampling distributions, but unfortunately does not support the parameterisation we are using here. To this extent it is necessary to implement a modelling trick in order to get our parameterisation to work (McGill [1]). There are two potential methods for coding an arbitrary distribution within BUGS a "*zeros trick*" and a "*ones trick*." Let us assume that we are looking for a sampling distribution in which an observation x contributes to a likelihood function that is defined by L.

For the *"zeros trick"*, we note that a Poisson (ϕ) observation of zero has likelihood $\exp^{-\phi}$. Thus, if we make our observed data a set of zeros and define $\phi_x := -log(L(x))$ we will obtain the correct likelihood contribution. The only complicating factor is that the ϕ_x needs to be positive as it is a Poisson mean so it is necessary to add a suitable constant in order to ensure that it is positive.

The *"ones trick"* relies on a similar perspective pivot. Here we assume that the observed data is a set of ones generated by Bernoulli trials with probabilities p_x. If we can make each of p_x's proportional to $L(x)$ then we once again obtain the necessary likelihood function. Here, we need to multiply by a scaling constant to ensure that each p_x is strictly less than 1.

Within the code utilised for this thesis the "zeros trick" was predominantly used, however some sensitivity analysis utilising the "ones trick" was also performed to ensure that this did not create any different convergence behaviour.

5.7 Component integration

5.7.1 Full method

With the four components of the model identified in isolation, it is possible to outline the full model which connects these together. Building on the aggregation/calibration elements previously outlined, we can create an algorithmic approach which defines homogeneity groups, calibrates and finally aggregates.

Posterior distributions for target variables are created utilising the following descriptive process:

Data going into the model:

- A set of experts $\mathbf{E}$, a set of seed variables $\mathbf{Y}$ and a set of target variables $\mathbf{X}$.
- A set of probabilities associated with quantiles that were elicited from each expert for each variable - P_L, P_M, P_U.
- Elicited quantiles from each expert for the target variables - $L_{Xeh}, M_{Xeh}, U_{Xeh}$ $\forall e \in \mathbf{E}, X \in \mathbf{X}$.
- Elicited quantiles from each expert for the seed variables - $L_{Yeh}, M_{Yeh}, U_{Yeh}$ $\forall e \in \mathbf{E}, Y \in \mathbf{Y}$.
- A parameterised distribution to be fit to the elicited data for each expert - g_e $\forall e \in \mathbf{E}$ with cumulative distribution function G_e
- The decision maker's prior for each target variable - π_{DM_X}.

Step one - Homogeneity group calculation

For each elicited seed variable mid-quantile, M_{Yeh}, rescale onto the unit interval. Term the new rescaled value rM_{Yeh}.

$$rM_{Yeh} := \frac{M_{Yeh} - min(\{M_{Yeh} : e \in \mathbf{E}\})}{max(\{M_{Yeh} : e \in \mathbf{E}\}) - min(\{M_{Yeh} : e \in \mathbf{E}\})} \tag{5.14}$$

Define $\mathbf{Y}_e$ to be the $|\mathbf{Y}|$ dimensional tuple:

$$\mathbf{Y}_e = (rM_{1eh}, rM_{2eh}, ..., rM_{|\mathbf{Y}|eh}) \tag{5.15}$$

Run an agglomerative hierarchical clustering process. Set each item $\mathbf{Y}_e$ to be its own cluster $\mathbf{C}_e$. Create a dendogram by merging clusters one at a time based on the Euclidean distance between them in $\mathbb{R}^{|\mathbf{Y}|}$.

i.e. Merge the two clusters $\mathbf{C}_i$ and $\mathbf{C}_j$ that minimise D($\mathbf{C}_i$,$\mathbf{C}_j$). Where

$$D(\mathbf{C}_i, \mathbf{C}_j) = max\ d(\mathbf{Y}_i, \mathbf{Y}_j) \quad \mathbf{Y}_i \in \mathbf{C}_i, \mathbf{Y}_j \in \mathbf{C}_j, \tag{5.16}$$

and

$$d(\mathbf{Y}_i, \mathbf{Y}_j) = \sqrt{\sum_{k \in 1:|\mathbf{Y}|} (\mathbf{Y}_i(k) - \mathbf{Y}_j(k))^2} \tag{5.17}$$

This defines an agglomerative process with a Euclidean distance metric and a complete linkage criterion (Sasirekha and Baby [2013]). These are standard metrics to use for this form of clustering but many others are available (Kumar et al. [2014], Madhulatha [2012]). Testing different metrics on the CWD data set from Cooke's database, Cooke and Goossens [2008], showed that the defined clusters were invariant to the choice of distance metric used, assuming that continuous distance measures were considered. Tested metrics which resulted in identical groupings were the euclidean distance, maximum distance, manhattan distance, canberra distance and the more generalised minkowski distance. The binary or Hamming distance metric (Norouzi et al. [2012]) which is a distance metric for categorical data and commonly used for error detection unsurprisingly results in different groupings. As expert judgments considered in this thesis are continuous in nature, this would be an unnatural distance measure to choose.

This merging process is repeated using the same criteria until all elements form a single cluster.

Final homogeneity groupings, $\mathbf{H}$, are then defined by selecting a cut of this dendogram which can be done either manually based on visual inspection or utilising a dynamic tree cutting approach such as in the R package NbClust.

The cluster groupings that sit along the cut are assignments of experts to homogeneity groups. Suppose along this cut ten experts were clustered in the following three groups. Group 1 - experts {1,3,5,7,9}, Group 2 - experts{2,4,6}, Group 3 - experts{8,10}. $\mathbf{H}$ would then be the array {1,2,1,2,1,2,1,3,1,3}.

Validate the homogeneity group choices by running a principal component analysis (PCA) over the seed variable space. Visualise the first two or three principal components pairwise and consider a scree plot of the PCA to understand the level of variance captured within these components. Determine whether there is agreement with the choices made algorithmically and then finalise homogeneity group assignments.

If the clustering exercise does not determine that there are any meaningful expert groupings then these should not be forced. In these cases the PCA will

either show no meaningful groupings in the first few components or will have been unable to meaningfully reduce the number of dimensions (each component will only describe a small amount of the variation from the original data structure). The lack of clusterings should be readily apparent. The generic model for calibration and aggregation should then be updated to remove the homogeneity group layer and reduced to a two layer model with an expert layer and a global layer. The model can then be run as before.

Step two - calibration and aggregation

Calibration

For each Y in $\mathbf{Y}$ and e in $\mathbf{E}$, assume the true realisation of Y (y_Y) are random draws from a distribution of structure g_e defined by the unbiased quantile estimates of the expert e.

$$y_Y \sim g_e(\cdot|L^*_{Yeh}, U^*_{Yeh}, M^*_{Yeh}) \quad e \in \mathbf{E}, Y \in \mathbf{Y} \tag{5.18}$$

where the unbiased quantile estimates are defined by:

$$\begin{aligned} L^*_{Yeh} &:= (1-\alpha_{le})M_{Ye} + \alpha_{le}L_{Ye} & e &\in \mathbf{E} \\ U^*_{Yeh} &:= (1-\alpha_{ue})M_{Ye} + \alpha_{ue}U_{Ye} & h &= \mathbf{H}(e) \\ M^*_{Yeh} &:= M_{Ye} & X &\in \mathbf{X} \end{aligned} \tag{5.19}$$

and the inflation factors for each expert are random draws from a distribution which is consistent across experts within a single homogeneity group:

$$\begin{aligned} \alpha_{le}|A_{lh}, B_{lh} &\sim \Gamma(A_{lh}+1, B_{lh}) & e &\in \mathbf{E} \\ \alpha_{ue}|A_{uh}, B_{uh} &\sim \Gamma(A_{uh}+1, B_{uh}) & h &= \mathbf{H}(e) \end{aligned} \tag{5.20}$$

where A_{lh}, A_{uh}, B_{lh} and B_{uh} are defined by:

$$\begin{aligned} A_{lh} \sim Pois(a_l) \;\; and \;\; B_{lh} \sim Exp(b_l) \\ A_{uh} \sim Pois(a_u) \;\; and \;\; B_{uh} \sim Exp(b_u) \end{aligned} \tag{5.21}$$

Hyperparameters a_l, a_u, b_l and b_u are consistent across all experts and homogeneity groups.

Aggregation

Assume that the elicited quantile for each expert target variable pair is a

function of the underlying unbiased quantiles and the inflation factors inferred.

$$\begin{aligned} L_{Xeh} &= (L^*_{Xeh} - (1-\alpha_{le})M_{Xe})/\alpha_{le} & e \in \mathbf{E} \\ U_{Xeh} &= (U^*_{Xe} - (1-\alpha_{ue})M_{Xe})/\alpha_{ue} & h = \mathbf{H}(e) \\ M_{Xeh} &= M^*_{Xe} & X \in \mathbf{X} \end{aligned} \tag{5.22}$$

Where the unbiased parameters $L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}$ are those such that:

$$G_e(L^*_{Xeh}|\gamma^*_{Xeh}) = P_L \tag{5.23}$$

$$G_e(M^*_{Xeh}|\gamma^*_{Xeh}) = P_M \tag{5.24}$$

$$G_e(U^*_{Xeh}|\gamma^*_{Xeh}) = P_U \tag{5.25}$$

Where each experts' unbiased parameterised values γ^*_{Xeh}, for a given target variable $X \in \mathbf{X}$ and expert e are random draws from a distribution, f, defined by the homogeneity group $h \in \mathbf{H}$ within which e sits. The appropriate functional form of f is determined by the functional form of g_e. Homogeneity group parameters are random draws from a global distribution, which have the decision maker's prior.

$$\begin{aligned} \gamma^*_{Xeh} &\sim f(\cdot|\gamma_{Xh}, \rho_{Xh}) \quad \forall e \in \mathbf{E} \\ \gamma_{Xh} &\sim f(\cdot|\gamma_X, \rho_X) \quad \forall h \in \mathbf{H} \\ \gamma_X &\sim \pi_{DM_X} \end{aligned} \tag{5.26}$$

The parameters γ_X are then used to infer the target posterior given by $g_{DM}(\cdot|\gamma_X)$. A graph of the combined calibration/aggregation model is shown in Fig 5.3.

Please note: In practice, when encoding in a language such as BUGS*, the logical determination in equations (5.22)-(5.25) is embedded within the first line of (5.26). Thus the data is encoded as a random draw. To this extent, this is modelled as* $L_{Xeh} \sim f(\cdot|\gamma_{Xh}, \rho_{Xh}, \alpha_{le})$ *and the functional forms of* g_e *and equation (5.22) determine the structure of this draw. Similar for* U_{Xeh} *and* M_{Xeh}*.*

5.7.2 DAG for connected calibration and aggregation models

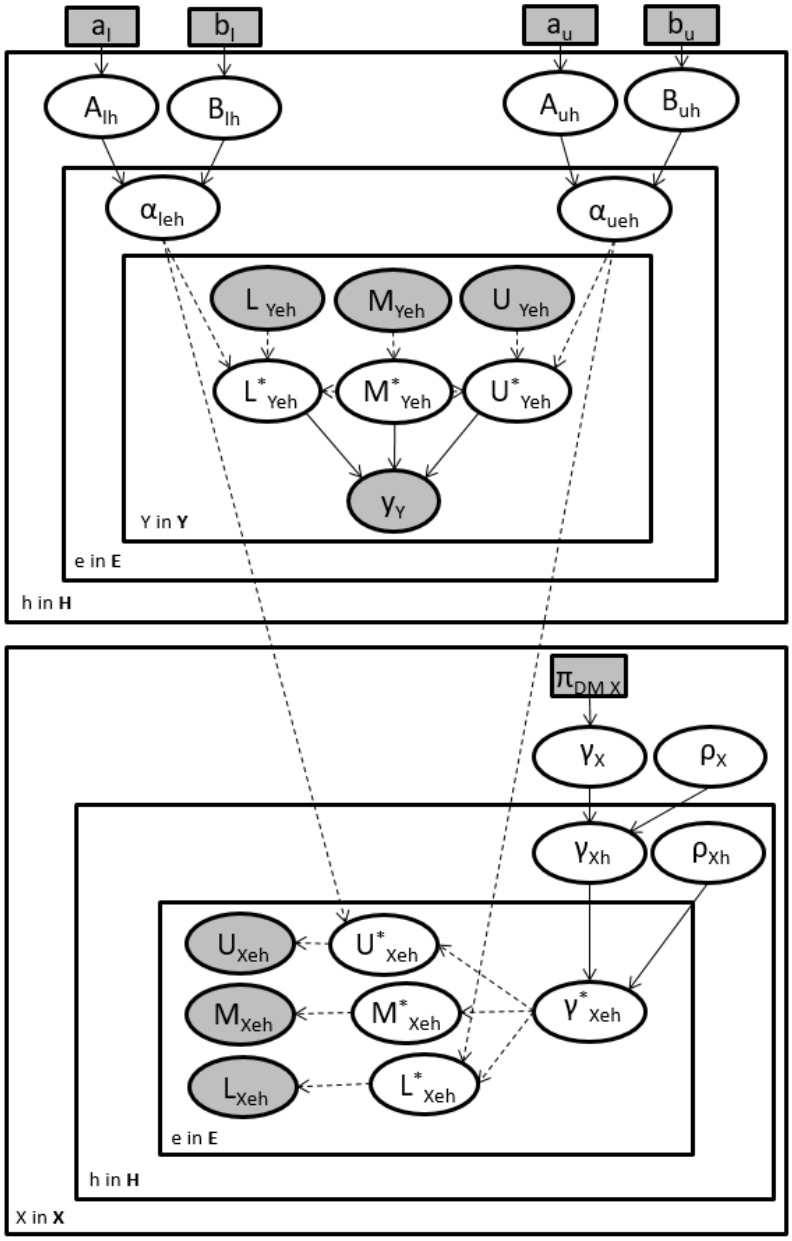


Figure 5.3: Directed acyclic graph for the linked aggregation and calibration models. The inflation factors from the calibration model are used to logically determine unbiased estimators in the aggregation model.

5.7.3 Full method outline - split normal parameterisation

When the distributions g_e are all defined to be a split normal then they can be represented by a single function such that:

$$g_e(x|L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}) \sim \begin{cases} \frac{1}{\sigma^*_{Xleh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-M^*_{Xeh}}{\sigma^*_{Xleh}})^2} & if x < M^*_{Xeh} \\ \frac{1}{\sigma^*_{Xueh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-M^*_{Xeh}}{\sigma^*_{Xueh}})^2} & if x \geqslant M^*_{Xeh} \end{cases} \tag{5.27}$$

where, the unbiased standard deviations σ^*_{Xleh} and σ^*_{Xueh} are calculated by:

$$\sigma^*_{Xleh} = \frac{M^*_{Xeh} - L^*_{Xeh}}{\delta_1} \quad and \quad \sigma^*_{Xueh} = \frac{U^*_{Xeh} - M^*_{Xeh}}{\delta_2} \tag{5.28}$$

and $\tau^*_{Xleh} := 1/\sigma^{2*}_{Xleh}$ and $\tau^*_{Xueh} := 1/\sigma^{2*}_{Xueh}$. Here δ_i represents the number of standard deviations between the elicited quantiles. The equation for g_e is identically defined for the seed variables and used in equation (5.18). The parameters γ^*_{Xeh} in equation (5.26) are then given by the triple, $(M^*_{Xeh}, \tau^*_{Xleh}, \tau^*_{Xueh})$. In this instance the generic aggregation component (equation (5.26)) is now replaced by:

$$\begin{aligned} L_{Xeh}|M_{Xh}, \alpha_{le}, \rho_{Xh} &\sim \mathcal{N}(M_{Xh} - \frac{\delta_1}{\sqrt{\tau^*_{Xleh}}}\alpha_{le}), \rho_{Xh}) \\ U_{Xeh}|M_{Xh}, \alpha_{ue}, \rho_{Xh} &\sim \mathcal{N}(M_{Xh} + \frac{\delta_2}{\sqrt{\tau^*_{Xueh}}}\alpha_{ue}), \rho_{Xh}) \\ M_{Xeh}|M_{Xh}, \rho_{Xh} &\sim \mathcal{N}(M_{Xh}, \rho_{Xh}) \end{aligned} \tag{5.29}$$

With **location parameter aggregation:**

$$\begin{aligned} M_{Xh}|M_X, \rho_X &\sim \mathcal{N}(M_X, \rho_X) \\ M_X &\sim \mathcal{N}(M_{DM_X}, \rho_{X0}) \end{aligned} \tag{5.30}$$

and **dispersion parameter aggregation:**

$$\begin{aligned} \frac{\tau_{Xlh}}{\tau^*_{Xleh}}|\tau_{Xlh}, \xi_{Xlh} &\sim \Gamma(\xi_{Xlh}, \xi_{Xlh}) & \frac{\tau_{Xuh}}{\tau^*_{Xueh}}|\tau_{Xuh}, \xi_{Xuh} &\sim \Gamma(\xi_{Xuh}, \xi_{Xuh}) \\ \frac{\tau_{Xl}}{\tau_{Xlh}}|\tau_{Xl}, \xi_{Xl} &\sim \Gamma(\xi_{Xl}, \xi_{Xl}) & \frac{\tau_{Xu}}{\tau_{Xuh}}|\tau_{Xu}, \xi_{Xu} &\sim \Gamma(\xi_{Xu}, \xi_{Xu}) \\ \tau^{-1}_{Xl} &\sim \tau^{-1}_{Xl0}\Gamma(a, a) & \tau^{-1}_{Xu} &\sim \tau^{-1}_{Xu0}\Gamma(a, a) \end{aligned} \tag{5.31}$$

The parameters now used to infer the target posterior are given by the triple

$(M_X, \tau_{Xl}, \tau_{Xu})$. These are used as inputs into equation (5.27) to create the full aggregate distribution.

To build the model the JAGs package embedded within R has been used. JAGs, built on the BUGs language, works using a Gibbs Sampling approach to MCMC, and in this combination allows relatively efficient calculations of results. Complete model runs typically take circa 10 minutes to complete 100,000 iterations. Modelling has been run in RStudio withRversion 3.6.1, JAGs version 4-10 on an AMD Ryzen 7 PRO 3700U processor, with 4 cores, 8 logical processors and 16.4GBs of virtual memory. These specifications are for a standard laptop. Timing was provided based on modelling only on a single core. Modelling was implemented in this way to mimic hardware available to study analysts whilst conducting an SEJ study live. If run-time is a concern, significant improvements can be made utilising a multi-threaded version of the code and deploying in a virtual environment with many cores.

Chapter 6

(Re)-Calibration

6.1 History of calibration

Calibration, sometimes known as re-calibration, of expert judgement has been studied since the inception of SEJ (Lichtenstein et al. [1977], Dawid [1982], Winkler [1981]). Formally, statistical accuracy (as highlighted earlier) refers to the assessment of long term frequency of outcomes relative to experts' responses. Re-calibration refers to the process of adjusting ones forecasts once statistical accuracy is known. Language in the broad expert judgement literature sometimes confuses these two concepts and so attempts have been made to create further linguistic clarity. Cooke used to term his *statistical accuracy* score as a *calibration* score, but moved away from this in order to ensure readers did not believe he was adjusting experts' judgements within his model.

Please note, the statistical accuracy definition we are using here has been given many names in the literature, such as, "reliability", "realism", "realism of confidence", "appropriateness of confidence", "external validity" and "secondary validity" (Lichtenstein et al. [1977]). All of these terms relate to the same mathematical phenomena.

Reminder, statistical accuracy is typically defined for continuous quantities in the following way. Suppose that we had an expert who was asked to make a judgement on the median quantiles for a number of variables. They would be responding with numbers which satisfy questions such as "What is x such that the following statement holds. There is a 50% probability that the kitchen will be fitted in less than x hours?" or "There is a 50% probability that more than x satellites will collide in orbit in the next decade. What is x?" Over a large number of such forecasts, you would expect to see 50% of the realisations end up above x and 50%

below x. If this is the case then an expert can be said to be well-calibrated. If across many such judgements the percentage of realisations below x is substantially different from 50% then the expert can be said to be mis-calibrated.

If the same expert had provided judgements over many quantiles for each variable it becomes possible to discern whether that expert is systemically over or underconfident. An *overconfident* expert would see significantly more realisations than expected above values they had given for quantiles greater than 0.5 and significantly more realisations less than expected for judgements they had provided for quantiles less than 0.5. An *underconfident* expert would display the opposite behaviour. If an expert is neither underconfident or overconfident on any quantile they are said to be *perfectly calibrated*. This behaviour can easily be visualised by examining the calibration curve of the expert, as highlighted in Fig.6.1 (Rausch et al. [2009], Dawid [1982], Lichtenstein et al. [1977]).

Bayesian approaches have considered calibration and its link to a related topic *coherence*, very carefully, (Clemen [1986], Dawid [1982], Dawid [1995], Winkler [1981], French [1986], Lichtenstein et al. [1977]). A *coherent* expert is one for whom there is internal consistency in their beliefs. For example an expert who deems the probability A of occuring to be 50% and the probability of $\neg A$ occuring to be 70% would be *not coherent*. Similarly, an expert who perceives there to be a 20% probability the height of a given tree is less than 100ft but 85% probability the height of the tree is greater than 150ft would also not be coherent. Coherence is perceived to be a relatively weak measure (Kadane. [1982]), but surprisingly it is readily visible in many SEJ studies, for example in Cooke's database (Wilson [2017]).

Dawid [1982], outlined mathematically how a coherent Bayesian expects to be well calibrated. Others have also more recently sought to shed light on the link between calibration and coherence (Wright et al. [1994]).

When an individual is not well-calibrated, recalibration is a potential method to adjust for these biases. Originally, correcting for ones own calibration issues was considered, particularly on sequential series (Dawid [1982]). Although, this process quickly becomes incoherent itself. Suppose that you had a calibration curve as described above which demonstrated that you typically overestimated that items you assessed to have probability 0.5, occurred 30% of the time. If you had a new event which you anticipated to occur 50% of the time, then the logical step utilising this curve would be to update your belief to 30%. This would imply that you had two separate beliefs about the probability here. One which is prior to the calibration process, the 50%, and one which is posterior to this, the 30%. These are conditional on different information. If the events under consideration are sequential and per-

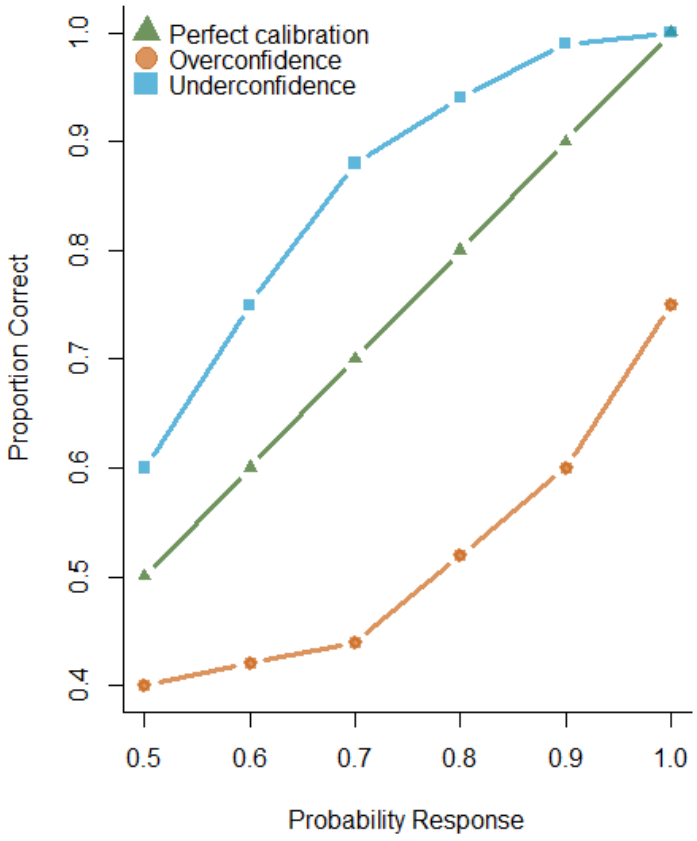


Figure 6.1: Calibration curves of a systemically overconfident and underconfident assessor compared to that of a perfectly calibrated assessor. Replicated from Lichtenstein et al. [1977].

ceived to be exchangeable then the posterior belief is also prior to the calculation of the calibration factor. This creates an inconsistency as definitions of exchangeability do not allow for accumulated experiences to alter probabilities.

Other challenges to the use of calibration have demonstrated that, assuming we are abiding by the standard rules of probability, if an expert is coherent, the only calibration curve that is rational globally (i.e. defined for every potential event of equal probability) and maintains coherence on the post calibration probabilities must be the line f(x)=x. This is true even for non sequential series. This means that there is no calibration adjustment curve that as an observer we can apply to the experts stated probabilities which will be coherent for all events ignoring the

trivial action of not adjusting at all (Kadane and Fischhoff [2013]).

It is trivial to see why this is the case at the median. Suppose we have an expert who has a calibration curve, f. Suppose further there is an event A which this expert believes will occur with probability 0.5, i.e. $P_e(A) = 0.5$. This event will thus have a post recalibration probability of $f(P_e(A)) = \kappa$ for some $\kappa \in [0, 1]$. From basic probability axioms however, $P_e(A) = P_e(\neg A)$. Here $\neg A$ is the event *not* A. This is because:

$$\begin{aligned} & P_e(A \cup \neg A) = 1 \\ \implies & P_e(A) + P_e(\neg A) = 1 \\ \implies & P_e(\neg A) = 1 - P_e(A) \\ \implies & P_e(\neg A) = 1 - 0.5 \\ \therefore & P_e(\neg A) = 0.5 \\ \therefore & P_e(\neg A) = P_e(A) \end{aligned} \tag{6.1}$$

The first statement holds here because the union of A and $\neg A$ is a sure event. The second line holds as A and $\neg A$ are, by definition, disjoint sets. Given our assumption that post calibration probabilities are coherent:

$$\begin{aligned} & f(P_e(\neg A)) + f(P_e(A)) = 1 \\ \implies & 2(f(P_e(A)) = 1 \,(\, Eq.6.1) \\ \implies & f(P_e(A)) = 1/2 \\ \therefore & \kappa = 0.5 \end{aligned} \tag{6.2}$$

The first line of the above is true as we have defined that post calibration probabilities must also be coherent. Thus we have shown that any global calibration map from coherent judgements to coherent calibrated judgements must pass through the point, (0.5,0.5). Kadane and Fischhoff [2013], expands upon this to demonstrate that such events (or groups of events) can be constructed for any probability in [0,1] and thus the calibration curve must be the line $f(x) = x$. Further challenges on the link between coherence and calibration are given in Oakes [1985].

Other potential challenges to recalibration are more philosophical. A key challenge highlighted earlier is with regards to ownership of judgements. In an expert judgement problem, if a DM were to recalibrate experts opinions before aggregating then in some sense the judgements are no longer those of the experts. Rather, they reflect the DM's belief of the uncertainty information the expert is providing, given their historic judgement performance. In the *rational scientist* context outlined in Cooke [1991] or EFSA [2014], this may appear inappropriate. In the event a DM

has to take sizeable risk given the uncertainty profile provided by an SEJ study, however, understanding the extent of how miscalibration can impact this result seems important. To bypass this issue completely would potentially risk wilfully ignoring critical information.

The tendency for overconfidence in expert judgements is well recognised (Brenner et al. [2005], Lichtenstein et al. [1977]). Whilst some disciplines, particularly weather forecasting appear to have well-calibrated experts (Dawid [1982], Lichtenstein et al. [1977]), this is not the norm. Lichtenstein et al. [1977], noted that for continuous variable elicitation:

"A nearly universal bias is found : assessors' probability density functions are too narrow. ..., This bias reflects overconfidence; the assessors think they know more about the uncertain quantities than they actually do know."

Despite the issues with recalibration raised above, the calibration issue cannot be ignored. One approach is to try and train experts to reduce the likelihood of mis-calibration (Lichtenstein and Fischhoff [1980]). This is one of the key aims of the procedural elements outlined earlier (Cooke [1991], EFSA [2014], Hemming et al. [2018]). However, in practice, even with these procedural improvements overconfidence is still readily visible in nearly all SEJ studies. There is more to be done to support DM's in dealing with this bias

The model outlined earlier has a recalibration component. Before advocating for the use of this however, particularly in light of the challenges outlined within this section, it is important to ensure that recalibrating in this way adds quality to decision making rather than detracts. Rather than tackle this from a philosophical perspective, the efficacy of this approach is demonstrated by running a cross-validation on empirical data.

6.2 Cross-validation of recalibrating experts

For a DM to trust that recalibration can improve their final understanding of underlying uncertainty, there must be empirical evidence that recalibrating enhances the statistical validity of the analysis. In order to generate such data for our method we have employed a cross-validation technique as outlined earlier.

The cross-validation approach we have followed conducts a leave-p-out analysis for 24 data sets from Cooke's broader database. These studies are a subset of those in the analysis originally published by Eggstaff et al. [2014] in which there are uniform scales for all of the variables. As highlighted already the mechanism of calibration we are using does not work for variables on multiple or logged scales.

As described in section.4.3, we perform an exhaustive cross-validation where we define p to be the the maximum of either 20% of the seed variables (80% of the seed variables remain in the training set) or 2. This ensures that we never just take 1 variable into our testing set.

For each training/testing set partition we run the calibration component of the Bayesian model outlined previously. This generates a *calibrated* version of each expert's judgements for the seed variables in the testing set, in addition to their original *uncalibrated* perspective.

Note, uncalibrated here refers to the fact that this is the original elicited judgement and has not been recalibrated. This does not imply that these judgements are inherently mis-calibrated.

For a given partition, we then aggregate all of the experts' calibrated and uncalibrated testing set perspectives together into a single dataset which we pass through Cooke's Classical model, setting the testing set as the seed variables. Statistical accuracy and information scores can then be generated by expert for this particular testing set. It is important we group all of the judgements together in this way before passing through Cooke's model as the information score is a relative number and therefore scores for all experts, both calibrated and uncalibrated, need to be calculated simultaneously.

If the calibration process adds incremental benefit the final combined scores for recalibrated experts should be higher then their original uncalibrated judgements.

Conducting the analysis in this way results in 53162 individual forecasts and 20386 individual expert testing set partitions. The final number of expert study combinations was 189. There is both calibrated and uncalibrated scores for each of these 189 combinations with statistical accuracy, information and combined metrics for each.

Fig.6.2 demonstrates the recalibration process delivered results in line with expectation. As experts are typically overconfident, (Lichtenstein et al. [1977], Burgman [2016]), the calibration process has created wider bounds for the expected uncertainty than the original elicited numbers. This results in higher statistical accuracy scores at the cost of information. In total the calibration process has increased the statistical accuracy score in 92% of the tested data sets. In Fig.6.2, if a dot lies above and to the left of the dotted x=y line then the calibration process has improved the score for this metric. Almost universally the dots in the left hand pane are above and to the left of the x=y line.

Conversely, the information score for each of the calibrated experts is worse

than their uncalibrated counterpart. It only improves in 1.6% of the tested experts. This is in line with what we would expect to see if experts were systemically over-confident. In these instances improving the statistical accuracy results in broader uncertainty bounds, which, by definition have lower information.

Fig.6.2 also demonstrates that the range of improvement is much broader for statistical accuracy than it is for information. The dots are spread further from the x=y line. It is important not to read too much into this, however, it is likely to be driven by the fact that statistical accuracy is a faster function than information. Small improvements here can result in radically different scores.

The calibration process has thus traded statistical accuracy for information. This alone is potentially very helpful for a risk averse DM. If a DM wished to understand the domain of feasible outcomes for a particular target variable $\mathbf{X}$, then to consider more statistically accurate judgements would give a better perspective on this range of uncertainty. This improvement however, has come at the expense of mathematical information. It is important to note that there is no loss of physical information and there is an argument that the "lost" information is vacuous in nature.

The key challenge therefore is to understand whether too much information

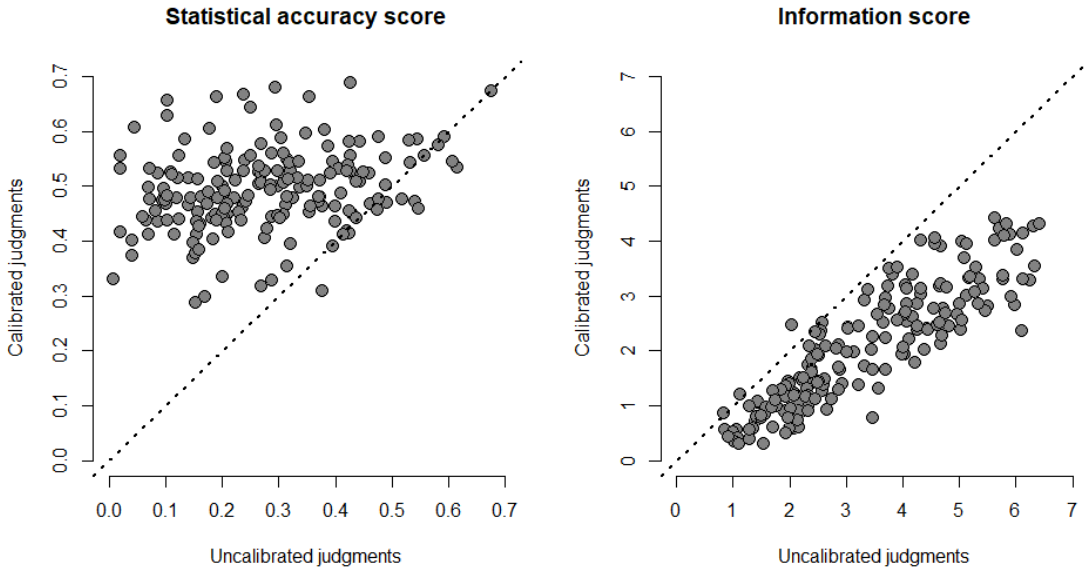


Figure 6.2: Recalibrated experts showed higher statistical accuracy than their original judgements in 92% of the expert study combinations tested. The information score however diminished as a result of the calibration process in all but 1.6% of cases.

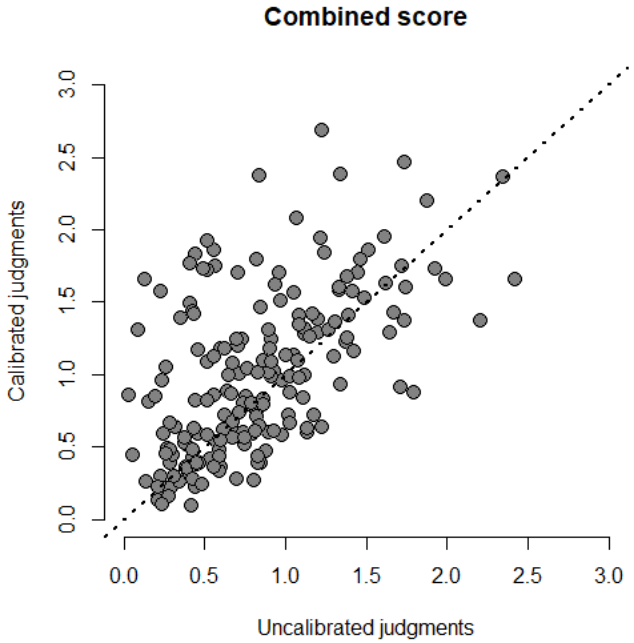


Figure 6.3: Recalibrated experts showed higher combined scores than their original judgements in 60% of the expert study combinations tested.

has been traded for the gains seen in statistical accuracy. This can be done by considering the combined score for each of the 189 combinations considered. Fig.6.3, shows the relative combined scores of experts prior and post calibration. 60% of the dots lie above the line denoting that the calibration process has improved the scores of the experts in over half of the cases studied.

Fig.6.3 also shows that the combined score varies more significantly from line y=x, above and to the left of the line than it does below and to the right of it. This suggests that when the calibration process has caused the overall combined score to decrease, it will typically do so less than the potential upside when it works. This is a positive for a potential practitioner of recalibration as it suggests that the potential risk reward ratio is favourable.

Examining the study level results allows us to draw some insight into when the recalibration exercise has improved the forecasts from the experts and when it has hindered. Fig.6.4 demonstrates on a study by study basis, the number of experts for whom the calibrated judgements had a higher combined score than the uncalibrated judgements and vice versa. The plot is ordered based on the proportion of experts for whom calibration improved their scores. Plotting the data in this way

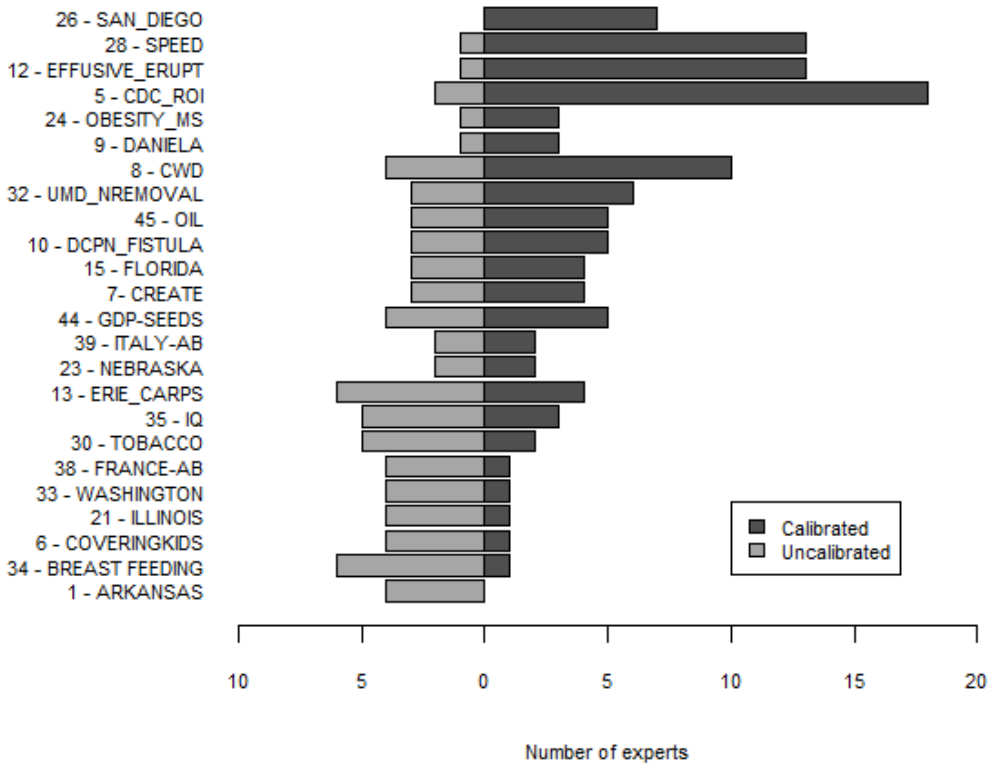


Figure 6.4: The number of experts by study in which either the calibration method improved the overall combined Classical model score (Calibrated) or diminished it (Uncalibrated). The recalibration exercise improved expert judgements more often in the larger studies. Study data is taken from Cooke's SEJ database, (Cooke and Goossens [2008]).

demonstrates that calibration certainly does not universally help, aligned as we might expect to the 60% figure mentioned earlier. Some studies have a significant number of experts for whom the calibration process does not improve their combined score.

Notable in Fig.6.4 is the relationship between the total number of experts and the impact of calibration. Studies which have more experts typically have a higher proportion for whom calibration has improved their scores. There are only 4 studies in the tested set who had more than 10 experts involved. In all of these 4 studies more than 70% of the experts would have seen improved forecasts if their recalibrated numbers were used in lieu of the uncalibrated original judgements. For 3 of the studies greater than 90% of experts saw improvement through recalibration.

Whilst there would need to be significantly more analysis conducted to ensure that this is a true phenomena, it is hypothesised that this is due to the hierarchical

nature of the Bayesian calibration process. Experts are calibrated individually but there are also latent hierarchical variables which assess and control behaviour at the total expert group level. With more experts in the study this shared knowledge about calibration increases. This sharing of information may have positive implications on the balance between statistical accuracy and information that the recalibration process drives.

6.3 Implications for the Bayesian model

The data outlined here does not imply that recalibration of judgements utilising the new model structure is a panacea for dealing with overconfidence issues. It does however suggest that recalibration is also not a fruitless task. Adjusting for systemic over/underconfidence can have tangible benefits to *a posteriori* analysis of SEJ data.

The mechanism for expert recalibration used in this analysis has been standard across all of the data sets considered. Namely, only upper and lower quantiles provided by the experts were recalibrated, specifically 0.05 and 0.95. No attempt was made to recalibrate internal quantiles if more than three were provided. A single set of relatively uninformative priors for the analysis was also used, aligned to those outlined in section.5.3. More targeted decisions regarding both the specific parameterisation approach and the priors used could feasibly drive further improvement. This is left for further research.

There is sufficient evidence in the above analysis to justify having a calibration module within our Bayesian approach. More work is necessary however to understand the causal factors which determine when it is beneficial vs. detrimental. Ultimately it would come down to context and DM preference to determine when and how to use it. For the next chapter, in which we analyse some applications of the broader Bayesian model, the recalibration module has been included.

Finally, it is important to note, that whilst this recalibration approach was built as a single step within a broader Bayesian model, there is no reason that it needs to be utilised in this way. If through more nuanced analysis of the data, it becomes possible to more concretely replicate the recalibration behaviour shown in the top performing studies analysed here; there is a strong argument to suggest that regardless of the aggregation approach taken, recalibrating judgements in this way could be a critical prior step within certain SEJ studies. If the experts in Study-26, the San Diego study for example, were recalibrated and then aggregated either through the equal weighted linear opinion pool or Cooke's Classical model, better overall prediction would be expected.

Chapter 7

Applications

7.1 Cooke's data

In this section we assess the effectiveness of the new combined model by running it against empirical studies from Cooke's database and comparing the resultant target variable forecasts to the more commonly utilised opinion pooling techniques. The results on five studies will be outlined. Studies have been chosen at random from a subset in which all forecasts are on a uniform scale.

The first study considered (Arkansas) formed part of a broader study conducted by the Centre for Disease Dynamics, Economics and Policy looking at grant effectiveness and child health insurance enrolment for the Robert Wood Johnson Foundation. The second case study we will review (CWD) was conducted for the University of Ottowa to assess infection transmission risks for chronic wasting disease (cwd) from deer to humans (Tyshenko et al. [2011]; Tyshenko et al. [2012]). The remaining case studies represent a volcanology study (Loughlin et al. [2012]) considering the risks of a Laki like eruption, an ecology study (Wittmann et al. [2015], Wittmann et al. [2014], Zhang et al. [2016]) looking at the risk of invasions of bigheard and silver carp in Lake Erie and a glaciology study (Bamber and Aspinall [2013]) examining ice sheet melting behaviour.

Whilst in principle we may wish to encode DM knowledge into the analysis, these studies have been conducted in the rational scientist context outlined in EFSA [2014]. In this context, relatively uninformative priors are appropriate, outlined below, and a consistent set of priors can be used for both studies.

The rational scientist context is a common one in SEJ. Here, rather than modelling the impact of elicited expert judgements to an individual DM's belief, the aim is to combine the judgements in a way that represents what a rational

scientist would believe given the experts' inputs. Typically, all of the knowledge for that hypothetical rational scientist should be encapsulated within the experts' judgements and hence relatively flat (and thus uninformative) priors are desired.

One of the advantages of modelling expert judgement in a Bayesian way is that this can easily be done, but, if DMs do have prior belief they wish to embed in the model the mechanisms inherently exist to do this. In many deployments of Bayesian models the modelling process can be quite distinct from DMs and model priors are defined by the analysts performing the work. It is strongly recommended for SEJ that this approach, for target variable priors, unless in a rational scientist context, is avoided.

If any prior knowledge on the target variables needs to be incorporated into the model, this should be elicited directly from the DM. This can be difficult, particularly as studies often run at a distance from decision makers, but I would contend that it is an essential step in deploying a Bayesian study correctly. The process of eliciting DM prior belief can operate as a mechanism to increase traction with stakeholders, help facilitators understand the sensitivity of the decision to judgements given and ultimately increase the chance of study outputs being utilised. Whilst admittedly challenging, I have successfully elicited prior beliefs from many decision makers in the past. When DM prior belief is completely unattainable other groups may act as a proxy. This is discussed in more detail in Chapter 8.

Incorporation of decision maker belief, often happens when SEJ is deployed in private enterprise. Public sector SEJ studies often take the rational scientist view outlined. As such, publicly available data-sets are often in this context, which, given the desire to compare to existing models, is the reason for the case studies outlined. SEJ is more commonly found in the public sector than the private sector, although the authors would like to see greater traction for these methods in private enterprises.

In a typical study there are often very few experts and elicited quantiles. The full set of hyperparameters, $(a_l,b_l,a_u,b_u, M_{DM}, \rho_0, \rho, \rho_h, \xi, \xi_h, \tau_0, a)$ consequently must be modelled carefully as their influence will be important.

The calibration component parameters $(A_{lh},B_{lh},A_{uh},B_{uh})$ will be set, as per (5.6), s.t. $A_{lh} \sim Pois(a_l)$ and $B_{lh} \sim Exp(b_l)$ (and equivalent for the upper bounds) and $a_l = b_l = a_u = b_u = 2 \; \forall h \in \mathbf{H}$. This provides a suitably diffuse prior, centred around 1, for the dispersion parameters of the calibration model. This is identical to the parameter values first proposed by Clemen and Lichtendahl [2002].

The aggregation component hyperparameters $(M_{DM}, \rho_0, \rho, \rho_h, \xi, \xi_h, \tau_0, a)$ are set with weakly informative priors.

Remark. given our split normal parameterisation there are actually hyperparameters for both the upper and lower model dispersion parameters, thus, there is effectively a ξ_u and ξ_l, and equivalent. In practice we set these hyperparameters to be identical, therefore for simplicity this distinction is omitted below.

Similar to Albert et al. [2012], we shall apply truncated Normal priors for the dispersion components as this is a useful we shall apply truncated Normal priors for the dispersion components as this is a useful simplification of a folded noncentral-t distribution that is conjugate for these parameters, as shown by Gelman [2004] :

$$\sqrt{\rho_0} \sim \mathcal{N}_+(0, \varphi_0); \quad \sqrt{\rho} \sim \mathcal{N}_+(0, \varphi); \quad \sqrt{\rho_h} \sim \mathcal{N}_+(0, \varphi_h); \tag{7.1}$$

Where $\varphi_0, \varphi, \varphi_h$ are selected to be weakly informative. Given the uniform scales of the examples that we are looking at, and the fact that in practice to ease modelling we normalise everything onto [0,1] (and then readjust back to the original scale at the end), setting $\varphi_0 = \varphi = \varphi_h = 1000$, provides a sufficiently diffuse prior for these hyperparameters. Similarly, we can switch the prior on the variance component τ^{-1}, from a gamma $(\tau_0^{-1}\Gamma(a, a))$ to a truncated normal prior $\sqrt{\tau^{-1}} \sim \mathcal{N}_+(0, \psi_0)$ where ψ_0 is selected in order to be reasonably diffuse, here we also select $\psi_0 = 1000$. For the intermediary hyperparameters, ξ, ξ_h, we stick with a gamma prior, as outlined in the orignal model, however, we do not use completely diffuse priors and set $\xi = \xi_h = 1.5$. As Gelman [2004] highlighted, utilising very small components for the terms in the gamma distribution puts a lot of the mass at zero, which for this model is unfavourable. Utilising the above structure, allows us to set a prior which has more of the mass centred at 1, building the assumption that there is similarity in intra homogeneity group dispersion parameters, whilst being sufficiently diffuse enough to learn the true nature of these intra group dispersion relationships from the data.

With these hyperparameters set, we review the results for each of the studies outlined earlier and compare these to opinion pooling methods. It is important to note here that the comparison is not aimed at showing superiority of the new method compared to the other methods. The desire is to better understand how the hierarchical modelling, and the focus on the consensus of experts, drives a differing perspective to the opinion pooling methods. Specifically, the two methods which we will compare to are an equal weighted linear opinion pool of the form (3.1) where $\omega_e = 1/|\mathbf{E}|$ (with a DM referred to as EWDM) and a performance weighted linear opinion pool where ω_e is defined by performance over the seed variables and is determined by Cooke's Classical model, outlined previously (PWDM).

7.1.1 Arkansas example

The Arkansas study, originally conducted in 2012, had 4 experts who were required to assess 10 seed variables and 20 target variables. An example of the seed questions utilised (with the values known *a priori*) were "What is the ratio between the number of children without health insurance in Arkansas / number of children without health insurance in Louisiana?" with a true realisation of 0.66. The target questions were of a similar nature; e.g. "What would the participation rate for public insurance be in 2020 if CHIPRA were not renewed in 2013?". Here CHIPRA refers to the Children's Health Insurance Program Reauthorization Act, which was signed into action by President Obama February 4th 2009. All of the data was elicited against 5 quantiles (0.05, 0.25, 0.5, 0.75, 0.95), although given the above parameterisation we will only utilise 3 of these within the model.

The first component to analyse is the clustering component (the outputs of which are then embedded within the broader calibration and aggregation model). Running the agglomerative hierarchical clustering approach over the seed variable space results in a proposal to split the experts in to three homogeneity groups. The dendogram for the hierarchical clustering can be seen in Fig.A.1 in the appendix. Expert 1 and expert 4 sit within their own groups and experts 2 and 3 are clustered within a single homogeneity group.

Following homogeneity group definition, we assess the impact of calibration. A very simple pre-analysis of the data suggests the experts are not well calibrated, with a significant bias towards overconfidence. If all of the experts were statistically accurate we would expect that circa 4 of the 40 seed variable estimations across the experts (10%) would be outside of the experts' elicited quantiles. In practice 43% of the realisations fell above the experts' 0.95 quantile or below the 0.05 quantile. This ranged from 80% for expert 1 to 30% of assessments for expert 3. To this extent, when we review the calibration parameters in the Bayesian model we would expect these to compensate for this behaviour and increase the expected uncertainty vs. what was outlined by the experts. The medians for the experts' miscalibration parameters range from 1.3-3.6 for a_{le}, and from 2.0-11.4 for a_{ue}. All of these values are greater than 1, indicating systemic over-confidence rather than under-confidence, aligned to the expectations from the pre-analysis. Expert 1, sits at the upper end of this range for both variables. The scale of these calibration parameters suggests the experts' forecasts are significantly over confident. The DM should consequently be very careful when assessing whether the originally elicited judgements from these experts give a true picture of the uncertainty present.

Focussing on experts' forecasts for a single target variable, Fig. 7.1 out-

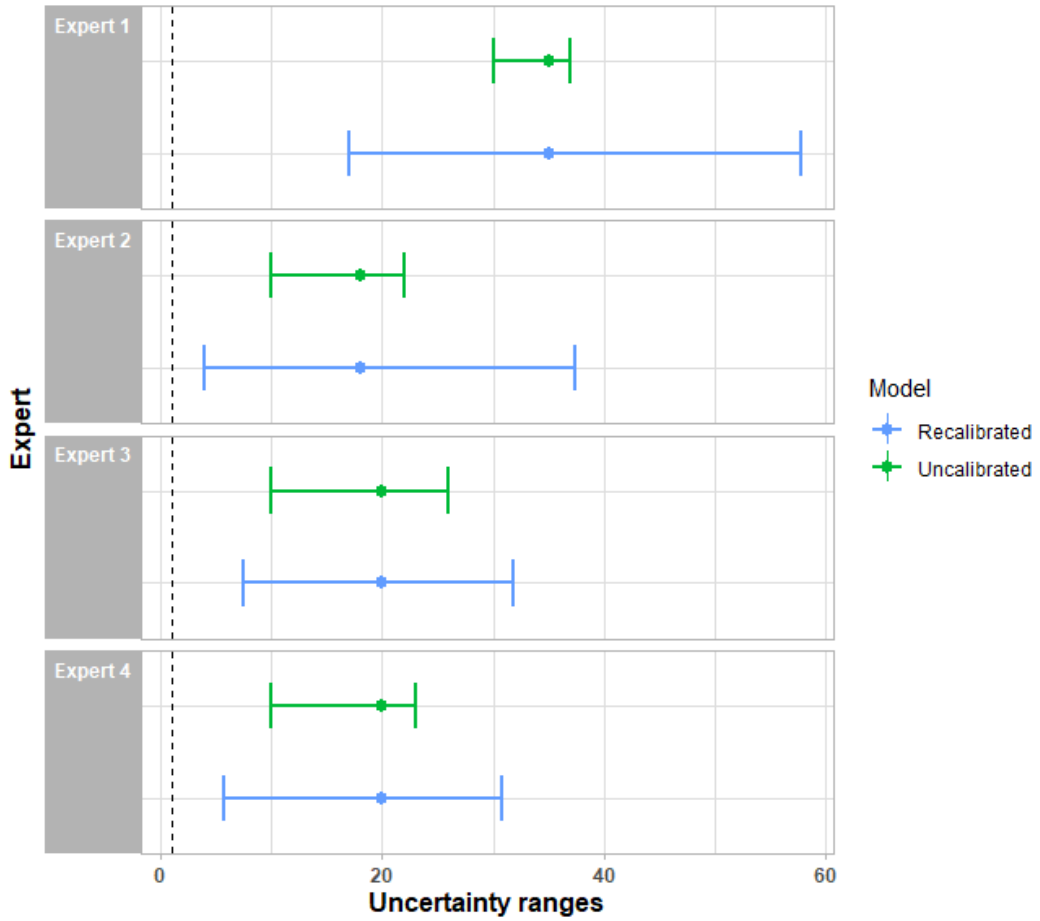


Figure 7.1: Effect of recalibration on experts' estimates within the Arkansas study on the question: "What would the adolescent well-care visit rate be in 2020 without RWJF Covering Kids and if CHIPRA were not renewed in 2013?" All experts show a significant broadening of their quantiles as a result of the recalibration, particularly expert 1.

lines the impact this recalibration has to the estimates used in the model. Variable 10 ("What would the adolescent well-care visit rate be in 2020 without Robin Wood Johnson Foundation (RWJF) Covering Kids and if CHIPRA were not renewed in 2013?"), is chosen here to demonstrate the difference in output for the three different modelling types. Of particular note, is expert 1 whom despite being very confident in their response initially, once recalibrated, as expected, display a much broader distribution.

It is interesting to also note the underlying structure of the responses to this variable. In the original forecasts expert 1 estimated with high certainty the true value will be greater than 30 whereas the other three experts estimate the median

of this variable at circa 20 (and strictly less than 30). The recalibration exercise has shrunk this discrepancy. The recalibrated judgement for expert 1 now overlaps considerably with the other three experts. We will return to this later when we assess the output distributions.

Finally, we focus on the aggregation component and the posterior distribution for the DM. One element to be considered when building this final posterior distribution is how to combine the posteriors of the components (M, τ_l and τ_u) into a single output. Within the initial MCMC these components are modelled separately; in each run we have a posterior distribution for each but no combined distribution. We create this combination by applying a secondary Monte-Carlo analysis drawing triplets from each distribution (here we actually use samples from the original model, post convergence) which we fit back to our split normal structure. We sample from this to give our combined posterior.

Sometimes there are irreconcilable differences in opinions between groups of experts' judgments which when equally weighted in a linear opinion pool would result in bi-modality. This can be driven by fundamental different causal models or divergent sets of beliefs between experts. Both may be entirely justifiable. It would be a mistake to lose this discrepancy in the modelling process. It is possible within

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Variable 1	51.9	74	97.5	56.2	95.4	97.9	46.2	75.6	95.3
Variable 2	62.1	92.5	99.5	85.5	96	99.6	58.7	87.9	98.7
Variable 3	54	76.5	98.6	56.2	94.3	97.9	51.2	79.1	96.8
Variable 4	35.4	84.4	95.6	54.3	92.9	96	39.2	73.7	95.6
Variable 5	32.1	70.1	95.4	36.6	92.8	95.9	33.8	67.2	92.1
Variable 6	10.2	17.5	25.2	10.3	20	25.9	6.6	18.5	30
Variable 7	15.9	28	38.1	16.8	26.2	37.5	12.9	29	44.1
Variable 8	11.9	26.2	38.1	10.2	21.1	37.3	10.3	28.1	45.6
Variable 9	11	23.2	36.4	12.5	25	30	10	25.7	41.9
Variable 10	10.4	20.8	36.4	10	19.1	26	7.8	23.8	41.6
Variable 11	5.9	16.1	30.4	5	18.1	32.6	-0.5	16.7	37.9
Variable 12	24.7	51.9	67.9	20.6	52.1	67.8	20.9	52.5	88.9
Variable 13	15.2	49	66.3	10.6	35.6	58.9	17.8	49.2	84.3
Variable 16	39.5	71	84.4	35.5	65.8	84.6	45	72.7	93.7
Variable 17	80.9	90.5	98.6	80.3	92.5	99	77.9	90.5	98.3
Variable 18	50.6	87.3	94.5	45.7	72.8	92.4	60.8	85.9	97.5
Variable 19	67.6	89.1	96.2	75.1	87.6	96.5	66.2	85.8	97.2
Variable 20	45.9	81.2	90.8	40.7	70.8	86.9	52.9	80.9	96

Table 7.1: Comparison of DM quantiles for different modelling approaches to the Arkansas study.

the Bayesian framework proposed here to always recover this multi-modal picture by examining the homogeneity groups posterior distributions in addition to that of the global output. An advantage of the Bayesian approach is that all of the latent elements within the model can be extracted and examined after the modelling is completed. This allows decision makers to move forward with confidence but also provides an opportunity to decide which elements to base decisions on. At times basing a decision on the multi-modal homogeneity group picture rather than the consensus distribution at the global level may be appropriate.

Reviewing all target variables in the study; Table 7.1 outlines the resultant posterior and how this compares to the estimates for the EWDM and the PWDM. (Please note. Variables 14 and 15 have been removed from this list as not all experts predicted these two target variables). It is clear from the data that there are similarities between the approaches in the mid-quantile assessment with the Bayesian decision maker (BDM), having a mean difference of 1.5% from the EWDM and 2.2% from the PWDM, (with mean absolute differences of 4.6% and 14.8% respectively). For the outer quantiles (0.05,0.95) the BDM model produces much wider final distributions than either of the other two models. Here the BDM suggests there is a significant probability that the true value of these realisations will lie significantly below or above the estimates of either the PWDM or the EWDM. This should not be surprising, given the calibration data reviewed earlier, however, it is important to review the full distributional forms rather than just the quantiles to understand the impact of these fluctuations.

Rather than review all of the distributions in detail, we will examine the distribution for the tenth target variable[1], which was outlined earlier. Variable 10 was selected as there was notable discrepancy between the EWDM and the PWDM distributions. This gives us an opportunity to understand how the BDM model compares to other models in cases where there is more complication in the underlying data structure. Distributions for all remaining target variables have been included in Fig.A.2 in the appendix.

The EWDM distribution in Fig. 7.2, is multimodal, aligned to the calibration point plot shared earlier. There is more mass under the first peak reflecting the lower

[1]From the nature of the questions asked within the study, some of the variables are bounded, i.e. the output is a % that must be between 0 and 1. Without intervention, the BDM in these cases may produce a posterior distribution that sits outside of these bounds as we do not constrain the model in formulation. Thanks to the constant in the Bayesian formula, we can simply do this by applying a further prior which is uniform on the unit interval (and zero outside of this) and then rescaling the posterior as necessary. All of the values in this study, when comparing to the other modelling types, have been adjusted accordingly. Another mechanism for imposing bounds, considering the generic model outlined, would be to select a parameterisation, such as a beta distribution, which constrains the bounds by default.

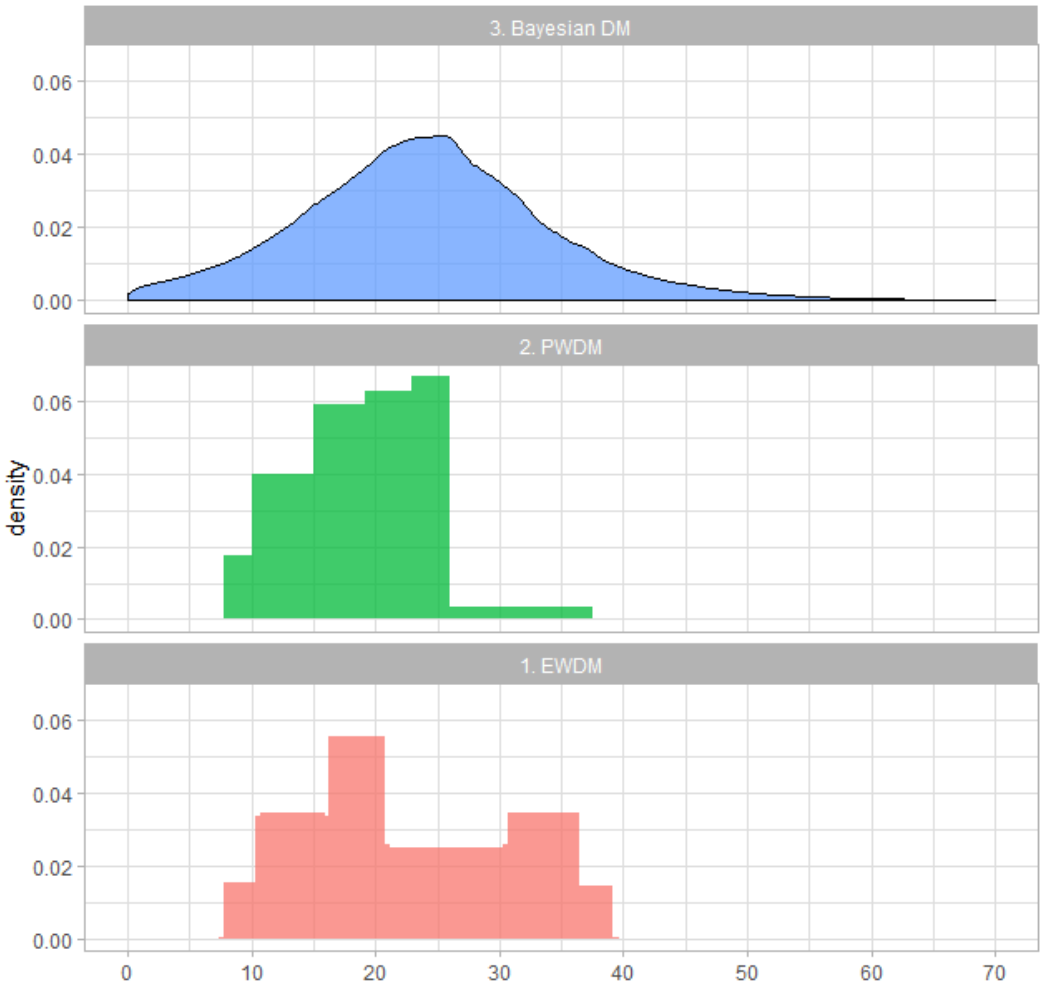


Figure 7.2: Comparison of final distributions to the question "What would the adolescent well-care visit rate be in 2020 without RWJF Covering Kids and if CHIPRA were not renewed in 2013?" The Bayesian model (blue) demonstrates a larger support, aligned to the overconfidence demonstrated by experts in the seed variables.

assessments provided by three of the experts. For the PWDM we see a unimodal distribution aligned to the lower estimations. This is driven by the relative weighting of Cooke's model which determined that 23% of the performance weighting should be assigned to expert 2 and 77% of the weight assigned to expert 3. Expert 1 and expert 4 were eliminated. Thus, the final distribution is dominated by experts who had a lower estimate of the variable as expert 1's judgements are not included. By design, as we are assessing the consistency in opinion, the Bayesian model also has a unimodal posterior. Evident from Fig. 7.2 is also the heavier BDM tails relative

to the other two models. The mode of the distribution, unsurprisingly, sits between the two peaks of the EWDM, however is skewed slightly to the lower peak. This is aligned to the calibration behaviour shown earlier and the relative homogeneity group weightings in the model. In this way, the Bayesian model smooths the variabilities between the experts, whilst still modelling the underlying differences in the estimations.

The beauty of the Bayesian model is that even though the homogeneity group structure has been compressed into this final posterior distribution, it is still possible to learn about the underlying model behaviour. We can examine the posterior homogeneity group parameters which can easily be recovered in addition to the complete posterior DM distribution. Fig.7.3 outlines the homogeneity group distributions. Group 1 comprises of just expert 1, group 2 includes expert 2 and expert 3, and group 3 is just expert 4. The rationale for the skew in the Bayesian model towards the lower peak of the EWDM is evident here. Group 1 has a distinctly different distribution to groups 2 and 3, driven by expert 1's differentiated belief compared to their peers. The final Bayesian DM distribution is weighted more to the common belief demonstrated in group 1 and group 2, but less so than if the experts had just been aggregated directly. In this way, the differentiated perspective of expert 1 has had increased weight. This is one of the advantages of the Bayesian model, the model design and software implementation facilitates a deepdive of the results, beyond just the final distributions, to support the DM decision making.

7.1.2 The impact of our elicited quantile choice

The Arkansas study originally elicited 5 quantiles (0.05, 0.25, 0.5, 0.75 and 0.95) for both the target and seed variables. Up to now, given the parameterisation choice outlined earlier, we have only been considering three of these (0.05, 0.5, 0.95) thereby effectively discarding some of the elicited information. Utilising the remaining data points and different modelling approaches allows us to determine the impact of our quantile choice and inform more efficient elicitation in the future. There are many possible combinations of quantiles which we could use to either inform the calibration inflation factors or the target variable aggregations. We have prioritised four combinations for further analysis and compare these to our original model structure:

- ***OuterQuantiles***: Our original parameterisation: quantiles 0.05, 0.5 and 0.95 are used to determine two inflation factors α_{le} and α_{ue}. The same quantiles are used for aggregation on the target variables.

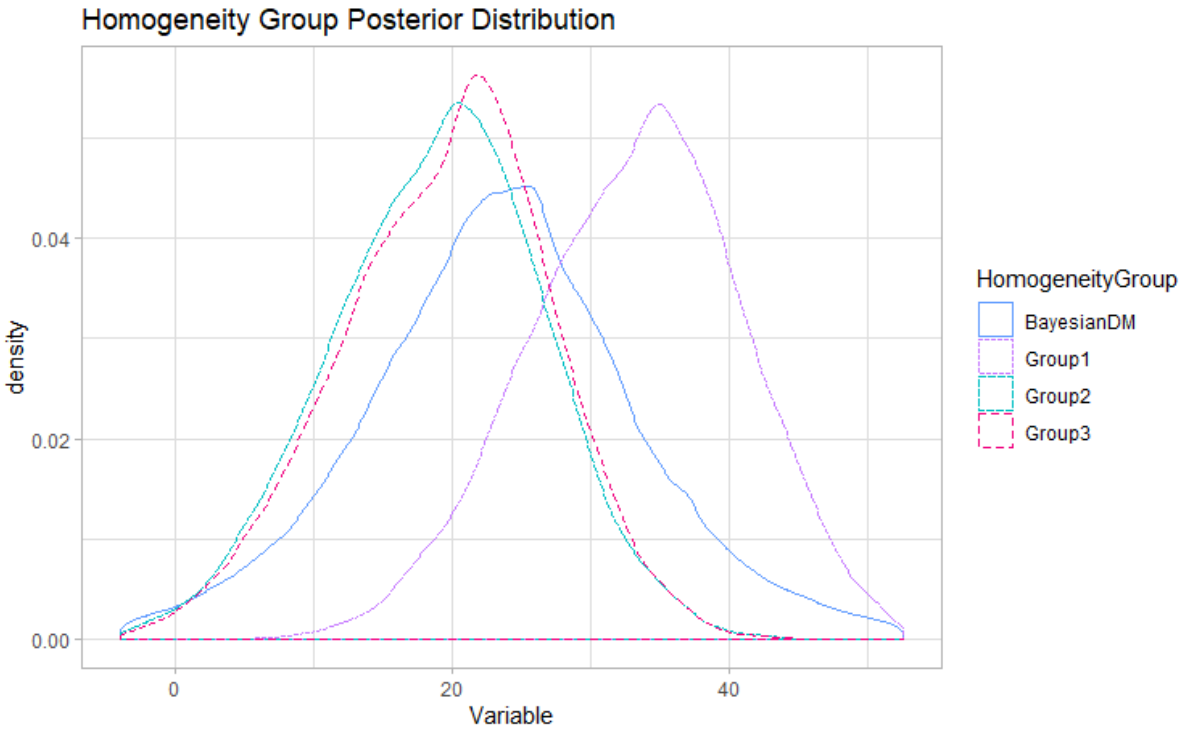


Figure 7.3: Comparison of homogeneity group distributions to the question "What would the adolescent well-care visit rate be in 2020 without RWJF Covering Kids and if CHIPRA were not renewed in 2013?" Group 1 = {expert 1}, Group 2 ={expert 2, expert 3} and Group 3 = {expert 4}.

- ***InnerQuantiles***: We still use three quantiles for determining α_{le} and α_{ue} but now consider the elicited quartiles 0.25, 0.5 and 0.75. The same quantiles are used for aggregation on the target variables.

- ***AllFiveQuantiles***: Use all five elicited quantiles for determining α_{le} and α_{ue}, effectively giving more power to the calibration model. Only the OuterQuantiles are used for aggregation on the target variables.

- ***BetaOuterQuantiles***: All five elicited quantiles are used for calibration. α_{le} and α_{ue} are calculated as before but we extend the model to allow positional uncertainty in the median by creating an inflation factor β_e s.t. $M_e^* = \beta_e M_e$. Outer quantiles (0.05, 0.5, 0.95) are used for aggregation on the target variables.

- ***BetaInnerQuantiles***:Identical to the BetaOuterQuantiles except the inner quantiles (0.25, 0.5, 0.75) are used for aggregation of target variables.

	Inner Quantiles	All-Five Quantiles	Beta Outer-Quantiles	Beta Mid-Quantiles
D Statistic	0.0395	0.0248	0.0852	0.0791
p-value	$< 1*10^{-9}$	$<1*10^{-9}$	$<1*10^{-9}$	$<1*10^{-9}$

Table 7.2: Kolmogorov-Smirnov test on the impact of quantile parameterisation for target variable 10 in the Arkansas study.

Applying these five model parameterisation types, to the target variable outlined early, results in different posterior distributions for the DM as demonstrated in Fig.7.4. What is striking in Fig.7.4 is that whilst there are some minor differences between the posteriors of each approach (particularly in the tails) these are relatively insubstantial. It is reasonable to assert that a DM is unlikely to make a substantively different decision regardless of the parameterisation choice used. We apply a Kolmogorov–Smirnov test to 10000 samples from each distribution and consider relative to the original model (OuterQuantiles) to assess the mathematical difference between the cumulative distribution functions (c.d.fs), Table.7.2. The p-value in this test demonstrates that in all cases the posterior distributions are with very high likelihood not the same (or strictly speaking, they are not both samples from an identical underlying distribution). Utilising the D statistic from the test does demonstrate however, that whilst they are not identical distributions they are very similar. The D-statistic can be interpreted as the maximum distance between the two tested c.d.fs. Thus if we consider the two parameterisations which do not have a beta term (InnerQuantiles and AllFiveQuantiles), the maximum distance between either of these distributions and our original parameterisation is circa 4%. Even if we assess calibration by placing an inflation factor on the median estimate (BetaOuterQuantiles and BetaMidQuantiles) we still do not generate massively different distributions. The maximum distance here relative to the original model is circa. 9%. If we discard the tails and apply a Kolmogorov-Smirnov (KS) test to the bulk of the distributions (samples in the 25th-75th percentiles), these numbers drop further to $< 2\%$ and $< 7\%$ respectively.

Extending this further to consider all of the target variables within this study (Table.7.3 and Fig.A.3-A.19 in the appendix), we can see that this distributional similarity is very consistent across variables. Indeed the maximum difference between any different parameterisation choice and our original model in any target variable at any point in the c.d.f is circa 15%. This is an incredibly small difference given the substantively different data sets, and calibration parameterisation choices used to generate these distributions. Thus, for the Arkansas study, we have seen posterior consistency when considering different parameterisations of the calibration

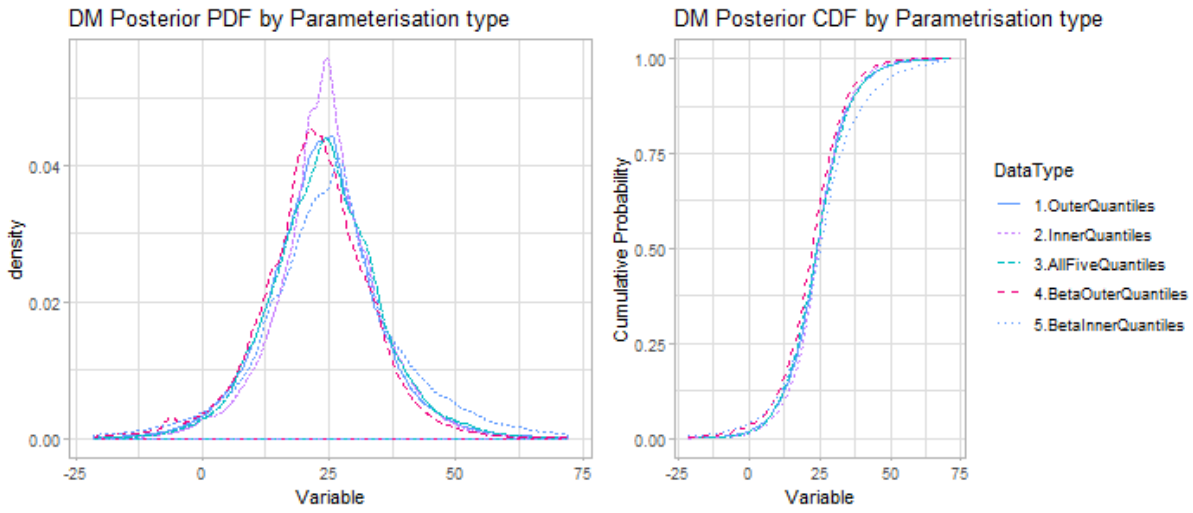


Figure 7.4: Comparison of the posterior DM distributions by parameterisation type for target variable 10 in the Arkansas study. The choice of parameterisation has had little impact to the overall uncertainty.

model and number of elicited quantiles being modelled.

Whilst not substantial, it is worth highlighting there is an appreciable differentiator for one of the parameterisations chosen. In 15 of the tested variables the BetaInnerQuantiles approach created the narrowest of the posterior distributions. This is significantly more than expected by chance. Even though the magnitude of difference seen here versus other parameterisations is small, the impact of this could be important under specific decision rules.

This is only a single study, and significant empirical analysis would be required to demonstrate that the consistency seen across parameterisations applies consistently. However, it is a very appealing result nonetheless. Firstly, this suggests the modelling approach we have outlined is robust and relatively consistent regardless of the quantile choice made. This builds confidence that the model is identifying underlying behaviour without being overly sensitive to our input choices. Secondly, if our choice of quantiles does not impact the robustness of the output, DMs can make a choice prior to elicitation on which quantiles they would like their experts to inform on. The cost of an elicitation exercise increases as the number of elicited data points increases. Eliciting more invariably takes more time from both facilitators and experts. Thus, if this model gives you consistently similar results regardless of whether you elicit three or five quantiles, you can make your elicitation exercise more effective by eliciting less, without loss on the output. Another way this behaviour could be beneficial is that experts may favour elicitation in different

Target Variable	Inner Quantiles	All-Five Quantiles	Beta Outer-Quantiles	Beta Mid-Quantiles
Variable 1	0.1004	0.0305	0.0456	0.0907
Variable 2	0.079	0.046	0.0631	0.0392
Variable 3	0.0586	0.0364	0.0473	0.0966
Variable 4	0.0378	0.0695	0.0767	0.067
Variable 5	0.0386	0.0311	0.0384	0.0838
Variable 6	0.0927	0.0308	0.0572	0.0963
Variable 7	0.1088	0.038	0.0142	0.0476
Variable 8	0.0553	0.019	0.0496	0.0759
Variable 9	0.0841	0.024	0.0336	0.0698
Variable 10	0.0395	0.0248	0.0852	0.0791
Variable 11	0.1305	0.0291	0.0142	0.0305
Variable 12	0.1589	0.0148	0.0167	0.0936
Variable 13	0.0839	0.0493	0.0466	0.1195
Variable 16	0.1595	0.0404	0.0395	0.1027
Variable 17	0.1258	0.0214	0.018	0.0799
Variable 18	0.0892	0.0406	0.0422	0.0889
Variable 19	0.0915	0.0348	0.0631	0.0945
Variable 20	0.0801	0.0259	0.0117	0.0953

Table 7.3: D Statistic from a Kolmogorov-Smirnov test on the impact of quantile parameterisation across all target variables in the Arkansas study.

ways, i.e. some experts wish to give their judgements on outer quantiles and others on inner quantiles. Given the model provides consistent outputs, a facilitator could tailor the elicitation exercise to each individual expert and then place these mixed quantiles into the model. The model could then adjust and standardise internally as required without loss of precision. If further testing discovers that these results do not generalise then running cross-validation with these different quantile choices and evaluating performance could offer a route to defining best practice for the quantiles to elicit and the appropriate inflation/bias factors to use.

7.1.3 All-in-one-method

The full modelling method outlined so far relies on two steps, a homogeneity group assignment step, followed by a combined calibration/aggregation step. In the interest of moving to a fully Bayesian method, an all-in-one method which links these three elements completely is desired. As described earlier, one approach to this is to consider mixture models for clustering over the seed variable space. We outline such a model and apply it to the Arkansas study here for two reasons. Firstly to assess how reasonable an approximation the two-step method is and secondly to address the suitability and challenges of modelling in this way.

Rather than use a standard mixture model to do our clustering, we utilise a Dirichlet process mixture model. A Dirichlet process is used here as not only the expert assignments but also the number of homogeneity groups may be unknown to the DM. The non-parametric nature of the Dirichlet process mixture model allows the DM to simply define a tuning parameter, o, based on an expected maximum number of homogeneity groups. The identified number of groups is then an output of the model.

The Dirichlet process mixture can be described by the following generative process (Chen et al. [2015]).

1. Generate a set of mixing weights, ν where $\nu = \{\nu_k\}_{k\in 1:|E|}$ according to a stick breaking process dependent on tuning parameter o.

2. Generate a set of parameters θ where $\theta = \{\theta_k\}_{k\in 1:|E|}$ for each potential cluster k, according to a prior distribution with parameters θ_0.

3. For each observation in the seed variable space $\mathbf{Y}_e$, assign a component label, h, according to the mixing proportion ν.

4. Generate $\mathbf{Y}_e$ according to the h^{th} component of θ.

When the model is a mixture of Gaussians, θ is defined as a mean and a precision matrix and the prior distribution on θ is modelled as a normal-Wishart. The model can be written algebraically as:

$$\mathbf{Y}_e \sim \mathcal{N}(\theta_h) \tag{7.2}$$

$$h \sim Cat(\nu) \tag{7.3}$$

$$\theta_k \sim Normal - Wishart(\theta_0) \tag{7.4}$$

$$\nu \sim GEM(o) \tag{7.5}$$

Where GEM denotes the stick break process (Sethuraman [1994]). The DAG of this process is outlined in Fig.7.5.

The DPMM can be integrated into the linked calibration/aggregation model by utilising the homogeneity group assignment h for expert e in the calibration and aggregation element, when $\mathbf{Y}_e$ is assigned to cluster h. Rather than being calculated *a priori*, clusterings are then defined at each step within the global MCMC.

We compare this all-in-one method to the two step method, with hierarchical clustering, outlined earlier by generating posterior distributions for all target variables in the Arkansas data. Running the all-in-one method takes considerably longer. On the same machine, modelling takes approximately 10 times as long.

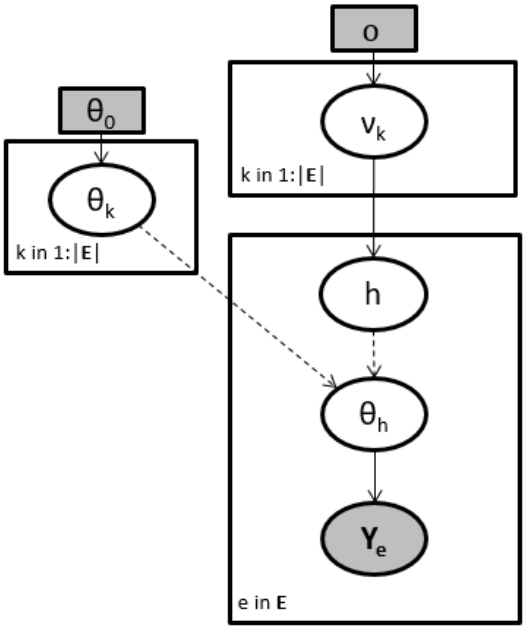


Figure 7.5: Dirichlet process mixture model for homogeneity group definition. Experts are points in the space $\in \mathbb{R}^{|\mathbf{Y}|}$ and are clustered utilising a mixture model (typically Gaussian).

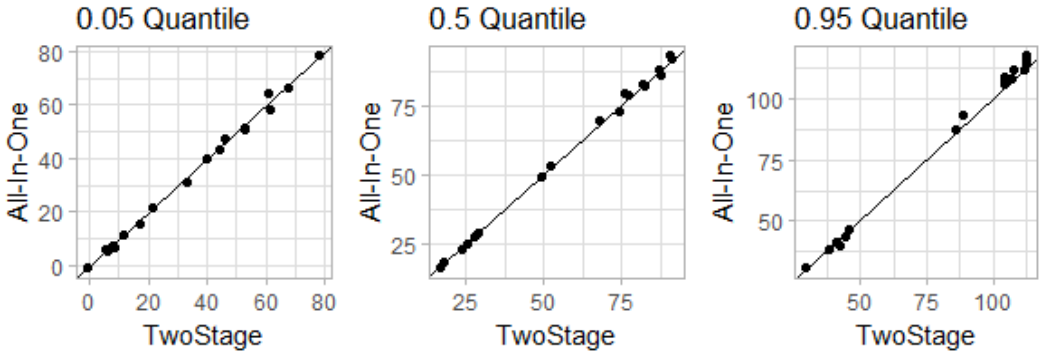


Figure 7.6: Comparison of the posterior DM quantiles between the one-stage and two-stage method.

Final results obtained using the two methods were extremely similar. Fig. 7.6 outlines the relationship between the quantiles of the final posterior distribution for all of the target variables in both the all-in-one method and the two-stage method. In each case the line x=y is plotted. Across all three variables we see very tight clustering of the results around these lines. The DM posterior has been largely invariant to the change in methodology. This is reassuring as it further builds the case for the robustness of the method, whilst providing options for analysts on how to perform the modelling based on both practicality (time needed to model, availability of data) and theory (modelling all-in-one allows us to avoid having to reuse the data).

The mixture model for clustering will work best when we have studies with a significant number of experts and seed/target variables in them, as we now have a very tangible increase in the number of model parameters. Methods for operating on a large scale are becoming more relevant due to mass participation expert judgement studies conducted over the internet (French [2011]), however, these are still the minority. As such, for most studies today, often the clustering space will be very sparse and potentially high dimensional. In such cases, principal component analysis (PCA) could be run, here as part of the clustering process itself, to reduce dimensionality ahead of cluster identification if an all-in-one method is considered and convergence is an issue. The analysis here however, suggests that the two-step method is a reasonable approximation and will likely be more appropriate in the short term. The issues with an all-in-one method overall are significant increase in model complexity (and consequently stability), the need for more data and risk that final posterior distributions become exceptionally diffuse with the integration of more areas of uncertainty, thereby reducing the value for the decision maker.

Whilst costs of this all-in-one method may outweigh the theoretical benefits today, as bigger studies are undertaken this balance is likely to shift. Application of this method to a mass participation expert judgement study is likely to test the enhanced efficacy it can bring. Given the scale/cost required to implement mass-participation studies, it would probably be best to integrate this test into a study conducted for other purposes, rather than designing a test study explicitly. This is left to further research.

7.1.4 CWD

The CWD study was outside of the health insurance domain, instead looking at the transmission risks for chronic wasting disease from deer to humans. The study was comprised of 10 seed variables and 13 target variables. With 14 participants,

this study had significantly more experts present than the Arkansas study. These experts were separated into 5 homogeneity groups by the model. Three experts (1,4 and 10) were placed in individual groups as they demonstrated consistently different behaviour over the seed variables than their counterparts. The model breaks the remaining experts into the following two sets, {2,6,8,9,11,12} and {3,5,7,13,14}. The separation of the individual groups is evident in a simple PCA plot of the first two components of the seed variable space, demonstrated in Fig. 7.7. Even within just these two components over half (∼54%) of the variability in the seed variable space can be explained.

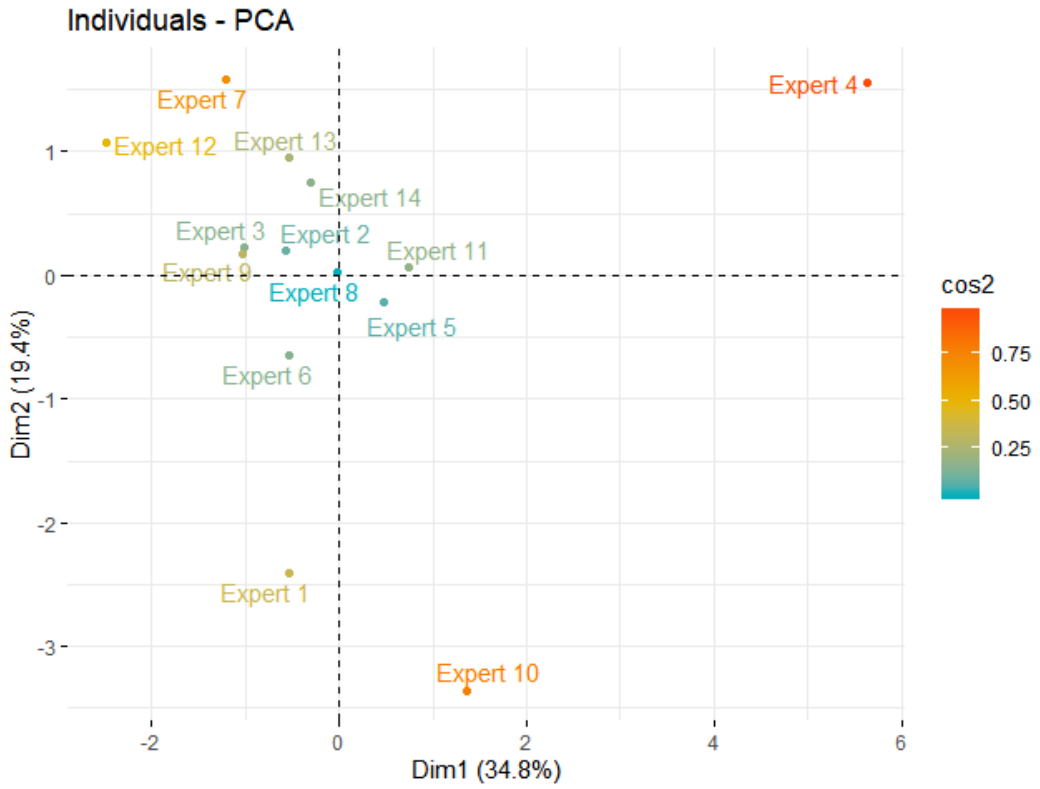


Figure 7.7: PCA plot of first two components for the seed variable space. Note the separation of experts 1,4 and 10 from the bulk of the remainder. These experts resultantly sit in their own homogeneity groups.

Whilst the three groups of individual experts are visible from the PCA plot of the first two principal components, there is no apparent logical seperation of the remaining experts. A split into two further groups is not readily visible. The third principal component of the PCA captures a further 15.4% of the variance within the seed variable space. Plotting the third principal component pairwise vs. the first,

the groups produced by the model become readily apparent (Fig.7.8).

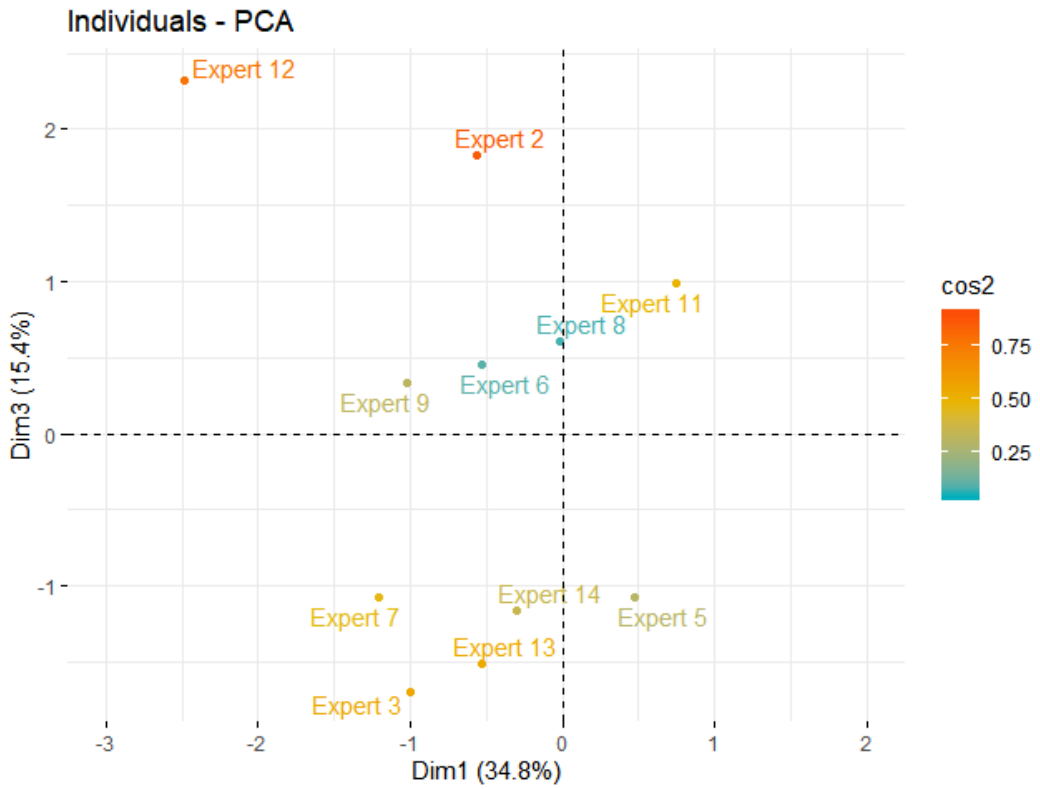


Figure 7.8: PCA plot of first and third components for the seed variable space. Two further expert clusterings are visible. One cluster consisting of experts, 2,6,8,9,11 and 12, all of whom sit above the line y=0. One further cluster with experts 3,5,7,13 and 14 is visible with members sitting below the line y=0. Experts 1,4 and 10 have been removed from this plot to reduce clutter.

Whilst the number of experts, the number of questions and the domain of the study in this case are very different to the Arkansas case, the statistical accuracy of the experts is similar. Here over 51% of judgements sit outside of the 5th to 95th quantile bounds. In the most extreme cases, 2 experts had 80% of their judgements outside of these bounds. This level of miscalibration suggests systemic overconfidence. Thus as anticipated in the majority of cases the Bayesian model has significantly broader tails across the aggregate target variables than the two opinion pooling methods. This behaviour can be seen in a plot of the distributions for each target variable, captured in Fig. 7.9. Table.A.1 in the appendix outlines the standard quantiles, by target variable, associated with each of these distributions.

One case in this study where we see different behaviour to any of the variables in the Arkansas study, is where experts predict significant imbalance in the upper

Figure 7.9: Final distributions for all target variables in the CWD study. The Bayesian decision maker (blue) consistently demonstrates a wider support than the PWDM (green) or the EWDM (pink). The Bayesian model is also always unimodal, emphasising the underlying consistency in the estimations.

and lower uncertainty. Variable 10 in CWD study, outlined in Fig.7.10, which looks at the link between wild and farmed deer, is a good example. The distance between the 50th percentile and the 95 percentile in the final DM estimation is 2%-3% across all three model types (EWDM, PWDM and Bayesian DM). This compares to up to 57.5% in the lower half of the distribution (between the 5th percentile and the 50th percentile). For this type of variable, the EWDM will often have uniform probability all the way out to the limits because there is a single expert who has had this extreme judgement. The PWDM may not include this expert in the final aggregation

(or significantly down weight them) and therefore does not recognise this tail. The Bayesian p.d.f however maintains the potential for the low value, but decays much more rapidly than the EWDM. This is due both to the chosen parameterisation and because the expert with the extreme perspective (expert 12) sits within a broader homogeneity group, this will effectively down-weight the effect of this expert in the model. This is an interesting counterpoint to the Arkansas study, in which the grouping up-weighted the differentiated view as that individual was grouped alone, here the view is included but down-weighted as their perspective is grouped with many others that are less extreme. This leads to a lower risk profile for the Bayesian DM which sits between the EWDM and the PWDM. The final result, in Fig.7.10 is that the Bayesian model acknowledges the additional lower uncertainty identified by a small subset of the experts (and highlighted in the EWDM) whilst maintaining the mass closer to the bulk of the estimations, similar to the PWDM. In these cases this behaviour can outweigh the impact of recalibration due to overconfidence and lead to tighter distributions than the EWDM in one tail.

As is evident across the studies analysed so far, it is very common to see overconfidence (as defined by low statistical accuracy denoting too narrow bounds) in expert judgement studies.

This cross-domain tendency for experts to be overconfident should give DMs pause for thought. Recalibration of expert opinions is a controversial subject, and there are certainly contexts when it is unwise; I would argue however that a DM should not neglect this critical information when assessing their belief in light of the elicited judgements. This would lead such a DM to a Bayesian model similar to the one outlined here.

Interestingly, also evident amongst a number of the final posterior distributions in the CWD study is the tendency for the Bayesian Model to have a mode which sits somewhere between the EWDM and PWDM model. Variable 1,6,7 and 9 are examples of this. It is hypothesised that this is more likely to occur as the number of experts increases. In such situations the PWDM will typically have a broader number of experts with non-trivial weights, thereby representing a mixture of many well-calibrated individuals, which is conceptually very similar to the Bayesian model. The over shoot percentages defined as part of the Cooke's Classical PWDM are analogous to the tails on the Bayesian model and it is conjectured, if these were relaxed further than the 10% that is commonly used today, then, as the number of experts were to increase, there is the potential for convergence between the posterior distributions for the PWDM and the Bayesian model.

7.1.5 Effusive eruption

Application to volcanology

Following the eruption of the Icelandic volcano, Eyjafjallajökull, in 2010 a scientific emergency group (SAGE) was appointed by the UK government (Loughlin

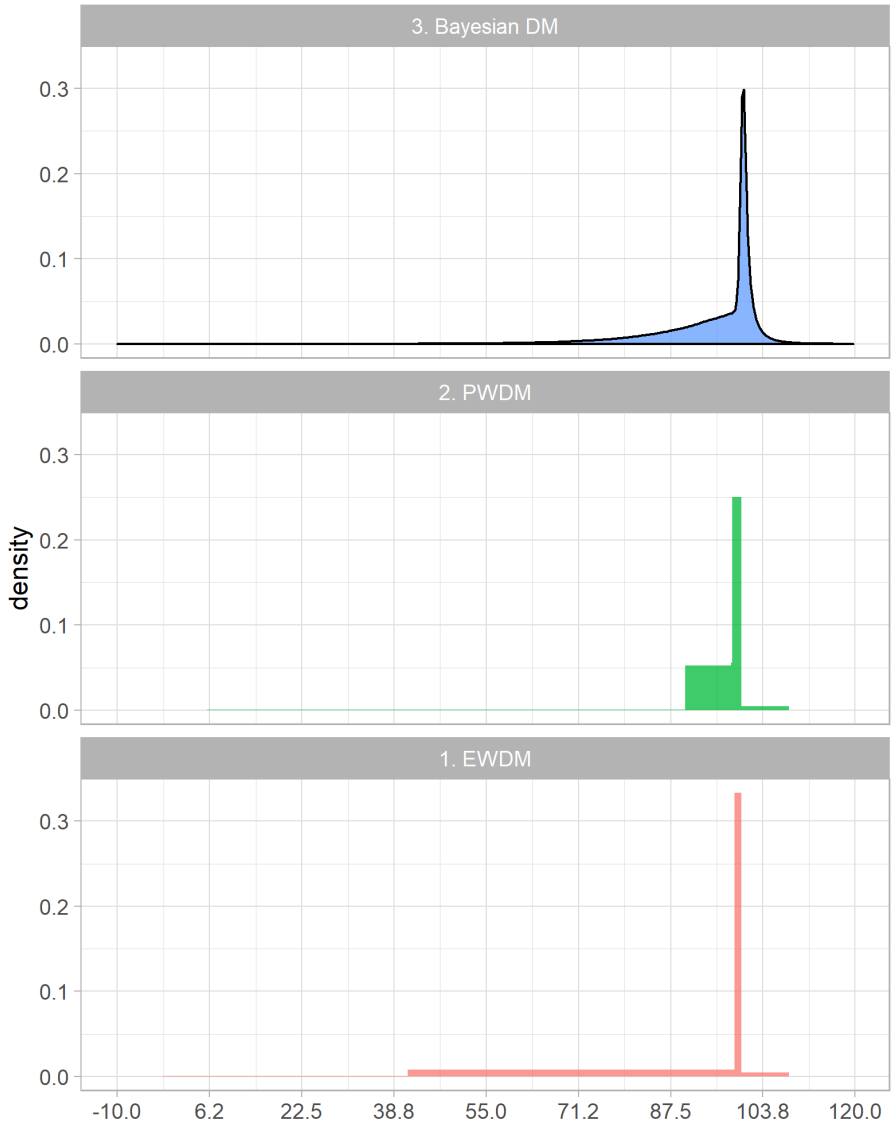


Figure 7.10: Final distributions for target variable 10 in the CWD study. The Bayesian decision maker (blue) lower portion of the pdf decays much quicker than the EWDM (pink). This leads to a posterior expectation on the tracked quantiles between the EWDM and the PWDM (green).

et al. [2012]). The group assessed future eruption scenarios that may impact the UK, and volcanic eruptions were subsequently added to the UK National Risk Register. A key scenario considered by the group was the eruption of the Grimsvötn volcano (commonly known as the Laki Eruption due to its presence within the Laki crater) which occurred in 1783-84. This volcano had a huge impact on Europe, particularly in Iceland where 60% of the grazing livestock died (predominantly by Flourosis) and 20% of the Icelandic population were also killed as a result of illness, famine and environmental stress. This eruption was considered to represent a "reasonable worst case scenario" for future eruptions.

Risk to the UK from such a scenario recurring would be in the form of volcanic gases, acid rain, aerosols, and acid deposition. These factors can have significant environmental impact (due to deposition on vegetation, buildings and potential impact to groundwater), or impact on transport, particularly aviation (as we saw with Eyjafjallajökull), where Suplhur dioxide and sulphuric acid can cause damage to airframes and turbines, engine corrosion, or put crew and passengers at risk of exposure. To model this complexity, meteorological (weather and atmospheric transport) models, in addition to chemistry models, are considered. In order to support this modelling and determine a set of prior values for some of the source characteristics an expert judgement study was conducted in 2012 (Loughlin et al. [2012]). This study followed an earlier study conducted on the same topic in 2010 (EFSA [2010]).

Structurally, the elicitation was conducted with 14 multidisciplinary experts. Experts were from academia, research institutes and other institutes with operational responsibilities. These experts were able to cover all of the modelling fields described earlier (meteorology, atmospheric dispersion, chemistry) in addition to specific volcanology expertise. Quantative responses were captured for 8 seed variables, alongside 28 target variables (22 volcanological in nature and 6 related to plume chemical processes). Not all questions were answered by all experts, with number of responses for each variable coming from between 5 and all 14 experts. For comparisons between the Bayesian framework and the classical performance weighted model, we shall consider only the 10 target variables which were responded to by all experts and for which details are captured within the Delft database Cooke and Goossens [2008].

Seed variables experts were asked to quantify were related to the historic Laki eruption, e.g.

- What was the area of the Laki Lava flows in km^2?

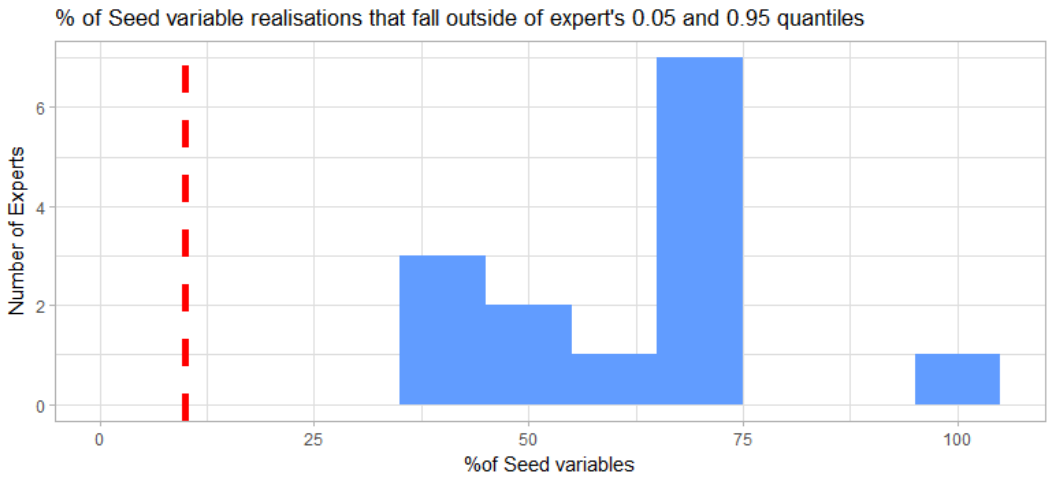


Figure 7.11: Across seed variables within the Laki effusive eruption scenario study, experts demonstrate significant overconfidence in judgements. All experts have more variables outside of the 0.05 and 0.95 quantiles than would be expected for high statistical accuracy. Red line indicates the expected % of variables for a perfectly calibrated expert.

- What was the estimated production of Laki in CO_2 megatonnes?

With true realisations of 500km^2 and 349 megatonnes respectively. An example target variable question was:

- What is the likelihood that in the next Laki-like eruption there is an episode which releases 10 times more SO_2 on the same timescale as the peak eruption episode during Laki?

With other questions similarly linked back to the Laki eruption. This link is important as it helps ensure that the seed variables are truly representative of the target variables and are thus suitable for use within the recalibration exercise inherent within the Bayesian model (and likewise for appropriate performance weighting in the Classical model).

Across the total 112 seed variable estimations, if we were to *a priori* assume that experts were well calibrated/statistically accurate, we should expect to see 11-12 of the seed variable realisations sitting outside the range given by the 0.05 and 0.95 quantiles provided by the experts. Individual experts would expect to have no more than one judgement where the true realisation sits outside of these bounds. In practice actually 64% (72) of the true realisations fell outside of the 90th percentile bounds given by the experts. For individual experts, between 37.5% and 100% of realisations fell outside of the confidence bounds given (Fig.7.11).

Running the Classical model over this dataset results in three experts getting a weighting (expert 10 - 53%, expert 14 - 31%, expert 12 - 16%) and eleven experts being removed from the final PWDM optimised DM quantile calculation altogether.[2] To compare the impact of this to the Bayesian framework, we can first consider the homogeneity groups that are created as a result of the first step within the model, the clustering exercise. Running this process identifies five core homogeneity groups within the expert pool (Fig.7.12), of which, two are formed of a single expert and three are each of four experts. The two experts who are grouped within their own pools have done so as a direct result of a significant divergence in judgement between themselves and the remaining groups as it pertains to the seed variables. Thus, the Bayesian model identifies that there is the potential for discrepancy in opinion on the target variables that should be considered and upweights these individuals relative to their peers. In this way the Bayesian model is capturing the diversity of thinking across the experts. Please note, at this stage no judgements have been recalibrated, thus we do not yet know whether this diversity is a result of different mental models by these experts or due to miscalibration. The recalibration exercise ensures that experts are well calibrated before aggregation, thus we minimise the risk of simply up-weighting a "poor" forecaster. Expert judgements are subsequently passed through the distribution fitting, recalibration and aggregation processes described earlier to create a single DM posterior distribution.

Before getting on to discussions regarding the uncertainty bounds provided by the Bayesian/Cooke's models, it is first interesting to assess differences between the posterior median for the Bayesian model and Cooke's optimised DM's 0.5 quantile. In many contexts, final decision makers will look to a point estimate from which to base their next best action. As the Bayesian model is trying to consider both the consensus and diversity in opinion, the hierarchical nature enforces uni-modality in the posterior distribution. This posterior mode (which, due to the parameterisation used, will be located at the median) reflects the most likely single value a DM would use to represent a point estimate, as long as the information in multimodal opinions is not lost. Whilst we recognise that ignoring uncertainty in this way is counter to the goals of risk management for which SEJ is typically employed, the use of point estimates is a common decision making reality and thus worth assessing.

Fig.7.13 outlines the final DM uncertainty ranges for each of the 10 target variables and each of the ascribed models[3]. It is important to note that across all

[2]Please note. whilst not included in the final calculation of the quantiles within Cooke's methods optimised DM, these experts assessments are still involved in determining the intrinsic range of the random variables.

[3]Table.A.2 and Fig.A.20 in the Supplementary material outline the same responses but also

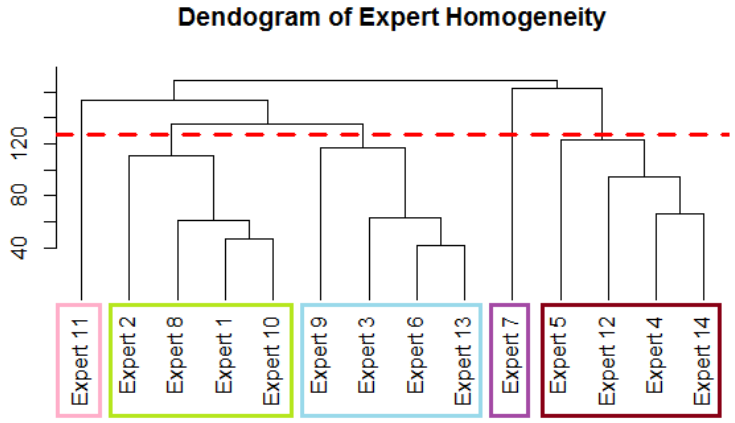


Figure 7.12: Experts are clustered into five homogeneity groups (coloured boxes) demonstrated by a horizontal cut on the dendogram (red line). These 5 groupings, based on seed variable responses, cluster the experts into three groups of four individuals and two outliers (expert 11 and 7) who sit within their own homogeneity groups as they regularly offer differing opinions to the other experts.

of the distributions the median for the Bayesian model sits within the uncertainty bounds of the PWDM. This is reassuring. Given the extent to which the PWDM has been utilised in practical studies, if there were fundamental concerns on this number these are likely to have been surfaced before. This suggests that a DM considering either model is not likely to make a significantly different decision based on the expected value alone. There is a noticeable difference however in the ranges given by the two models. The PWDM has consistently narrower bounds than the Bayesian DM. This is as we would expect and is caused by two predominant factors. Firstly, the PWDM selection criteria, by design, is optimising for statistical accuracy *and* information, and will often trade minor reductions in statistical accuracy for significantly improved information. This occurs due to the fact that information is a slower responding function than statistical accuracy. Secondly, the Bayesian DM is recalibrating the experts' judgements. Given that experts have demonstrated overconfidence, the DM has correspondingly increased uncertainty ranges.

Whilst the median and the uncertainty bounds themselves are critical, it is also important to understand the shape of the final DM distribution for each model.

include the results from considering an equal weighted linear opinion pool (EWDM), omitted initially for brevity and clarity. Comparisons vs. the EWDM are considered later when assessing the distributional forms as this provides greater clarity on the difference in modelling approaches than simply the uncertainty bounds alone. All EWDM results have been taken directly from the tool EXCALIBUR used to calculate the PWDM optimised DM.

Fig.7.14 and Fig.7.15 outline two such distributions, selected as these show different behaviours of the models. The equal weighted decision maker (EWDM) distribution has also been added to these slides for comparison. The equal weighted decision maker is the result of a linear opinion pool with identical weighting given to each expert.

Target variable 3 (Fig.7.14) demonstrates common behaviour of the Bayesian model vs. Cooke's performance weighted approach and the equal weighted decision maker. Notably a single modal point with a Gaussian decay in either direction (rather than multi-modality), narrower shoulders, and a broader support. One of the outlined aims of the Bayesian framework is to identify underlying consensus in

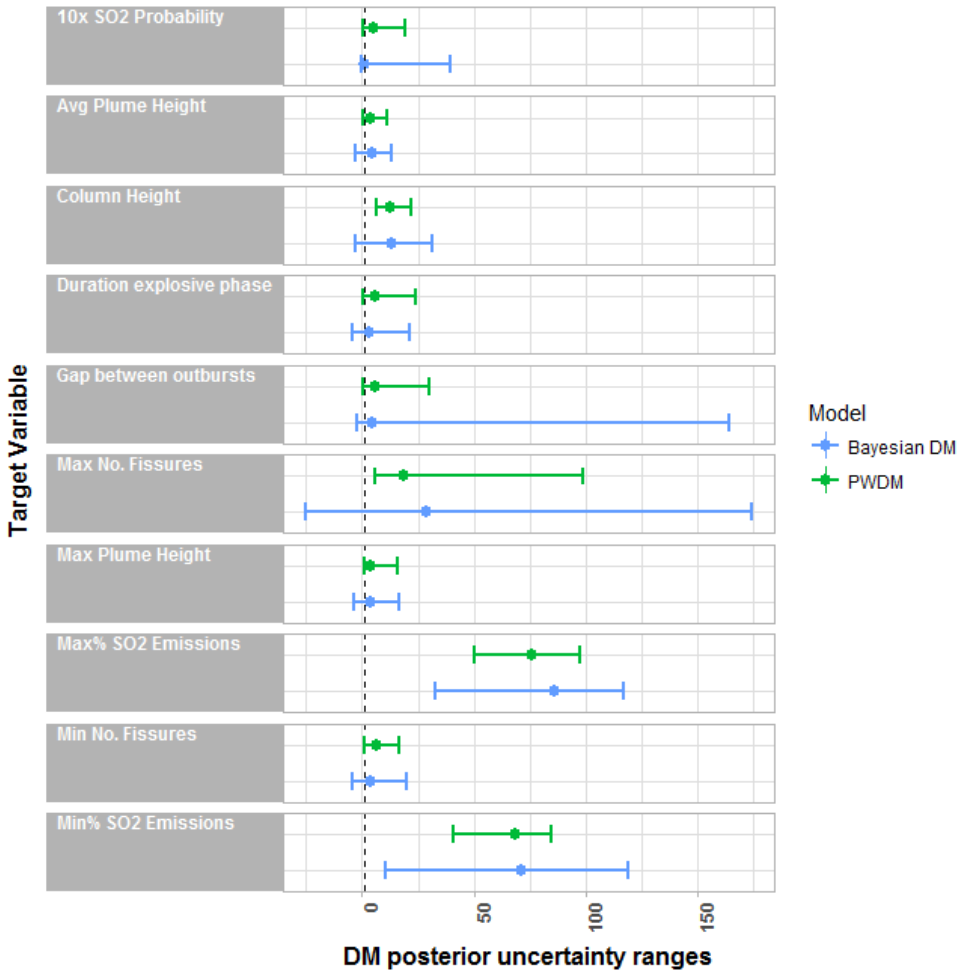


Figure 7.13: The Bayesian model produces median estimates similar to that of the PWDM and always within the PWDM uncertainty bounds. The Bayesian DM however suggests a higher level of underlying uncertainty.

opinion from the experts and thus this distributional shape is by design and reflects a starting assumption that the Supra-Bayesian DM's belief is of this form. Note, this is a decision to be made in setting up the model, and the framework is generic in nature to support many other possible parameterisations. Broader support is

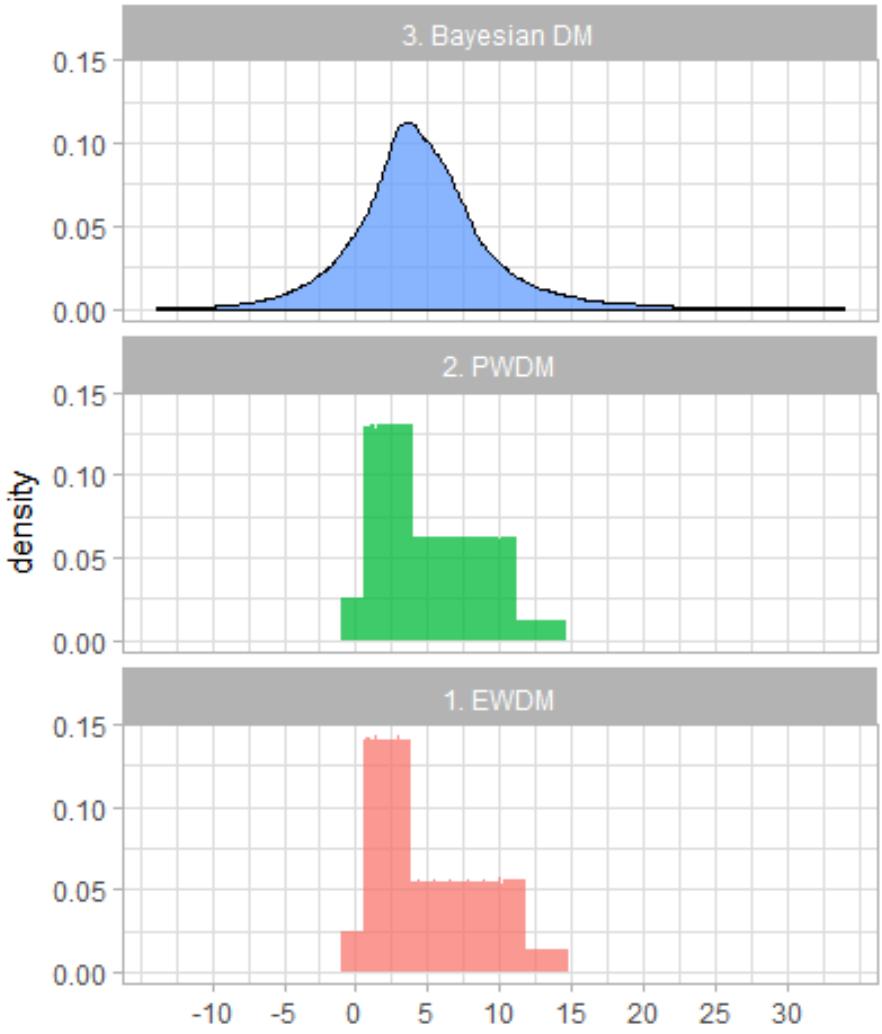


Figure 7.14: Final distributions for target variable 3: 'After the explosive phase (i.e. first days), what is the likely average sustained plume height for gases above the vent for the remainder of the active episode?' Note the similar shapes between the PWDM and EWDM distributions. The Bayesian model demonstrates a slightly higher modal point, and more uniform shape as it is focussed on the underlying consensus in opinion. Note also the larger support of the Bayesian DM as this recalibrates for overconfidence.

driven by the recalibration portion of the model and the overconfidence displayed by the experts on the seed variables. If experts were systemically under confident, we would expect to see narrower tails on the Bayesian model. In practice, as we have already seen overconfidence is much more common.

Target variable 10 (Fig.7.15) demonstrates a slightly different picture, here, once again the modal point of the Bayesian model is similar to that of the PWDM,

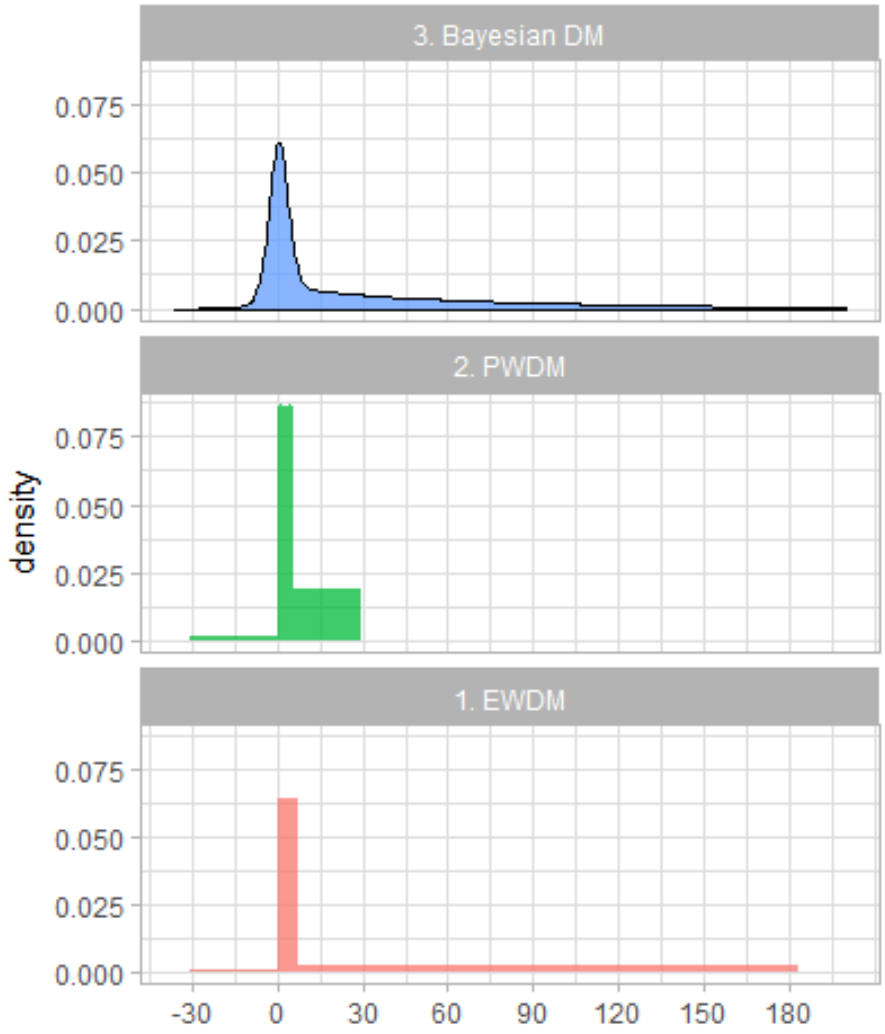


Figure 7.15: Final distributions for target variable 10: 'What is the typical gap between major gas outburst episodes?' Note, similar to the PWDM (although not the same extent) the Bayesian DM puts more of the distributional density in the region just greater than the modal point. Whilst the Bayesian model includes the greater range suggested by the EWDM it tapers off much faster.

and the unimodal shape is maintained. However, in this instance the EWDM is demonstrating a slightly different picture of the uncertainty. Both the PWDM and the Bayesian model put a very significant amount of the density at the modal point with little probability to a value below this and a limited but positive probability of more extreme values. The EWDM however has significantly less mass around the modal point, a larger probability of a lower realisation and more significant density in the upper tail. It is a positive sign that similar distributional shapes are visible here for the Bayesian and PWDM as there is no *a priori* reason that this should be the case and suggests they may both be pointing to similar underlying consensus between experts. This is similar to the behaviour seen in the earlier CWD study.

7.1.6 Invasions of bighead and silver carp in Lake Erie

Application to ecology

Forecasting the likelihood of, and damage caused by, invasive non indigenous species within many natural environments is difficult and poses a problem for those responsible for natural resource management. Similar to the context outlined before, often, the data necessary to build comprehensive decision models is incomplete and thus expert judgement can be used to supplement what data is available. A recent study (Wittmann et al. [2015], Wittmann et al. [2014], Zhang et al. [2016]) utilised expert judgement through the Classical model to forecast the impacts of Asian carp in Lake Erie. Asian carp is non indigenous and currently believed not be established within the lake. Assessments were made to quantify potential aspects of the Asian carp population (biomass, production and consumption) as well as impacts to existing fish species, in the instance that these carp become established within the lake. Establishment could occur as a result of release by humans, bait contamination or through waterway connections which are linked to currently established populations.

Structurally, the study comprised of 11 experts, each of whom was asked to assess 84 variables (20 seed and 64 target) within the elicitation questionnaire. In practice, for 5 of the seed variables, actual realisations did not become available and for 1 expert, only 11 of the seed questions were responded to. Hence within this analysis, to ensure consistency across modelling approaches, these 5 seed variables and this 1 expert have been removed, to leave 15 seed variables and 10 experts. Please note, this selection choice differs from the original paper in which the expert was left in the study but a further 4 seed variables were removed. Elicitations were made against the standard three quantiles (0.05,0.5,0.95).

The clustering of experts defined by seed variable responses suggested 3 core

homogeneity groups within the expert pool. The largest group consisted of 6 members (experts 3,4,7,8,9 and 10), the second group 3 members (experts 1, 2 and 5) and finally expert 6 sat within their own group as their responses consistently differed to those of the remaining groups, suggesting that they may be using a different set of reference data or mental models through which to base their judgements. Supporting a DM in identifying why a particular homogeneity grouping may have arisen could be difficult as the space in which the clustering is performed may be high dimensional (in this case 15 dimensions). Fig 7.16 outlines some key PCA outputs for this study and identifies the emergent groups in a visual way.

Similar to the prior studies outlined, overconfidence was common in the experts across the Lake Erie study. 47 of the 150 (31.3%), seed variable/judgement combinations had realisations sitting outside of the bounds given by the experts. Significantly more than the ∼15 (10%) we would have expected assuming the experts were all well calibrated. Unlike the effusive eruption example given earlier however, the range of calibration across the experts was broad. Expert 4 demonstrated strong statistical accuracy, only 1 of the realisations (7%) fell outside of the judgement bounds they gave. As expected this translates into significantly less recalibration within the Bayesian model. Expert 4 had the lowest recalibration parameters within the group. Classical model analysis of the study put all weighting to Expert 4, thereby effectively removing all other experts' judgements from the quantile aggregation within the PWDM optimised DM. Note, as before all experts are still included in the calculation of the intrinsic ranges.

The key finding of the original elicitation was that given the right starting condition, there is significant potential for the establishment of Asian carp within Lake Erie. In particular, they have the potential to achieve a biomass level similar to some already established fish species currently harvested commercially or recreationally (yellow perch, walleye, rainbow smelt and gizzard shad). These findings remain when considering the final posteriors proposed by the Bayesian model. The Bayesian model estimations of peak biomass levels for bighead and silver carp, in scenarios where they are the sole invader, suggest higher medians than those predicted by the EWDM but lower than those predicted by the PWDM for both bighead and silver carp (Table.7.4).

Uncertainty ranges within the Bayesian model are slightly narrower than either of the other two. Equilibrium biomass estimates in the same scenarios were lower than peak biomass levels but displayed consistent behaviour between models. The PWDM estimates that the median equilibrium values were approximately 1/3 of the peak value compared to approximately 1/2 in the Bayesian or EWDM models. In

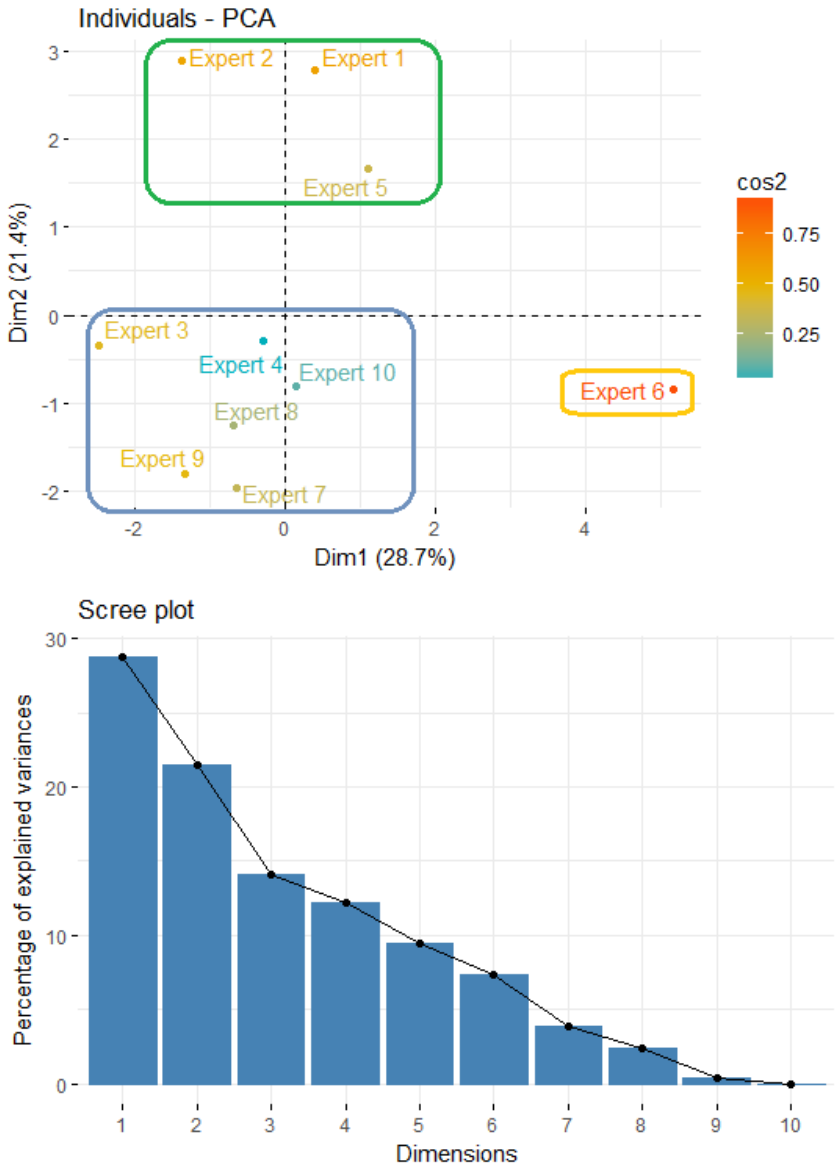


Figure 7.16: A scree plot of the Principal Component Analysis demonstrates that the first two identified components explain 50.1% of the variance across the original 15 dimensions within the seed variable space. When we isolate these two components and look at where the individual experts sit, the homogeneity groups identified by the model (coloured boxes) emerge. Expert 6 is separated from the remainder and thus is within their own homogeneity group, as they systemically give differentiated responses to the other experts. Note, groupings here are for visualisation purposes only, actual clustering occurs over the full set of seed variable dimensions.

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Peak biomass									
Bighead carp	0.0	2.4	17.2	1.6	8.9	25.9	0.7	4.2	13.0
Silver carp	0.0	2.3	17.0	1.6	8.8	25.9	0.7	4.1	11.9
Equilibrium biomass									
Bighead carp	0.0	1.2	9.1	0.4	3.0	12.2	0.3	2.0	6.2
Silver carp	0.0	1.1	8.0	0.4	3.0	12.2	0.3	2.3	6.8

Table 7.4: Biomass levels (t/km^2) predicted for the Lake Erie study sole invader scenario.

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Equilibrium biomass	0.0	2.2	12.3	0.4	3.0	12.2	0.6	3.6	10.4
Proportion Bighead carp	0.0	0.3	0.9	0.1	0.5	0.9	0.1	0.5	1

Table 7.5: Estimates for the Lake Erie joint invasion scenario.

the joint invasion scenario (where both bighead and silver carp are established) the Bayesian estimation of the equilibrium biomass was marginally higher than both the PWDM and the EWDM (Table.7.5). Mid quantile estimations of the proportion of the total biomass that is bighead carp in the joint invasion scenario was identical in the PWDM and Bayesian models and marginally higher than that of the EWDM (Table.7.5, Fig.7.17).

Overall, similar to what was seen in the preceding examples, the quantities of interest resulting from the Bayesian model do not vary significantly (where significance is defined as implying a radically different conclusion from the judgement data) from those of the PWDM, and in this case the EWDM. All models have suggested that there is significant potential for establishment of tangible biomass of these carp, in relation to existing fish populations, although each model has demonstrated a slightly different posterior distribution of the uncertainty as they emphasise different underlying elements of the judgements.

This is reassuring for a new model, such as the Bayesian framework, as existing models have been used and tested extensively. If radically different values had been found significant justification would be required.

7.1.7 Cross study median comparisons

In both this and the earlier examples, we have seen that the median estimate was similar in the performance weighted and Bayesian approaches. If we look at a broader subset of the Delft studies, specifically a subset where all variables are on a

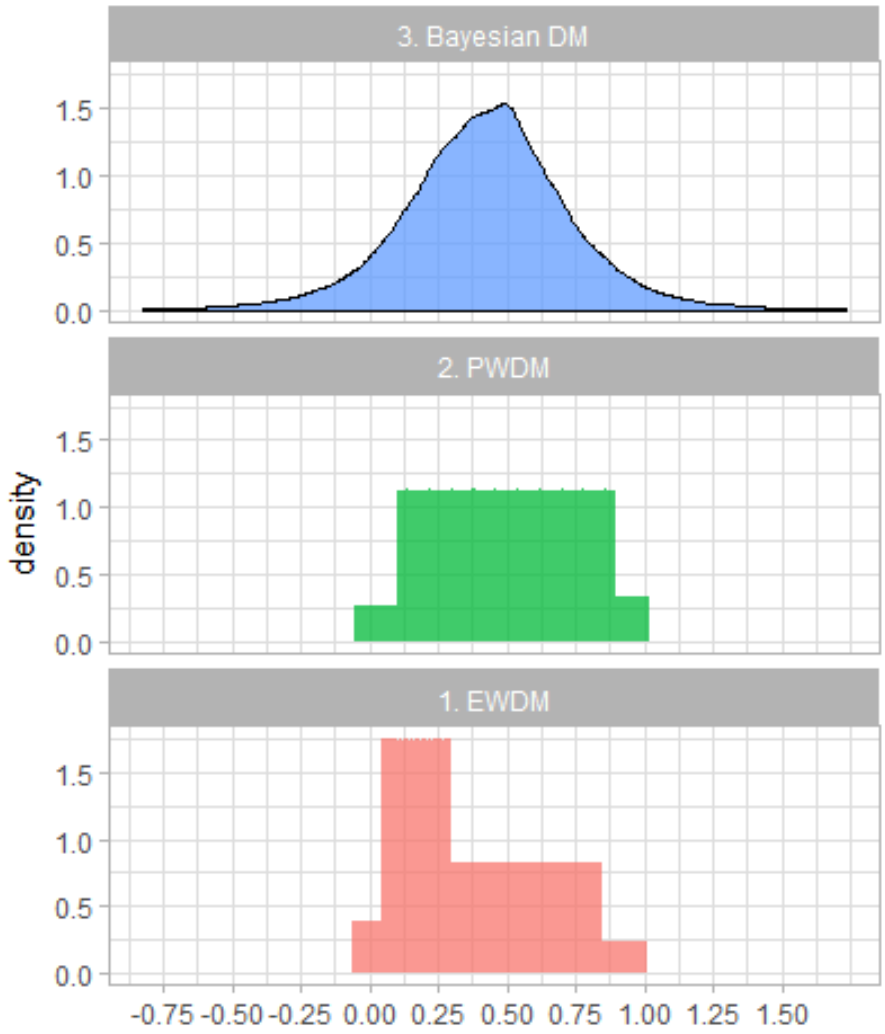


Figure 7.17: Final distributions for target variable 10 in the Lake Erie study: 'What is the proportion of the total biomass that is bighead within the joint invasion scenario?' The Bayesian model and the PWDM suggest a marginally higher proportion of the biomass will be the non indigenous bighead carp than EWDM predictions. Note the narrower shoulders and broader tails of the Bayesian model, consistent with other target variable estimations.

uniform scale to ensure the recalibration algorithm is applicable, we can assess the final decision maker medians for each of the target variables. In total there are 20 considered studies with 548 forecasted target variables. In Fig.7.18 we can see that this similarity between median estimates is true more broadly as there is a strong correlation (0.82) between the final median estimates of the two approaches (0.99,

removing outliers).

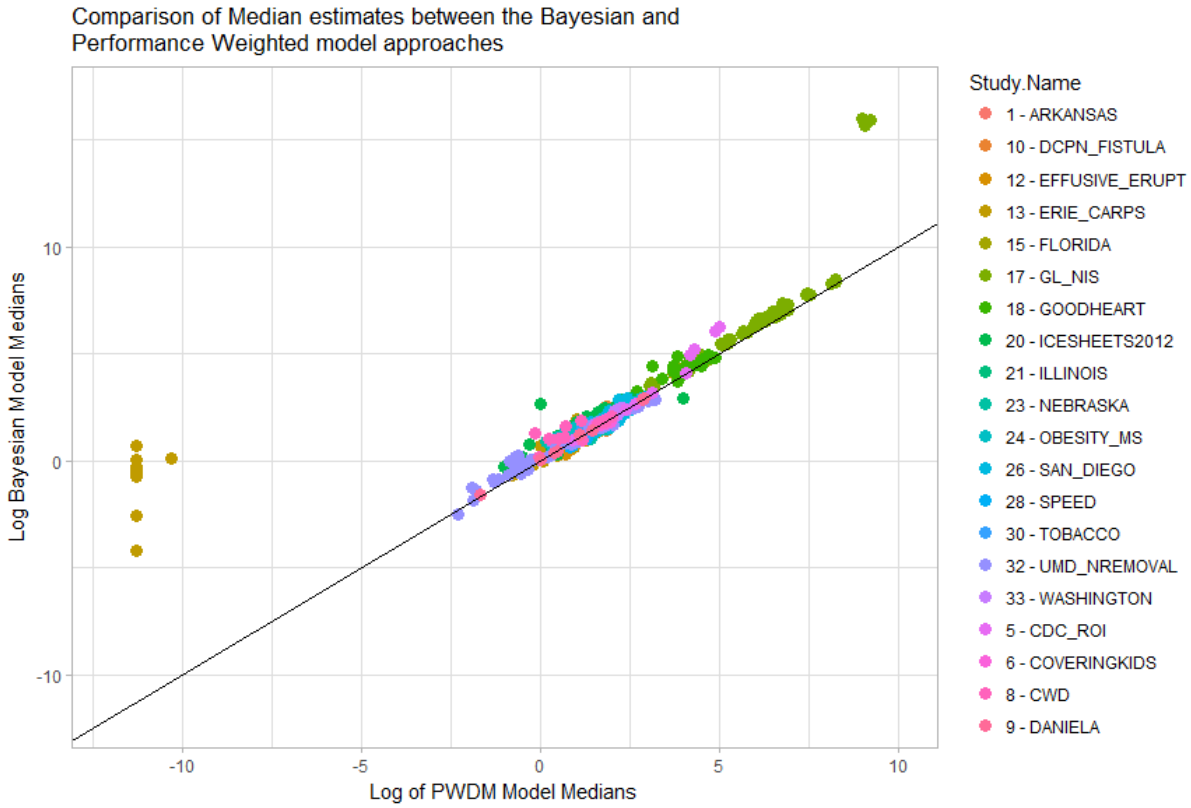


Figure 7.18: Median estimates from the final Decision Maker distributions are highly correlated between the Bayesian and performance weighted approaches across studies within the Delft database.

A log scale is used in Fig.7.18 to allow us to compare across studies. Whilst variables are on a consistent scale within a single study, they may be on very different scales in different studies. There are a small number of outliers at either end of the plot, from the Lake Erie study and the GL_NIS study, where the Bayesian model has a final median of a significantly different order of magnitude to the PWDM. Those at the lower end, from Lake Erie, have been driven by the fact that the performance weighting approach selected a single expert who had a value for these variables many orders of magnitude lower than some of their compatriots who were also included in the Bayesian aggregation. On the upper end, the discrepancy is driven by a small number of target variables within the GL_NIS study whose estimated values were many orders of magnitude higher than other target variables. This will have broken the constraint of scale uniformity from the recalibration process within the Bayesian model and potentially projected higher than realistic values here. Aside from this

small number of outliers however, there is broadly good consistency between approaches on this median value. Whilst important for decision makers, the median is not the only element that is being considered by those utilising the output of expert judgement studies, with the way that uncertainty is being expressed also of critical importance. A recent study on the impact of melting ice sheets, and the subsequent commentary papers, emphasised some of these considerations.

7.1.8 Ice sheets

Application to glaciology

Climate change is one of the major issues of the current age and as such is an area where strong scientific insight is fundamental to building the case for necessary political decision making and public behavioural change. Unsurprisingly, despite the wealth of geological datasets and sophisticated models, understanding the complexity in and predicting the outcome of many climate change problems relies heavily on expert judgement. Willy Aspinall, who performed the effusive eruptions elicitation study, alongside Jonathan Bamber, conducted a glaciological study to predict the impact of melting ice sheets, due to global warming, on rising sea levels (Bamber and Aspinall [2013]).

This research has been extensively cited and has positively contributed to the ongoing debate regarding the appropriate use of expert judgement within the geological community. One commentary that came from this paper de Vries and van de Wal [2015], and the subsequent discussion papers (Bamber et. al. [2016],de Vries and van de Wal [2016]), assessed and questioned a number of key elements of applying the Classical model in this context. In particular:

- The correct way of assessing the lack of consensus in the interpretation of post-processing the experts' answers.
- The reduction in the "effective" number of experts caused by the Classical model weighting process
- The choice of underlying distributions

Please note. One of the other topics raised in the commentary paper was regarding the choice of variables to elicit from the experts. In this case the primary elicited variables reflected the experts' predictions on the impact to sea level rise from three separate ice sheets (East Antarctic, West Antarctic and Greenland) and were then combined *post hoc* to create a total sea level rise estimate utilising a Monte Carlo model. Questions were raised whether this would accurately reflect

the experts' underlying belief of the final target variable as the choice of model can impact the total uncertainty bounds. Hence, it was suggested, the total sea level rise estimates should have been elicited explicitly. This is a question of study design and we will not tackle it here, except to comment that it is very common for expert judgement to be used both to make judgements on final decision making variables, or variables which are then inputs into a broader model. The choice selected here may be largely context dependent. One of the design elements of the Bayesian model (a fully parameterised posterior) is a support tool for DM/analysts utilising the output of expert judgement studies as priors in other models.

Many of these topics are not unique to glaciology and have been commented on elsewhere in the literature with respect to the Classical model. The Bayesian hierarchical model, by design, takes a philosophically different approach to each of these areas than the Classical model. Thus, whilst it will not address all of the comments posed in de Vries and van de Wal [2016] it would be interesting to consider how the application of the Bayesian aggregation model to the same data performs relative to the performance weighted approach.

In the effusive eruption and Lake Erie example we compared some of the forecasts for target variables however made no comment as to the validity of the final estimates nor how this varies between models. To be confident in any forecast a DM should have prior validation of a model's results. To this extent we shall not assess the two models over the target variables within the Ice Sheet studies as we have done before but will look for ways of assessing how well the models perform (in this context and the previous studies) using some cross validation techniques.

7.2 Cross-validation of Bayesian approach with Cooke's model

Applying the cross validation technique outlined earlier to the three studies just discussed and using the 80% subset rule results in 527 separate subsets and 1529 individual forecasts. Statistical accuracy and information scores for these forecasts are then calculated within R (Skronska [2014]). The R code was validated by taking a sample of these forecasts, rebuilding it from scratch within EXCALIBUR and ensuring consistency of the output. All numbers for Cooke's Classical model (PWDM, and EWDM when relevant) have been drawn directly from Eggstaff et al. [2014] supplementary material, kindly provided by Roger Cooke, to ensure consistency. Mean statistical accuracy scores, Fig.7.19, show that the Bayesian model (0.53 Effusive eruption, 0.54 Lake Erie, 0.57 Ice sheets) scored higher in each study than the PWDM

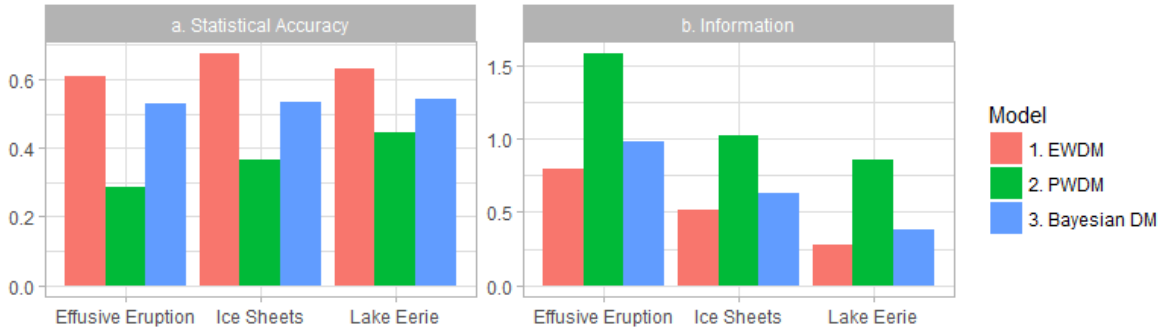


Figure 7.19: Arithmetic mean of the statistical accuracy and information scores for each tested model across the three studies previously discussed. The Bayesian model typically demonstrates better statistical accuracy but lower information than the PWDM, as we would expect. Perhaps surprisingly, given the broader support, in these studies the Bayesian model demonstrates higher informativeness than the EWDM. This is due to the Bayesian model having narrower shoulders. Please note that information is a relative measure and absolute informativeness numbers are not relevant cross studies and should only be considered across models within a single study.

(0.29, 0.45, 0.31). Conversely, Cooke's model (1.6, 0.85, 1.01) performed better than the Bayesian model (0.98, 0.38, 0.63) according to the information criteria outlined. This highlights exactly the behaviour we might expect to see, given the distributions we saw earlier, the fatter tails of the Bayesian model as a result of calibration, and the inherent trade off made within Cooke's model.

Perhaps more surprising is the performance relative to the EWDM, which has also been included in Fig.7.19 to provide another reference point. Across the studies outlined, the Bayesian model had higher information scores than the EWDM (0.80, 0.28, 0.52) but lower statistical accuracy scores (EWDM; 0.61, 0.63, 0.35). This may seem counter intuitive, as the Bayesian model has been specifically recalibrated, whereas the EWDM has not. The reason for this behaviour is the consensus focus that the Bayesian model has, rather than diversity which is emphasised in the EWDM approach. Despite the fatter tails, by looking for a consensus view, the Bayesian model typically has narrower shoulders than the EWDM, as we have seen in some of the earlier distributions e.g. Fig.7.14. Narrower shoulders are likely to reduce the statistical accuracy score but increase information. In this way, the Bayesian model is also trading off between statistical accuracy and information. If we were to only apply the recalibration component of the Bayesian framework and then aggregate utilising the EWDM, we should expect to see the highest statistical accuracy scores

but the lowest information of any of the models discussed so far.

Expanding the cross validation technique to the broader set of studies within the Delft database can help us ascertain whether we see the above behaviour consistently. As the recalibration within the Bayesian model cannot currently deal with variables on different scales a subset of 28 studies which were utilised by Eggstaff and within the Delft database were considered. In each considered study all of the variables were of similar order of magnitude. Study names align to those in the original paper. In total this equated to 2706 forecasted subsets and 6882 individual variable forecasts.

The Bayesian model outperformed the PWDM on statistical accuracy in 71.4% (20 out of the 28) studies based on the arithmetic mean (Fig.7.20). The PWDM outperformed the Bayesian model in mean informativeness in 93% (26 of the 28) of the cases. This is reassuring as it demonstrates that the model behaves consistently across studies relative to the PWDM and aligns with what we saw earlier. One of the studies (Study 9 - DANIELA) is clearly an outlier with an extremely low statistical accuracy and high information for the Bayesian model. This is because there was a convergence issue with this model, believed to be due to the combination of a low number of experts (4) and seed variables (7, given the hold out sample, only 5 of which would be included in each subset). Whilst more work is necessary to understand the impact of the number of seed variables to expert judgement models, performance weighting guidelines suggest that at least 10 seed variables are considered. Bayesian models with recalibration will similarly require minimum numbers to reach appropriate convergence which meaningfully reflects underlying expert bias.

The above analysis highlights that the Bayesian model and PWDM model are trading off between statistical accuracy and information to different degrees. The choice of which model to use in practice for a DM may depend on the context in which the study is being performed and the sensitivity of the decision they are making to either information or statistical accuracy. To get a better sense however, whether the trade off that the Bayesian model is making is reasonable, we can consider the combination score, as per Cooke's performance weighting method. Here, the statistical accuracy and information scores are multiplied together to give a combined score. This metric for cross validation is based on the same motivations that lie behind performance weighting. It is thus important to ensure that it is not biased towards a performance weighted DM. More research is needed to confirm this is the optimal unbiased cross validation approach. We note this challenge and agree that more work should be done to define a set of cross validation metrics and

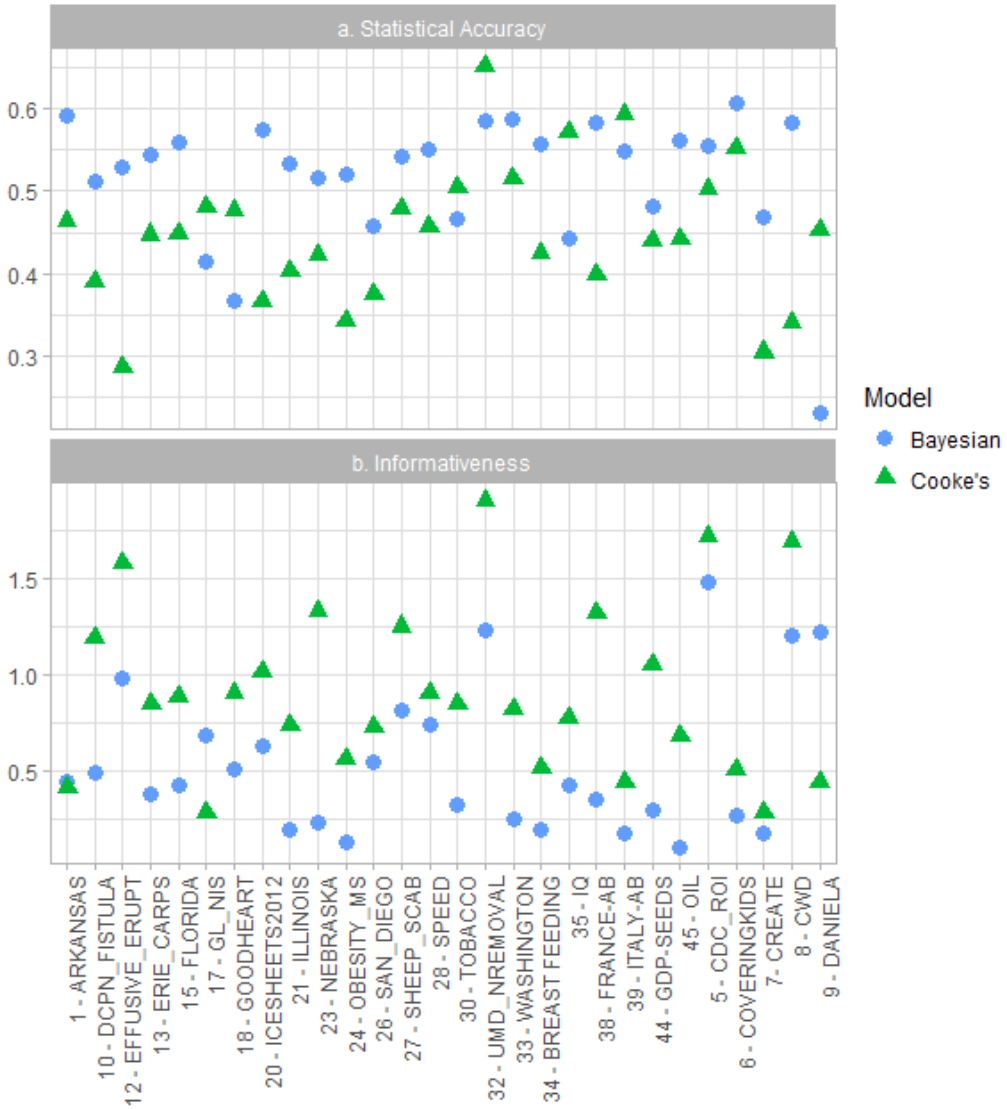


Figure 7.20: Statistical accuracy and information plots for each analysed study within the Delft database. The Bayesian model demonstrates consistently higher statistical accuracy than the PWDM but lower information scores.

processes that are independently ratified, model agnostic and applied consistently to such studies. In the short term however, this does remain the best available approach and gives us access to a body of knowledge built in the previous listed studies for comparison. Rather than considering the aggregate combined score, which may mask some of the underlying behaviour, we will consider the combined score of each forecasted subset for each study. Fig.7.21 plots the combined score of the PWDM vs. the Bayesian DM and an x=y line to help identify relative performance.

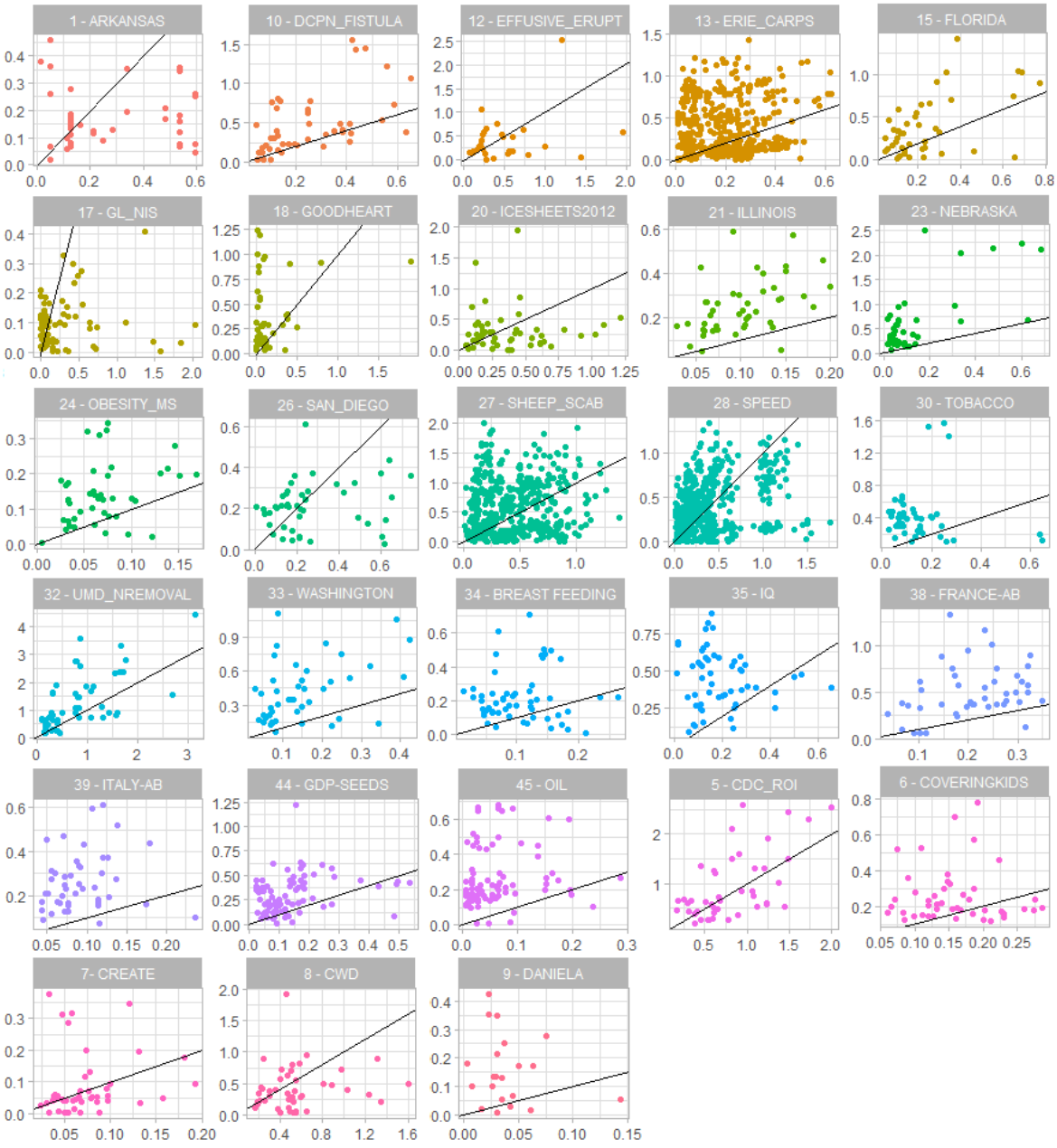


Figure 7.21: A plot of the combination scores for each analysed study subset. The performance weighted model (y-Axis) demonstrates higher combination scores than the Bayesian model (x-axis) as a significant mass of the points are above the x=y line. There are studies however, e.g. Study 20 (the Ice Sheets example), where the Bayesian model typically has higher combination scores.

This plot highlights a number of interesting elements about the performance of the two models across these subsets. Firstly, of key note, is that across many of the studies outlined, a significant portion of the subset forecasts sit above the line x=y (e.g. Study 23, or Study 35). This implies that for these studies the PWDM has outperformed the Bayesian model on aggregate whilst considering such a combination measure. Whilst this might appear disheartening for the Bayesian framework, it provides further evidence of the robustness of a performance weighted approach. On the positive side for the Bayesian framework, however, is that there are studies in which the mass of points have been more balanced (e.g. Study1, Study 27 and Study 28) and a few studies in which the Bayesian model appears to be a better predictor across the given subsets (e.g. Study 1, Study 20, Study 8). In fact there is only one study (Study 23) in which the Bayesian model did not outperform the PWDM on some subset of the seed variables when we consider a combination metric. In total across the 2706 subsets the Bayesian model outperformed the performance weighted model in approximately a third of cases (912) with the PWDM demonstrating higher combination scores in 1794 subsets. It is reassuring for the Bayesian approach that there is a substantial number of cases where the model can meet the aims of providing a consensus distribution which is fully paramaterised, whilst performing well against the PWDM when considering a combined statistical accuracy and information score. To be a fully viable model, however, more research is required to understand the drivers of what causes certain combinations to perform better in the Bayesian context than others. One potential option, originally posited in Hartley and French [2018], is that performance here could be linked to the number of experts/seed variables present within the study. This assessment is left for future research.

Chapter 8

Embedding the model into a broader elicitation framework

8.1 Overview

To embed Bayesian models as a key tool for DMs takes more than a mathematical framework. The modelling component is only one step within an SEJ study. Bayesian models need to be seamlessly integrated with the other process steps in order to make using them efficient and effective.

Significant work has already been conducted to optimise the definition, elicitation and documentation protocols of SEJ, (Cooke [1991], EFSA [2014], Hemming et al. [2018]). It does not make sense therefore to build a completely new elicitation paradigm solely for the purpose of introducing Bayesian models. The logical approach is to take one of the existing set of elicitation structures and replace the aggregation method currently utilised with the Bayesian method.

Bayesian methods typically analyse elements not included in other techniques, such as inter-expert correlation. Consequently they require different inputs. They also necessitate some pre-analysis decision making with regards to priors, which also needs to be included within the process steps taken. To this extent, whilst existing processes should form the basis of the elicitation framework within which a[1] Bayesian model sits, they will need to be adjusted to make it clear how Bayesian specific components should be included.

In addition to embedding a Bayesian model within a broader process, sup-

[1] *"a"* Bayesian model rather than *"the"* Bayesian model is described here as the process adjustments recommended are likely to be needed for any Bayesian approach. Whilst targeted to the model outlined within this thesis, they could be applied to other Bayesian modelling paradigms, such as the model outlined in Perälä et al. [2019], with only minor adjustment.

porting practitioners to run the analysis and play back information in a visual form real-time is also critical. All of the analysis conducted to date for the Bayesian model has relied on many lines of code and pieces of analytical software such as R, BUGs and Excel. Without better packaging, the methods for running the Bayesian model will become a hurdle to utilisation.

This chapter proposes a mechanism by which the Bayesian model outlined could be integrated into one of the existing elicitation protocols, specifically the IDEA framework (Hemming et al. [2018]). The chapter will start with an outline of the IDEA protocol and then describe the process steps necessary to embed our Bayesian model within it, noting some of the implications that Bayesian approaches have to processes and procedures more generally. Finally, the following chapter will describe a customised piece of software, the (B)ayesian (E)xpert (A)ggregation (M)odel or BEAM, built in R-Shiny, that allows the Bayesian model to be run efficiently within the updated IDEA process.

8.2 The benefit of mixed methods

Chapter 3 outlined the differences between *behavioural* and *mathematical* approaches to expert judgement combination. As briefly noted, in Section 3.2, there exists a third type of expert judgement protocol known as *mixed methods*.

In behavioural methods, there is significant face-to-face engagement between experts throughout the process. Consensus is reached through discussion. This allows for dialogue and debate on both the questions being asked and the answers to these questions. Experts' information, and perceived knowledge, can be shared and debated.

Mathematical methods are characterised by their intent to be explicit, auditable and neutral to key issues such as expert status. They typically have significantly less discussion between experts. This is normally limited to training and study briefing. Dialogue, it is perceived, increases dependence between experts and therefore artificially decrease the uncertainty stated.

Mixed method protocols are a balance between behavioural and mathematical concepts. They allow for sufficient dialogue to ensure there is common understanding of the questions being asked, but they do not seek for consensus through discussion. Mixed methods have some form of mathematical aggregation component and therefore retain the auditability that is so advantageous from these approaches. The most common mixed methods are characterised by multiple rounds of structured discussion with independent elicitation in between. Some versions of

the DELPHI process are arguably mixed methods, although as DELPHI is now an umbrella term for hundreds of individual methods, some are arguably more akin to behavioural methods and others to mathematical methods.

Whilst naturally a mathematical approach, Bayesian models can be embedded neatly into mixed method protocols. Behavioural approaches typically strive for a consensus on risk, whilst mathematical approaches attempt to highlight the range of diversity. One of the stated aims of our Bayesian approach is to support in the creation of a posterior which reflects both the consensus and the diversity. Discussion elements within mixed methods allow for opportunity to embed the formulation of some key inputs, such as priors, that Bayesian models need. It is important however, through the discussion elements in mixed methods to be careful with the impact on inter-expert dependence, we will discuss how these can be managed appropriately.

8.3 The IDEA protocol

The IDEA protocol was developed at the Centre of Excellence for Biosecurity Risk Analysis (CEBRA), at the University of Melbourne, (Hemming et al. [2018], Hanea [2017], Hanea [2018]). Similar to other structured processes for expert elicitation, the protocol is divided into three distinct sections, Pre-Elicitation, Elicitation and Post-Elicitation with explicit sub-components within each.

The name IDEA stems from the flow of key activities through these phases. The process encourages experts to *Investigate* the problem before providing a first set of initial estimates. This is followed by a period of *Discussion*, which can be either face to face or remotely, to clarify reasoning and assumptions. There is a second round of *Estimates* provided by each expert. Finally the opinions of the various experts are *Aggregated* together using some form of mathematical aggregation method, Fig.8.1.

The Pre-Elicitation component of IDEA is very similar to other structured methods, compromising of many of the activities outline in the Steering Group pre-elicitation phase of the EFSA approach. Namely, problem definition and framing, expert identification and expert training. One notable element of Pre-Elicitation within the IDEA framework however is a step of empirical validation. As highlighted in Cooke [1991], for expert judgement to be considered as a robust scientific discipline it must be subjected to quality assurance and testing in the same way as other empirical methods. Similar to Cooke's Classical method, expert validation and testing can be integrated into the IDEA framework by eliciting seed variables (Hemming et al. [2018]). The IDEA framework is the only broadly adopted struc-

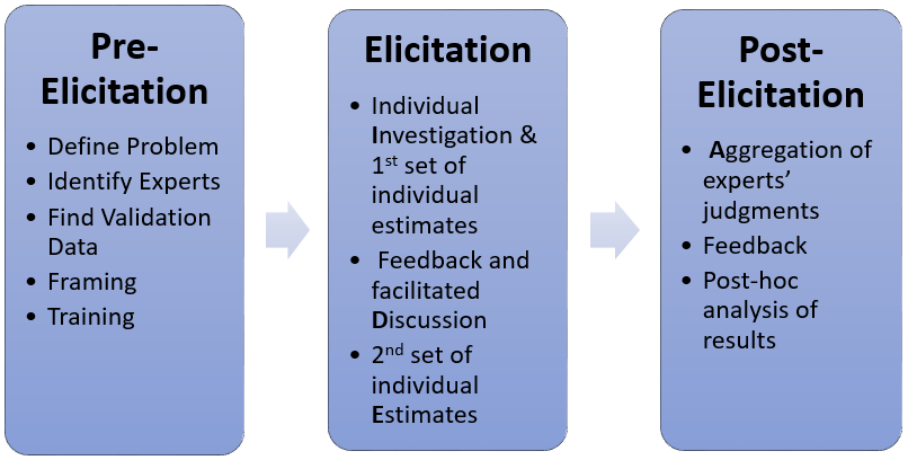


Figure 8.1: Structure of theIDEAprotocol. Replicated from Hanea [2018].

tured protocol for expert elicitation that considers such elements outside of Cooke's Classical method.

Within the elicitation process, a strict questioning process is utilised to support experts in providing their quantitative assessments. This questioning process consists of either three or four estimates. Three questions are utilised for eliciting the probability of discrete events and four questions are used when considering quantities measured on a continuous scale, Fig.8.2 . The four question format comprises of three questions to elicit different values for the variable, at different quantiles. The fourth question is then utilised to identify the probabilities associated with the upper and lower range of what has been provided. Similar to other elicitation approaches, this method is designed to counter biases that may appear as a result of unstructured or poorly structured questioning.

One notable element not highlighted previously is that the result of this questioning approach does not necessarily stipulate that all experts must provide identical quantiles. Experts are first asked for the values and then the probabilities associated with these. Each expert could feasibly provide a unique sized credible interval for the information they provide. These will typically be standardised, through linear interpolation, before either being played back to experts ahead of the second round of elicitation or as part of the aggregation process at the end. Standardisation here allows for direct comparison between experts. Whilst it may seem more direct to simply ask for the desired standardised quantiles initially, Speirs-Bridge et al. [2010], shows that allowing experts to specify their own quantiles and then standardising can reduce overconfidence.

The discussion and second round elicitation process is where the IDEA frame-

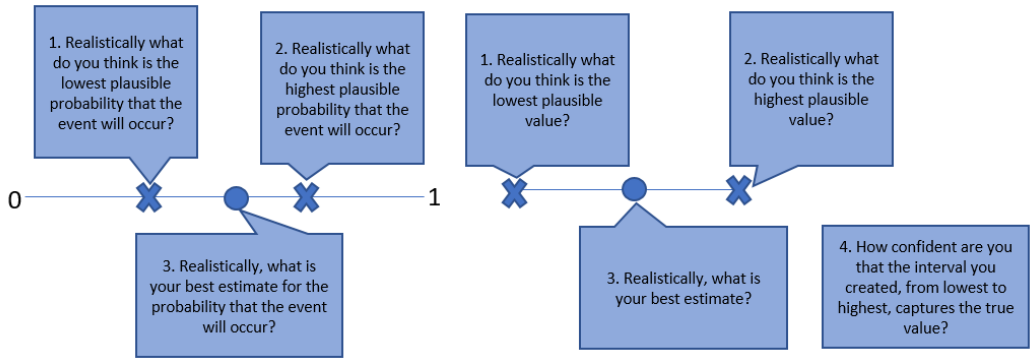


Figure 8.2: Questioning formats for eliciting discrete event probabilities (left) or quantities on a continuous scale (right) within the IDEA framework. Replicated from Hanea [2017].

work differs substantially from existing structured processes. The multi round approach of IDEA is akin to the basic premise of the DELPHI method (Linstone and Turoff [1975]), however the aim is not to arrive at consensus. There is also the opportunity for expert engagement between elicitation rounds, which is uncommon in DELPHI approaches. Conversely, having discussion post elicitation is aligned with behavioural approaches, such as the SHEFFIELD/SHELF method (O'Hagan [2006]), but the structure and aims of the discussion is very different.

The discussion is facilitated, by a trained facilitator, in a way which maintains anonymity with regards to ownership of judgements throughout. Experts reconcile differences they may hold in linguistic interpretation and context during the discussion, but they do not share any adjusted belief within this forum. The second round of elicitation occurs individually and the experts re-elicited perspectives remain anonymous. Maintaining anonymity during the discussion, allows for practitioners of the IDEA framework to benefit from the advantages of behavioural approaches, whilst mitigating against some of the more debilitating elements. Pressure to conform to dominant or authoritative voices within the group is reduced. Evidence from existing applications of the IDEA protocol demonstrate that the multiple rounds with discussion markedly improves the final group accuracy (Hanea [2017]).

The aggregation method within the IDEA framework typically takes the form of quantile aggregation as outlined in Chapter 3.3. It is recognised that whilst this method is fast to implement, it has typically performed poorly in relation to other mathematical aggregation methods when empirically tested (Hanea [2017]). More recently, Cooke's method has been implemented within the IDEA framework.

Analysis of Cooke's method relative to equal weighted linear opinion pools in this context found better performance of Cooke's method (Hanea [2018]), similar to results of, Eggstaff et al. [2014], Colson and Cooke [2017] and Cooke et al. [2020].

Concepts within the IDEA protocol have been used with significant success for the past decade, most notably within public health, conservation and ecology (Speirs-Bridge et al. [2010], Burgman et al. [2011b]). The most comprehensive use of the end to end protocol however was in a tournament set up by US Intelligence and Advanced Research Projects Activity (IARPA), in which a broad number of contexts were considered (Hanea [2016]). This is the same tournament that was ultimately won by the Good Judgement Project (Mellers [2014]), the method outlined in the book on Super-forecasting (Tetlock and Gardner [2016]). There are some conceptual similarities between the two approaches, notably the focus on making sure that experts thought multiple times about the problem and considered bounds on plausible outcomes. IDEA was unique in the whole tournament, however, in considering ranges whereas the other methods only looked for point probabilities, (Hanea [2018]). The Bayesian model outlined within this thesis is concerned with forecasts on continuous variables, and as such creates a logical link with the IDEA framework.

Whilst this captures many of the basic concepts behind the IDEA method, which provides a sufficient outline for the purpose of this thesis, there is a set of comprehensive protocol elements available. Hemming et al. [2018] is recommended reading for a description of these.

8.4 Integrating the Bayesian model into the IDEA protocol

Integrating the Bayesian model outlined within this thesis into the IDEA protocol gives a practical mechanism by which the method could be used by DMs. As already highlighted, it allows the Bayesian approach to be embedded within a process which seeks to consider both consensus and diversity of expert opinion.

The aggregation process of the Bayesian model is the easiest element to utilise within IDEA. Here, the Bayesian model outlined can simply be used in lieu of one of the existing IDEA aggregation methods. To enable this however the other steps within the Bayesian method outlined need to be considered carefully, and the points at which to address these within IDEA is a little more nuanced.

The first key decision that a DM or practitioner has to make is which modules within the Bayesian framework would they wish to employ. Whilst, *homogeneity group definition, distribution fitting, calibration and aggregation* represent the mod-

ules that can be utilised, not all will be appropriate within each context. To this extent the first direct change to the IDEA protocol would be to include a *component selection* element of pre-elicitation. This would involve discussion and agreement between the practitioners and the DM on key questions such as whether to calculate homogeneity groups algorithmically, or, to switch off this module and utilise existing information, such as expert affiliation to determine this. Similarly the validity of adjusting for expert overconfidence within this context could be decided. The Bayesian model overall is very flexible to these choices, but these key decisions then determine how and where the Bayesian approach will appear within the overall study deployment.

8.4.1 Homogeneity group definition

Discovering the homogeneity groups within our Bayesian framework has been described as a mathematical exercise, albeit with subjective judgment still required in a couple of specific structured places (i.e. the choice of clustering, the cut-off distance and the PCA review). It is a critical component to ensuring that dependence between experts does not cause uncertainty to be underestimated. The rationale behind creating a mathematical approach, is that practitioners of SEJ need support in defining these critical groups and existing mechanisms for doing this were too ambiguous. This does not mean that getting feedback and discussing the groupings proposed is not recommended. The mathematical approach is designed to provide structure and to support as an aid in the identification of dependence structures that might be obtuse.

The major prerequisite for homogeneity group calculation is the existence of the seed variables which can examined for clustering structures. The point at which these seed variables are elicited therefore becomes the defining factor in where the homogeneity groups can be finalised.

There are two points that seed variables can be gathered within the IDEA process flow. Seed variable responses can be captured in the Pre-Elicitation phase as part of the Final validation data gathering exercise or they can be captured as part of the first set of individual estimates.

If a longitudinal study is conducted, after a certain period of time, enough target variables may have been realised to generate meaningful seed data. In this case there is not the need to elicit seed variables specifically within subsequent implementations of the IDEA framework, the data can be used post its generation. This can be considered as capturing the seed variables in the *find validation data* step of the pre-elicitation phase. This is the mechanism that was used for the

application of Cooke's model within the IARPA tournament. Cooke's model was not implemented until sufficient participant judgements had been generated and realised to calculate the performance weighting scores of statistical accuracy and information.

Point-in-time studies are those in which experts are not asked for many forecasts over time, but are asked to comment on a specific set of target variables at a given moment. In this context the seed variables can typically not be gathered within the pre-elicitation phase as expert framing and training has not yet occurred. Here, seed variables must be captured within the Elicitation phase. This can either be done as a third elicitation round which is conducted prior to the initial set of individual estimates of the target variables, or it can be embedded within this first elicitation step. Given that expert fatigue is a common phenomena with SEJ, it is recommended that seed variables are elicited alongside target variables. Whilst this may cause a risk of distraction for experts within this individual round is does minimise the overall time taken for the process. Risks to the individual round can be minimised by careful ordering of the seed and target questioning. In this case the *find validation data* step in the Pre-Elicitation exercise is utilised to identify the seed questions against which the experts will be tested and gather the validated realisations.

Regardless of the point at which the elicitation of the seed variables occurs, the recommended point at which the homogeneity groups are calculated is the same. The homogeneity groups should be calculated post the initial estimates and before the feedback and facilitated discussion section. The rationale behind this is that there is perceived value to sharing the proposed groupings back with the experts.

Sharing the mathematically calculated homogeneity groups with experts as part of the facilitated discussion exercise provides two key benefits. Firstly it is an opportunity to increase expert confidence in the methodology utilised. Experts rarely like black box approaches and transparency is often key to getting the best engagement. Secondly, discussing the homogeneity groups with the experts provides opportunity for refinement on the purely mathematical outputs of this step. Experts may have visibility to potential causes of correlation which are not apparent in the seed variables utilised. Taking any such elicited information and adjusting the homogeneity groups accordingly may improve the accuracy of the final judgements.

It is important that sophisticated players do not use this as an opportunity to game the system. If such experts ensured that like minded individuals sit in different groups for example, this would increase the representation of that common belief in the final posterior. This can be mitigated however, through strict use of

facilitation in this stage and clear documentation on rationales for any changes. It is important that the final groups used are auditable, and so the mathematical process should be used as a base with logged and noted changes providing a complete audit trail.

Please note, the calculation and finalisation of homogeneity groups in this way results in the two-step Bayesian model being used. That is, homogeneity groups are calculated independently first and then calibration and aggregation follow separately afterwards. It would not be possible to utilise a single step model, in which homogeneity group definition, calibration and aggregation all occur within a single pass of the MCMC in this way. The single step approach does not allow for changes to the homogeneity groups from the algorithmic results. It is perceived that benefits gained from having the opportunity to discuss outweighs the slight loss of theoretical validity, as data is ostensibly used twice. The case study earlier demonstrated that the impact of this is likely to be benign.

Of course, the above assumes that the DM wishes to use the homogeneity groups calculated by the algorithmic method outlined. If the DM wanted to calculate homogeneity groups through another mechanism, or wanted to assign experts directly personally, this can be done earlier in the process. Providing visibility and the opportunity to discuss these during the first set of feedback is still recommended.

8.4.2 Fitting the distributions

Distribution fitting within the outlined Bayesian model is conceptually generic. Whilst a split normal has been utilised for the purposes of the data generated within this thesis, to ensure consistency in application, this may not be the appropriate choice in all contexts. In cases where there are fixed plausible bounds for example, an adjusted beta distribution might be a more natural choice. Similar to homogeneity groups, the choice of distributions utilised should be transparent to all those involved in the study, including the experts.

In principle experts may each have a different perception of the structure of the uncertainty on any given variable. When asked to encode their judgement explicitly within a distributional shape, each may have a different perspective on what this should be. The closest fitting distribution, g_e, for some experts for a single variable may be encoded as a beta, for others as a Student-T, for example. Aggregating mixed parametrised forms hierarchically for a single target variable within our Bayesian framework would create significant complexity, if it is even feasible. This approach is not recommended.

Having consistent distributions across experts, but variability across tar-

get/seed variables, however, is recommended. If some of the variables are bounded and others are not, then using adjusted betas for some variables and split normals for others may be appropriate. This can be encoded simply within the mathematical model, with careful consideration to the calibration parameters α_{le} and α_{ue}. Only a single set of these calibration parameters will be generated for each expert and applied across the multiple distributional forms. Ensuring that the application of these is robust and consistent is key. It is recommended that a common set of distributions with compatible calibration parameterisations is produced. This is not tackled within this thesis.

Ultimately, a single set of distribution forms to be utilised need to be produced and ratified by both the experts and the final DM. The distributional form elicited from experts has a direct impact on the posterior distributional form the hierarchical model produces and the form of the prior. All of these need to be aligned. This alignment necessitates the consideration of distributional forms at two points within the IDEA framework.

The first point of engagement is during the pre-elicitation problem definition section. Here, the study facilitators should be engaging with the problem owners to identify and agree, for each of the target variables, the functional form that will ultimately be modelled.

The second point of engagement, similar to the homogeneity groupings, should be post the first set of individual estimates. Within this section of the IDEA framework is where the standardisation of credible intervals typically occurs (Hemming et al. [2018]). Rather than utilise the linear interpolation assumption, the elicited data can be passed through the parameterisation process and the standard intervals that are shared back to the experts can be read off of the resulting distribution accordingly.

There is also choice in the visual method of sharing the standardised intervals back with experts. Whilst distributions are used to generate the standard intervals and ultimately will be used in the Bayesian aggregation, they can create complexity initially when discussing and comparing anonymous results with experts. Sharing Forest plots of standardised intervals, similar to Fig.7.13, provides a simple visual method for elicited quantile comparison. It is recommended that during this section of the facilitated discussion, Forest plots are shared first and then, as required, distributional plots are shared after, where this serves the underlying discussion the experts are having. Regardless of when and how the distributional shapes are shared, transparency with the experts on the process used to generate the standardised intervals is critical. Experts are gifted the opportunity to change their estimates

in the second round if the result of parameterisation and standardisation does not represent their true belief.

8.4.3 Calibration

Embedding the calibration component of the Bayesian model into the IDEA framework is the element that needs to be most sensitively addressed. Recalibration of experts' judgements can easily be misinterpreted and runs the risk of creating dis-engagement if managed poorly.

The first choice a DM must make is whether calibration is appropriate for the problem at hand. If not, this step within the Bayesian model should be excluded. This decision should form part of *component selection* within the pre-elicitation process outlined earlier.

When calibration is appropriate, similar to all of the other elements, it remains important to be transparent with experts on what calibration is, how it is used, and why it should be perceived positively rather than negatively by all involved.

The first time that calibration should be bought to the attention of experts is during the up front training exercise. Here, as part of standard training on bias reduction through structured elicitation, it should be highlighted that whilst vast improvements can be made, naturally it is not possible to remove all biases in this way. Statistics on overconfidence in well trained expert elicitations should be shared. It should also be highlighted that in order for the final picture the group provides to best represent the sum of all available information and their judgements, any further systemic confidence biases should be addressed.

The calibration model within the Bayesian framework should be clearly articulated. It should be stressed during this articulation that the method is entirely algorithmic, there are no subjective judgements from either the DM or the study facilitators with regards to the calibration of any individual.

The only subjective component of the calibration exercise is the choice of priors that are used. Stronger priors for α_{le} and α_{ue} here, set more weight on the original judgements from experts and require more evidence of mis-calibration within the seed variable testing data to diverge from this. The choice over the priors here should be guided by the number of seed variables that are available and the extent to which the DM believes these are representative of the target variables. Many, highly relevant, seed variables would suggest weaker priors. If only a few, loosely related seed variables are available, then putting stronger priors is more appropriate. More research is required to provide guidance on how to set these priors under different

seed variable cases.

Following the training exercise, it is recommended to share the calibration data back with experts at two key points. The first is as part of the discussion and review exercise and the final time is as part of the post-elicitation and aggregation feedback.

Following the collection of all of the first round assumptions and ahead of the discussion exercise, it is recommended to perform a round of aggregation in which all of the first round values are collected and passed through the full Bayesian framework, creating a final posterior at this point, ahead of discussion. It is useful to do this exercise as it helps experts to identify the current working assumption on the final output for the target variables.

The calibrated information should be shared back at this point, similarly at the end of the whole process, on the aggregated distribution but not on the individual experts. Individual expert calibration levels, even anonymised, should not be shared explicitly within the working forum. Plots like Fig.7.1. should be avoided. It is recommended that a very specific order of visual play back occurs. Firstly, standardised judgements from individual and anonymised experts should be played back, both in forest plot and distributional form as required. Following this an overview of the aggregate un-calibrated distribution should be shared. This should represent the view of the experts with no adjustments made. Finally a recalibrated aggregate distribution should be shared. This demonstrates to experts the extent to which the calibration exercise has impacted the final proposed uncertainty profile.

Explicitly not sharing expert level calibration impacts within the working forum is not to avoid transparency and scrutiny over this component. The rationale for not sharing this is that the aim of calibration is to identify how unexpressed uncertainty may impact the final distribution. Individual experts levels of calibration is not the focus. Sharing expert level calibration outputs with the group during discussion can result in dis-engagement from individuals if they feel that their very confident views are not being considered as such. Whilst experts are not able to identify each other in anonymised feedback they are often able to distinguish themselves, as they are aware of what they provided. Visibility into the level of recalibration that would occur on their judgements, either significant or not, may then lead to undesired behaviour, including, if the facilitator is not careful the calibration being subjectively adjusted.

Anonymised expert level recalibration effects can and should be shared as part of the final documentation and write up of any study. This ensures complete auditability in the future. Sharing such anonymised results with experts post study

also builds credibility without risk to undesired impacts to the final understanding of the uncertainty around the target variables within. For longitudinal studies in particular this needs to be carefully managed.

My personal experience with expert recalibration has been that, whilst experts are understandably conscious of their own judgement and have personal motivations, their primary desire whilst partaking in SEJ is to support in understanding risk. Many experts feel nervous about recalibration, but careful and explicit explanation of why it is important, ensuring them it is not subjective and making it clear that at no point will their performance be made public, can mitigate a lot of these misgivings. Transparency on process but protection of the individual is key. With these few simple steps, many experts will gladly have their judgements recalibrated if it leads to better outcomes.

8.4.4 Impact to process flow

To simplify implementation, proposed adjustments to the IDEA protocol can be implemented in two additional macro steps. The first additional step would be a *model configuration* step within pre-elicitation, post *finding validation data*. This step would include the *component selection* and *homogeneity group definition* elements outlined earlier as well as the decision on the distribution to be fitted for each variable (both seed and target). Finally, this component will included a step for *prior definition*. This is a critical component that can only be decided for certain hyperparameters once the target and seed varaible questions are well understood. It is useful to cluster all of these steps into a single process element within pre-elicitation as the study facilitators will need to get direct input into these elements from the DM. In practice the *model configuration* process will be two steps. There will need to be some analysis conducted by the facilitators to identify options and some recommendations collated. This should be followed by approval from the DM on the choices made. DM time is often very limited and so ensuring these components can all be covered with a single engagement is key.

The second additional step would be a *preliminary analysis and aggregation* step. This should encapsulate the *distribution fitting* component and calculation of *aggregate* and *calibrated-aggregate* posterior distributions.

The combination of these two steps should give all of the material necessary for study facilitators to feed the additional Bayesian methodological components into the facilitated discussion. Fig.8.3. highlights where these two new steps occur in the process and outlines how other subcomponents are specifically addressed within existing process steps.

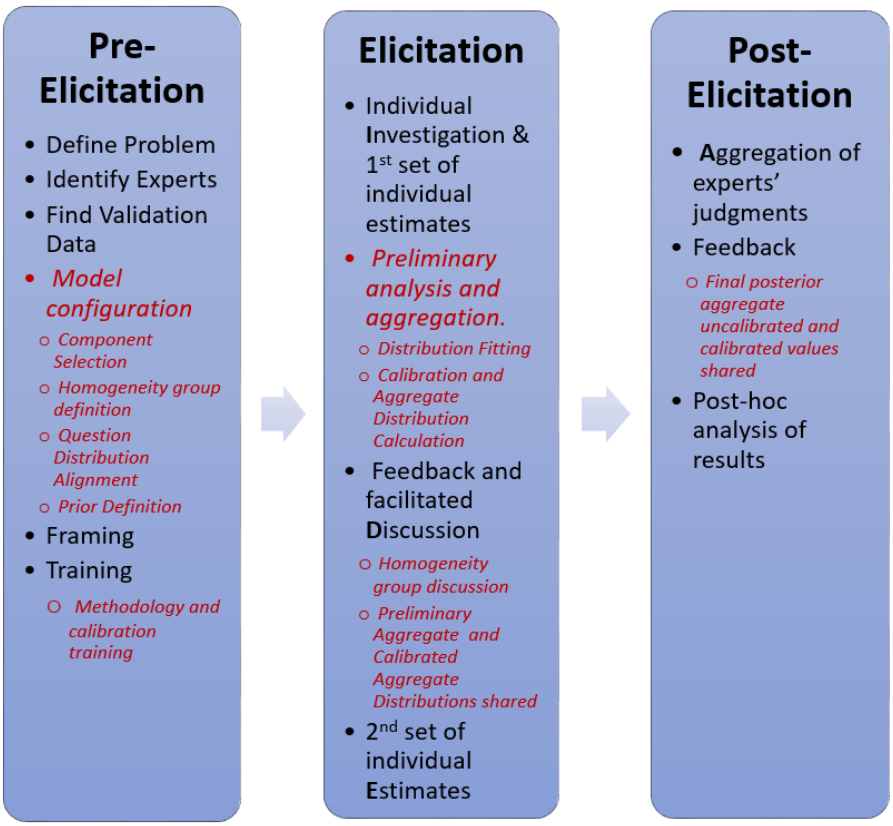


Figure 8.3: Proposed adjustments to the IDEA protocol to embed a Bayesian model with both calibration and homogeneity group definition.

The topics outlined in the new proposed steps are not new to IDEA. Model selection, and preliminary analysis always occur within this framework however are not explicitly called out. The reliance of the Bayesian model on some of these more complicated areas, and the more nuanced ordering in which they must be performed, generates a need for practitioners to have greater clarity. Elevating these topics to core process steps makes this explicit.

Overall, there is a natural fit for Bayesian approaches within the IDEA framework. The focus of the discussion process on removing linguistic uncertainty and promoting clarity on context, provides opportunity to improve judgement quality whilst minimising the risk of correlating experts and under-calling uncertainty by driving to complete consensus through group discussion. If a Bayesian IDEA practitioner wished to be even more cautious with regards to the risk of increasing correlation, experts could be split into their homogeneity groups and separate facilitated discussions held for each group. This would still provide opportunity for

context and linguistic uncertainty to be removed, but would ensure that inter-group correlation would not be increased. This would only be possible in large elicitation exercises, where each homogeneity group is sufficiently populated to have quality discussion. It would also rely on very careful facilitation as the discussion facilitator would have to ensure consistent responses with regards to lack of clarity on the question themselves whilst not providing information that would lead the discussion and inadvertently drive further correlation.

8.5 Further implications of Bayesian approaches to practical procedures

There are a couple of other aspects of Bayesian thinking that it is important to consider when embedding a Bayesian framework within an an existing elicitation protocol. These are not unique to any single elicitation approach, and so we will consider the impact on both the IDEA framework and the EFSA protocol outlined in Chapter 2.4. The clear delineation of roles within the EFSA framework provide opportunity to examine different mechanisms by which some of these elements can be judged.

8.5.1 Prior ownership and definition

The key difficulty in the Bayesian approach is often the development of tractable likelihood models, $p_{DM}(p_E|x_X)$, as these are the most mathematically complex elements. However, when considering any set of elicitation procedures the decision on who should own the prior for the model π_{DM_X} is an equally important consideration. In a fully subjective context, i.e. when there is a single decision maker who is ultimately accountable for the output of the SEJ, e.g. a commercial leader who is utilising SEJ to invest their own money, it is very reasonable for this individual to own the prior personally. Many protocols (EFSA [2014], Cooke [1991]) are typically looking for the view of a rational scientist. In these cases, a Supra-Bayesian approach, as utilised in Albert et al. [2012] and described in the model outlined in this thesis can be used. This hypothetical Supra-Bayesian would still need to have priors assigned and so the problem remains.

If the EFSA definition of roles within an SEJ study are considered, there are several places for potential prior ownership:

- **The Working Group** – *Ultimately accountable for the output of the* SEJ *as it feeds into the risk assessment model, one option would be to have the Working*

group define and own the priors.

- **The Steering Group** – *Closer to the refined parameters being elicited, the experts and the final method utilised, the Steering Group would also be a logical owner for the priors. If any quantitative assessment of the variable of interest was made during the initialisation of the pre-elicitation phase then this could be considered as a prior.*

- **The analysts performing the modelling** – *If naive priors*[2] *were used and therefore limited knowledge encoded within the prior specification, the analysts could feasibly act as a proxy for the rational scientist.*

- **The experts themselves** – *Another source of priors would be the experts themselves or a subset of other experts. Utilising the experts for the prior however, would draw us closer to the group decision problem, rather than the rational scientist expert judgement problem typified in the* EFSA *model. This is not inherently a problem however does bring about a number of other constraints to be considered and blurs the boundary of the role of the expert vs. the decision maker. If this approach were to be taken, a very different set of processes would need to be considered. For an expert problem all knowledge from the expert's should be codified in the likelihood function.*

The most compelling of these options would appear to be the Steering group, as they represent a body close enough to the problem whilst still in a position of accountability. In the model described in this thesis there are also multiple priors being assessed. There are priors over the variables of interest but also over the experts (and their potential correlations) themselves. It would be a considerable risk for the analyst to be accountable for these priors due to the potential impact on the output and the legal ramifications discussed before.

The choice of prior ownership is also impacted by the affects of power and authority in decision making, (Sagi [2015], Fiske and Berdahl [2007]). Diverse groups work well because of the variety of social perspectives they bring. In cases where ownership of priors is salient but the decision context means that there is not a single DM from whom they should be elicited, then diverse ownership would be preferred. Without this, the analysis may become a vehicle for one group to impose their values on another. This is another argument for having the priors owned by a group such as the steering group, as long as this has a diverse composition.

[2]Naïve priors are very flat distributions which seek to represent complete lack of knowledge or something close to it.

Once the owner of the priors have been finalised, decisions need to be made on to what extent informative or uninformative priors can be utilised. Priors can be used to encode knowledge into the model when informative or represent lack of existing knowledge when uninformative. Uninformative priors would focus the final output much more directly on the expert judgement, however, would clearly reduce the amount of data encoded into the problem.

The context itself here is important, in a fully subjective model with a specific decision maker it would be unwise to utilise uninformative priors as ultimately you are trying to update someone's belief in the light of expert opinion and the decision maker's belief is naturally a critical starting point to this. Explicit elicitation of decision makers priors in these contexts can ensure that results from a study are ultimately utilised. Eliciting these initial beliefs ahead of study deployment is important for any SEJ practitioner in these contexts and is often missed. DMs are often reluctant to share their perspectives *a priori* as they do not want to bias the study. However, understanding this critical information may be important to study design and ultimately deciding whether an SEJ study is appropriate at all.

The extent to which DM priors are informative/uninformative can be translated into their willingness to change their position given the experts' judgements. A DM who insists on priors that are very tight will in a Bayesian model, consequently not see significant divergence from this given the experts' judgements, even if these judgements differ substantially from those of the DM. In these cases it may be appropriate to insist that it is not worth the time and investment to conduct a study and the DM should proceed with the decision they would ultimately take regardless. Throughout my career, I have made a point of eliciting these prior judgements on many occasions. A significant number of projects have not been commissioned as a result. This time saving can then be translated into other value driving activities.

It is also important to note the impact social power has when a DM considers tight priors within a model, (Fiske and Berdahl [2007]). At times, whilst this might be a DM's initial perspective during a pre-elicitation exercise, it may not be appropriate and the impact to the outcome of the exercise may not be well understood. It is recommended that study facilitators engage with the DM about the effects of social power, the value of diverse opinions and the implication that prior choices may have to the study.

For the rational scientist viewpoint it would appear sensible to aim for uninformative priors over the variables of interest. However, we would argue that priors over the experts should not be completely uninformative. If for example we consider calibration; starting from the belief that experts are well calibrated and

only recalibrating with significant evidence (where significance here is determined by the application of Bayes rule with a calibration data set) rather than starting as agnostic to calibration issues, would appear an appropriate decision for the Steering Group.

8.5.2 Evidence dossiers

Another critical component of the EFSA process, and other protocols, that must be analysed from a Bayesian perspective is documentation. One constituent of the EFSA guidance is often a shared evidence dossier. This dossier captures all of the known data regarding the parameters of interest, and the risk assessment model, ahead of the elicitation exercise and is shared with all of the experts. This is important for transparency; and from an auditing perspective, it would appear to be unethical to with-hold evidence from an expert before they are due to make judgements that may impact critical decisions. The legal ramifications of a decision being recommended when data was withheld may be substantial. However, there is both a technical and philosophical issue with this approach.

The technical issue is that in creating this evidence dossier the Steering group may inadvertently increase the correlation between experts as, by definition, they are given a shared body of knowledge from which to base their judgements.

From a philosophical perspective there is also an issue with this approach as it makes assertions about the evidence base that may not be complete. Indeed as we are engaging in a SEJ study it is incomplete by definition. Sharing this partial data with the experts may further bias the results, for example increasing the risk of the availability bias being demonstrated, or increasing the likelihood for experts to become overconfident. There are a couple of ways that this issue could be handled in the Bayesian model:

- **Evidence Dossier shared with experts pre-elicitation** as per the standard EFSA guidance – *Increases the risk of bias and correlation but ensures that no data is withheld.*
- **Evidence Dossier shared with experts during elicitation** – *It would in theory be feasible to elicit the experts knowledge before they see the Evidence Dossier and then perform a second elicitation after this has been shared. This would allow the analysts to directly ascertain the impact of the evidence dossier and consider this within the final recommendation, however, would put a significant burden on the elicitation process.*

- **Utilise the Evidence Dossier in prior definition** – *As empirical evidence, it would be potentially feasible for the evidence dossier to be used by the Steering Group, rather than the experts, in the definition of the priors. This would ensure that priors encoded knowledge, but only empirically generated knowledge and would also help to ensure that the data is utilised in the process. Here the experts never see the evidence dossier and therefore there is no increase in cross expert correlation, however, the issue of data being withheld is reduced as any final recommendation will be net of any existing evidence.*

The decision of which approach to use in any analysis may again be context dependant.

A further decision to be made is; How to pass back qualitative knowledge along with the consensus distribution? The final output of any structured elicitation exercise should not just be the consensus distribution itself but also the qualitative knowledge experts utilised to inform their decision making. This qualitative knowledge can be critical in getting decisions ultimately implemented and to enrich/explain the outputs of the analysis. Any documentation requirements should consider these elements in addition to the distributions.

Chapter 9

BEAM software

In addition to processes and procedures it is also important to consider the software utilised to support any elicitation or to perform any analysis. Software plays a critical role in minimising effort for facilitators, providing visual ways for analysis to be digested, and can be a key tool in supporting process adherence. Structured decision tools typically have a modular workflow. Rigidity in this workflow can support facilitators in ensuring that process is followed. Perhaps surprisingly, for facilitated discussion and elicitation purposes tool flexibility can be a risk rather than a benefit.

To support the implementation of the new Bayesian model into the IDEA protocol, an R-Shiny app has been built utilising the structural components outlined earlier. Not all elements of the IDEA protocol will have a tool counterpart, for example *expert identification* is an activity for which a tool is likely to offer limited benefit. There are other elements which currently do not have a tool component but could in the future, e.g. *training*. In an idealised world, the BEAM tool would operate as a comprehensive audit trail that underpins the deployment of the IDEA protocol for a given study. All training materials could be integrated into the tool as well as other core materials for study documentation. To enable this it may be appropriate to refactor the tool into something other than R-Shiny. This is left for further research.

The work-flow elements implemented into the tool, and mirrored in the tab selection hierarchy of the model (Fig.9.1) are as follows:

- **Model Configuration** - A section of the tool to capture key study meta-data, such as the study name, number of experts, etc. Also a place to define which components of the the Bayesian model will be utilised.
- **Primary Elicitation** - Post the first elicitation session, a section to load and

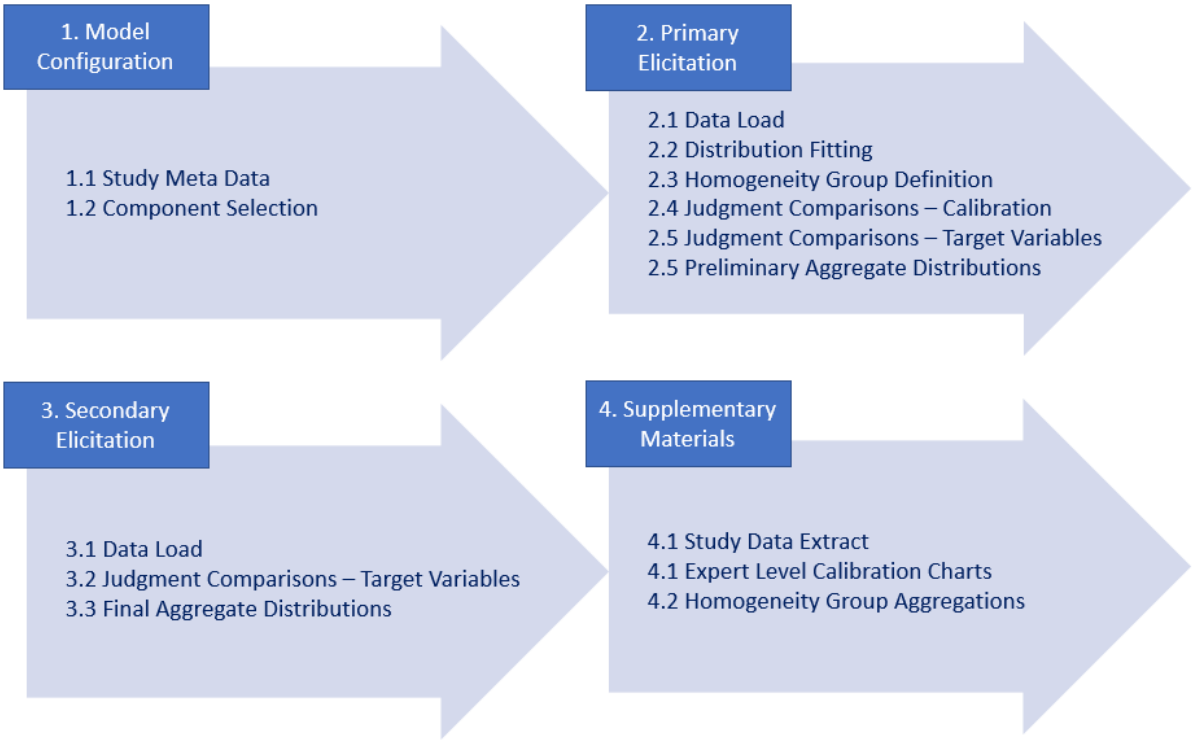


Figure 9.1: Workflow and tab hierarchy of the (B)ayeisan (E)xpert (A)ggregation (M)odel.

analyse these preliminary results. Visuals to support in the discussion section are generated within these tabs.

- **Secondary Elicitation** - A summary of the final individual and group judgements arising from the study.
- **Supplementary Materials** - Other visuals that can be used by facilitators to either analyse the results that are being presented or to be included in post study documentation. These visuals are typically not shared during the live discussions with experts.

The value in SEJ projects utilising IDEA is from the conversations that experts have and how this translates into adjusted judgements. The primary focus of the BEAM software is to support this discussion process. The key user of BEAM is intended to be the study facilitator. A simple user interface has been constructed to ensure that navigating and utilising the modelling capabilities with BEAM enables rather than detracts from this discussion. Graphical user interface design is a well studied area (Blair-Early and Zender [2008], Bhaskar et al. [2011], Nielsen [1995]). Key considerations such as clarity, comprehensibility, consistency and control must

be ensured throughout any tools process flow (Bhaskar et al. [2011]). To this extent four core principles, based on the theories in the outlined texts and tailored to the SEJ context, have been generated and applied to the graphical user interface within BEAM:

- *Tab proliferation (with clear signposting on the role that each tab plays) is preferred over increased complexity within single graphical panes.*
- *Choices presented to users should be structured to have the minimum options feasible. Binary choices are preferred.*
- *Mathematical/technical complexity should be hidden during normal utilisation. Although, it must be possible to get visibility to this information if desired.*
- *Automation should be applied wherever possible to minimise the number of touchpoints required. Users should only need to perform an action once and results proliferate through to the remainder of the tool.*

Fig.9.2 demonstrates these principles applied to the process of capturing study meta data and in deciding the modelling components to utilise.

Unlike other structured expert judgement software, BEAM does not have functionality to support the actual elicitation of the numbers from individual experts. Elicited judgements are loaded into the tool in bulk. Outside of behavioural methods, individual expert elicitations often occur on templated paper which are subsequently transcribed into an aggregate data table, typically stored in Microsoft Excel, manually. BEAM ingests data in either Excel or CSV form. There is a standard template that needs to be completed. The dimensions in this template include:

- **Target/Calibration Variable**[1] - a name for the item under consideration.
- **Description** - a more detailed description of the variable considered.
- **Expert** - the unique identifier for the expert. This should be post anonymisation. For data privacy reasons unblinded expert names should not be loaded into BEAM.
- **Round** - a flag for whether the data is from the first or second round of elicitation.
- **Lower** - the minimum value provided by the expert for this item in the specified round.
- **Mid** - the median value provided by the expert for this item in the specified round.

[1]Seed variables within BEAM are known as *calibration variables* to make their link to the recalibration process more explicit.

- **Upper** - the maximum value provided by the expert for this item in the specified round.
- **Lower.Percentile** - the elicited percentile associated with the minimum value provided.
- **Upper.Percentile** - the elicited percentile associated with the maximum value provided.
- **Realisation** - the true realised value of the variable considered. *Only applicable for calibration variables.*
- **Rationale** - Any reasoning the expert has given to justify the elicited figures.

Lower and upper percentiles are entered separately to allow for the case that experts provide an unbalanced interval around the median. The unique key for each row in the dataset is the triple (expert, variable, round). Fig.9.3 outlines how data templates are loaded into the system via standard directory file selection methods and are immediately visualised so that facilitators can ensure no errors occurred in transcription or load.

Once data has been loaded into the system, percentiles are standardised. This is done within the distribution fitting tab outlined in Fig.9.4. Each expert can be individually selected to highlight the impact standardisation will have on their elicited judgements. The original confidence interval and the standardised quantiles (and associated probabilities) are shown both graphically and in tabular form. Providing both formats allows for easy digestion of the changes made and supports different cognitive processing methods. Throughout BEAM, data is shown both in graphs and tables. This aligns with different thinking styles and makes both specific and relative values transparent.

The standard quantiles are produced using two methods, the first is a linear interpolation method (Hemming et al. [2018], Adams-Hosking et al. [2016]). Here:

$$\begin{aligned} \textit{Lower Standardised Quantile} &= M_{Xe} - ((M_{Xe} - L_{Xe}) * (\frac{P_M - sP_L)}{P_M - eP_L})) \\ \textit{Upper Standardised Quantile} &= M_{Xe} + ((U_{Xe} - M_{Xe}) * (\frac{sP_U - P_M}{eP_U - P_M})) \end{aligned} \tag{9.1}$$

Where M_{Xe} = mid value elicited, L_{Xe} = lower value elicited, U_{Xe} = upper value elicited, P_M represents the percentile of the mid value, typically 0.5, sP_L/sP_U represent the standardised percentiles desired and eP_L/eP_U represent the percentiles associated with the originally elicited values. The linear interpolation method is often utilised when considering a quantile aggregation method.

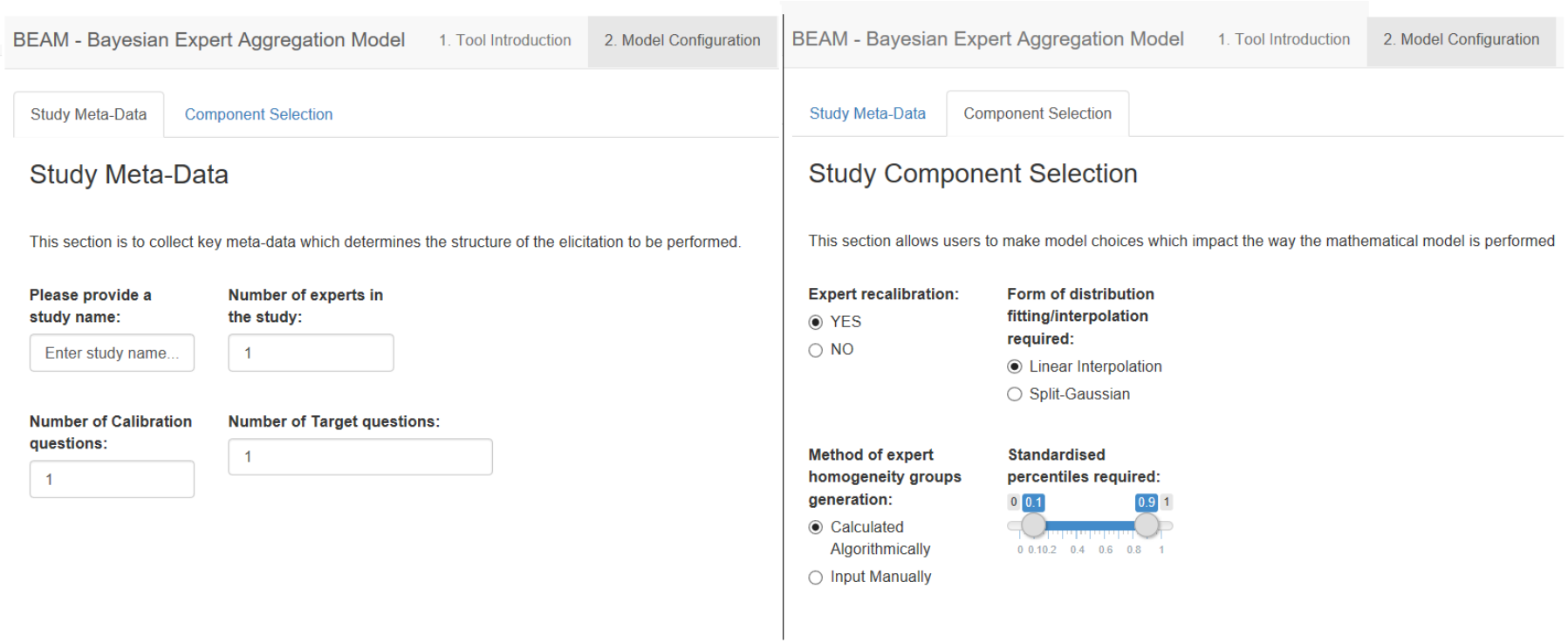


Figure 9.2: Screenshot of the two BEAM model configuration tabs. To ensure ease of use, only a few choices are presented to users. When decisions are necessary, options given aim to be binary.

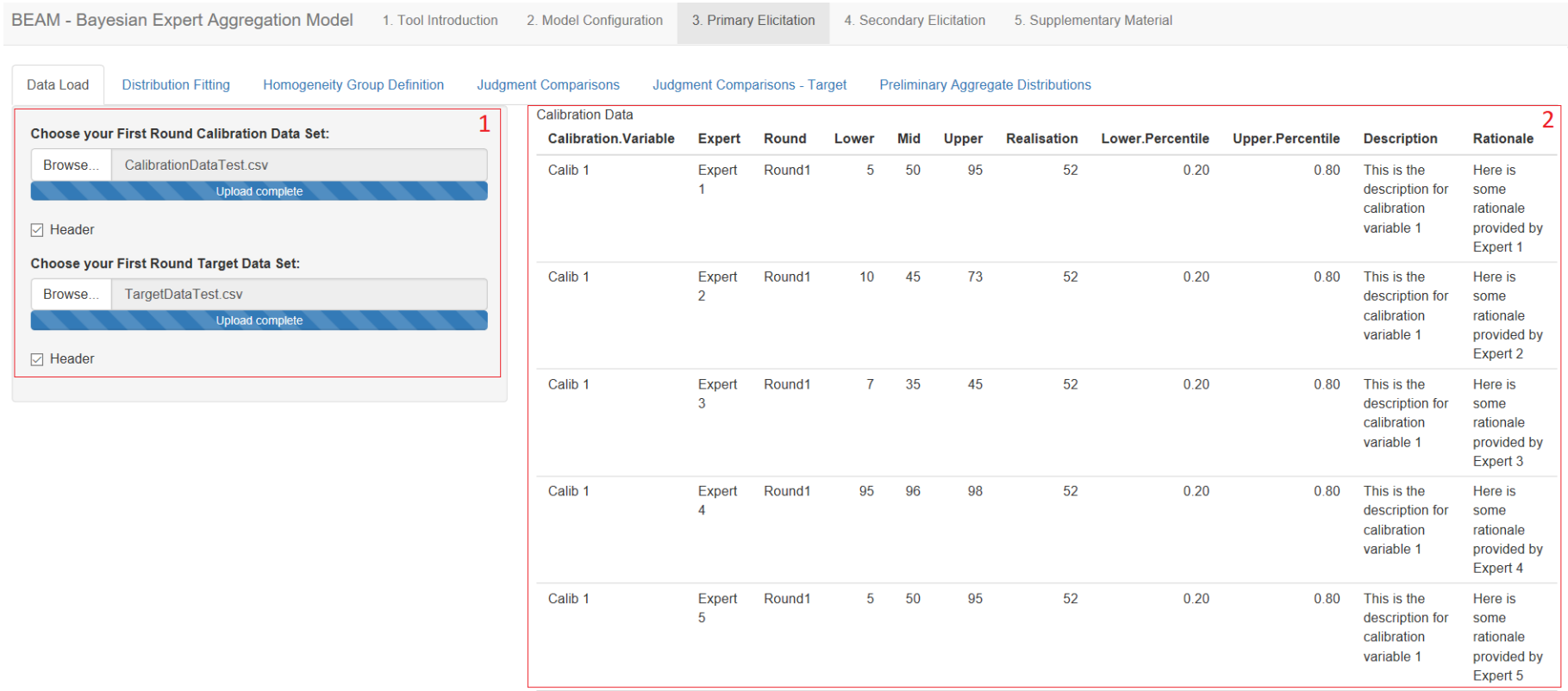


Figure 9.3: Screenshot of the BEAM data load tab. Data is loaded into the system via standard folder browsing and file selection (1). The loaded data is immediately visualised to ensure that the load processing did not generate errors and that template transcription was completed appropriately (2).

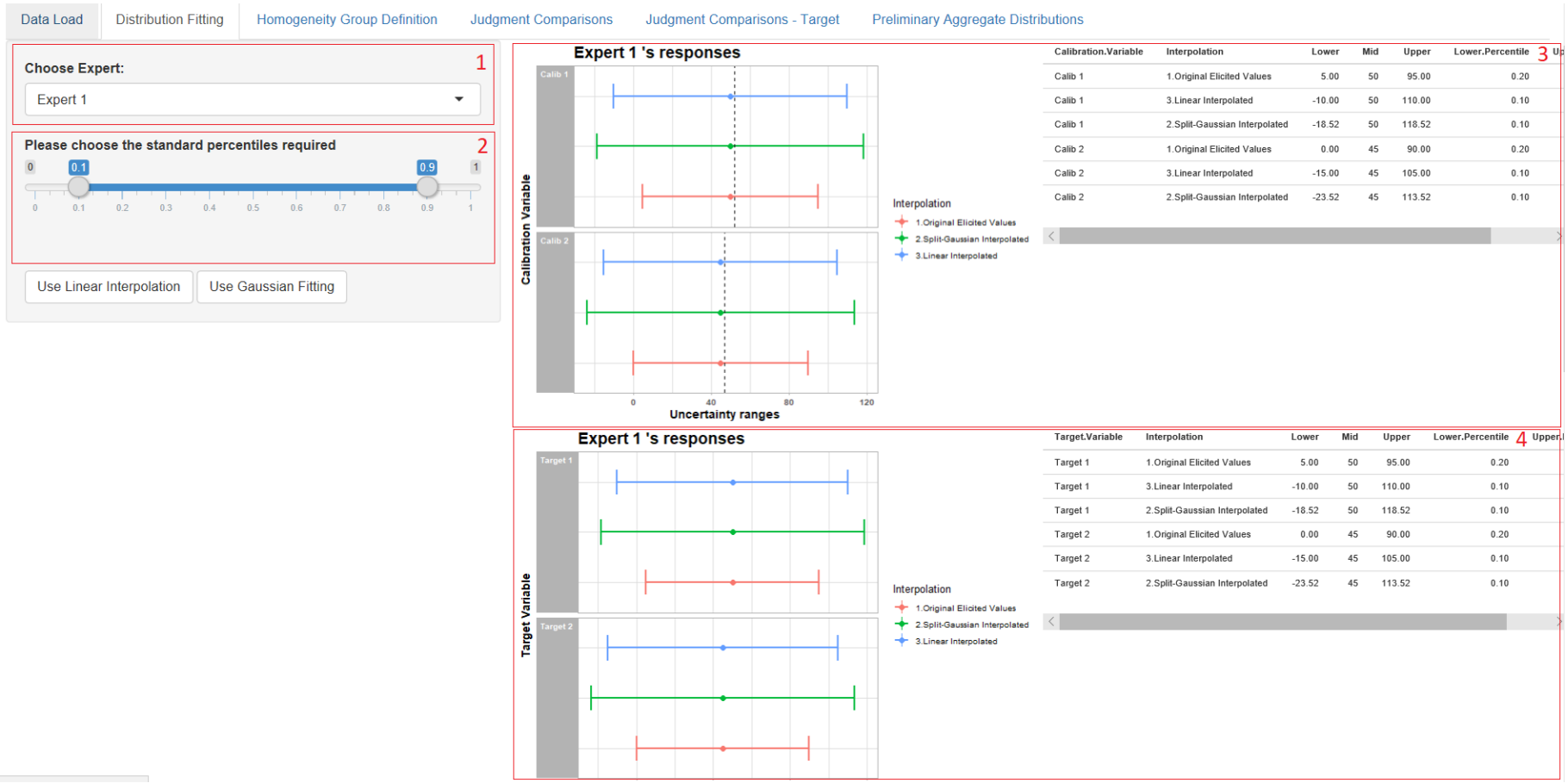


Figure 9.4: Screenshot of the BEAM distribution fitting tab. Individual experts can be selected (1). The impact of standardisation through linear interpolation and split normal fitting can be viewed graphically and in numeric form. The impact to both calibration (3) and target (4) variables is shown. A dynamic slider (2) allows for the desired standards to be changed and the impact immediately rendered.

The second method of distribution fitting embedded in BEAM is by fitting the split normal distribution, as per (5.27). This method does not require interpolation, the distribution can be fit to the elicited values and then the standardised quantiles simply read-off. Despite only one of these distribution fitting methods being utilised in the modelling both are visualised on this tab so the relative impact of each method can be considered. The facilitator decides on the method type, through discussion with the DM. This can be set either in the model configuration screen or selected through the action buttons on this tab. It is recommended that the linear interpolation method is utilised if the final desired aggregation takes place with quantile aggregation and the split normal method is utilised when the Bayesian model is to be considered. If more parameterisations of the Bayesian approach are created then the options should be extended accordingly.

The distribution fitting tab also has the opportunity to investigate different choices of standardised percentiles. A slider is provided (populated originally with the values defined on the model configuration page) which allows a study facilitator to change the desired standardised percentiles and dynamically render the resulting impact. The selection here does not overwrite the original standardised values selected but allows the impact of this to be understood. If upon reflection a different standardised value is desired this can be implemented by going back to the model configuration tab and updating.

Homogeneity group allocations are made easy in BEAM by allowing either automatic algorithmic calculation or manual entry. Fig.9.5 outlines the realisation of this in the tool. Groups are generated automatically using the unsupervised learning based clustering methods outlined earlier. Users are shown a plot of the first two dimensions of the PCA to help understand the expert locations within the calibration space. Scree plots of the PCA are also provided to demonstrate the percentage of explained variance within the two shown dimensions. As before, the clustering recommendation is performed over the whole calibration space and the PCA is just included to help users understand the rationale for recommendation. A dynamic slider is available to allow the user to set limits on the number of homogeneity groups. This does not define the exact number of clusters but sets a maximum.

When facilitators feel uncomfortable with the algorithmically calculated groups, or there exists logical clusterings based on information not captured within the calibration data, manual groups can be added. These are loaded into the system via a very simple Excel data template. In such cases it is recommended that the algorithmically generated groups and PCA are still considered. Facilitators may identify other underlying structures in the data which were not considered previously.

The two judgement comparison tabs, Fig.9.6, within the Preliminary Elicitation section of BEAM allow individual variables to be assessed across experts. These tabs are designed to provide visibility to the diversity of judgements provided in the preliminary elicitation and to highlight the variety of rationales provided for the numbers given. Everything shown on this tab reflects the standardised values to ensure it is comparable. Specific numbers are shown for the anonymised experts for completeness but these are not the focus.

These tabs should be shared with experts as part of the group discussion. In addition to the numbers elicited, BEAM renders the unstructured text provided within the rationale field of the input dataset. Conversation around the reasons provided for the judgements given allow opportunity for different linguistic interpretations to be recognised and uncertainty here eliminated. It also allows for discussion around the different sources of information considered, and the relative robustness of each. The facilitator has an important job to do here in ensuring that discussion balances the important aim of reducing uncertainty but simultaneously not increasing correlation too significantly.

Whilst not of primary concern, it can be helpful to highlight and discuss the calibration variables in addition to target variables. This provides opportunity to test the discussion process and can make experts more efficient in the target variable conversations. It also allows for the opportunity to re-elicit calibration variables and identify if the discussion has impacted correlation and therefore the homogeneity groups that should be considered. Discussion and re-elicitation of calibration variables is uncommon as it significantly increases effort required. SEJ studies are often time bound and there is a risk that this exercise increases expert fatigue which could reduce quality on target variables. BEAM provides the functionality to support this, if necessary, by including a judgement comparison' tab for calibration variables which is identical to that for target variables and by including the option for a calibration dataset load in the secondary elicitation workflow.

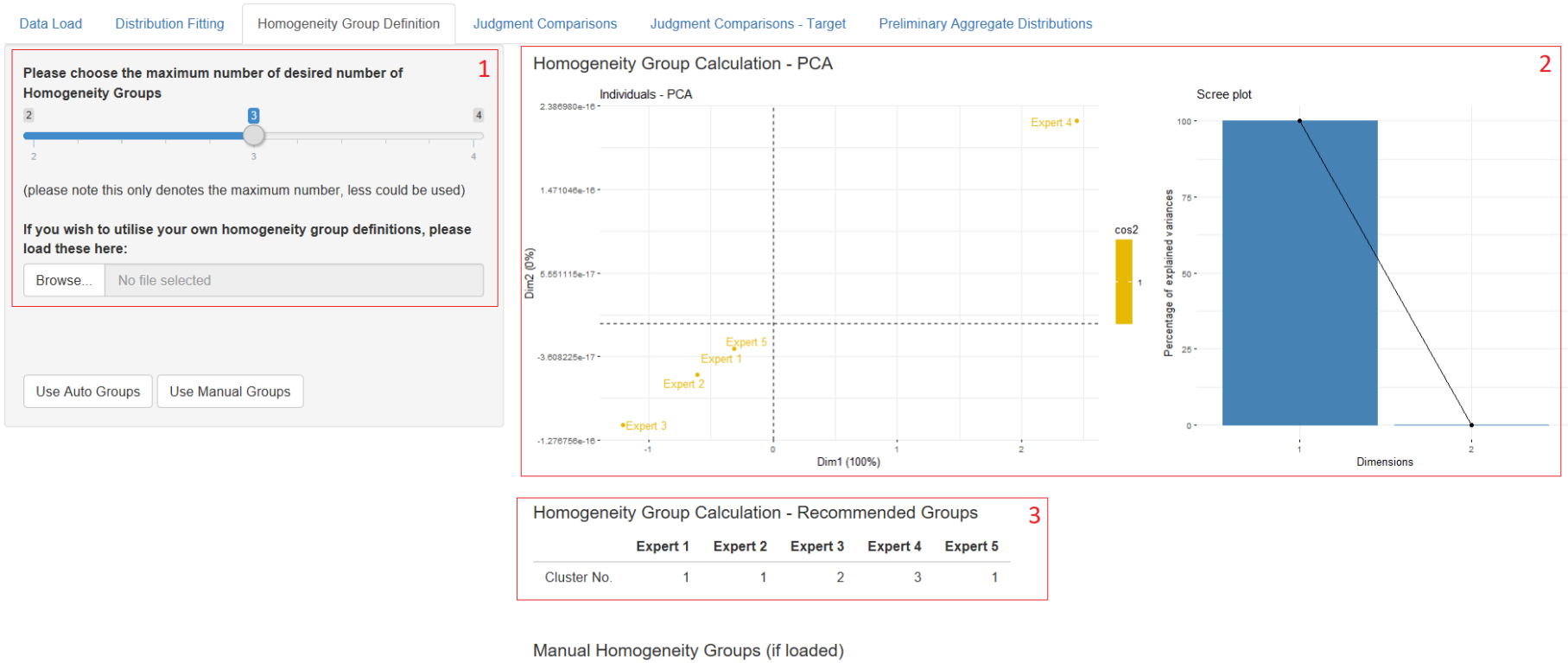


Figure 9.5: Screenshot of the BEAM homogeneity group calculation tab. A PCA plot of the calibration space (2) demonstrates the structure within the first two dimensions. A Scree plot highlights the percentage of explained variance shown. A recommended grouping is provided (3) based on an underlying hierarchical clustering algorithm. As in the pictured example the link between the plot and the recommended groupings helps users identify the rationale for the recommendation. Users can define a maximum number of desired groupings (1). When manual groupings are required these can be loaded via a standard excel template.

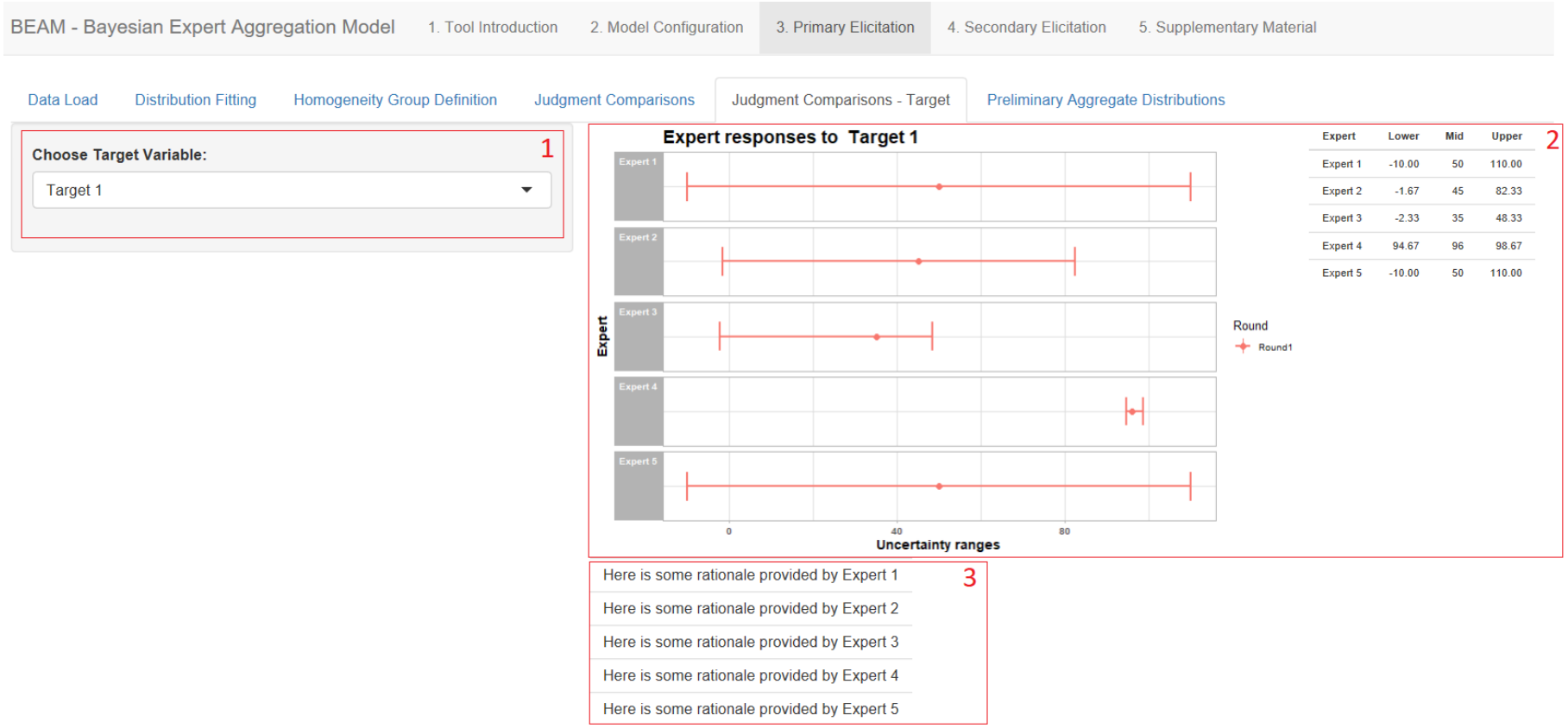


Figure 9.6: Screenshot of the BEAM judgement comparison target tab. Each variable is selected individually (1). Standardised judgements across experts are shown in the commonly used forest plot format with specific numbers shown in a table (2). Unstructured text for the rationale provided by each expert is shown (3).

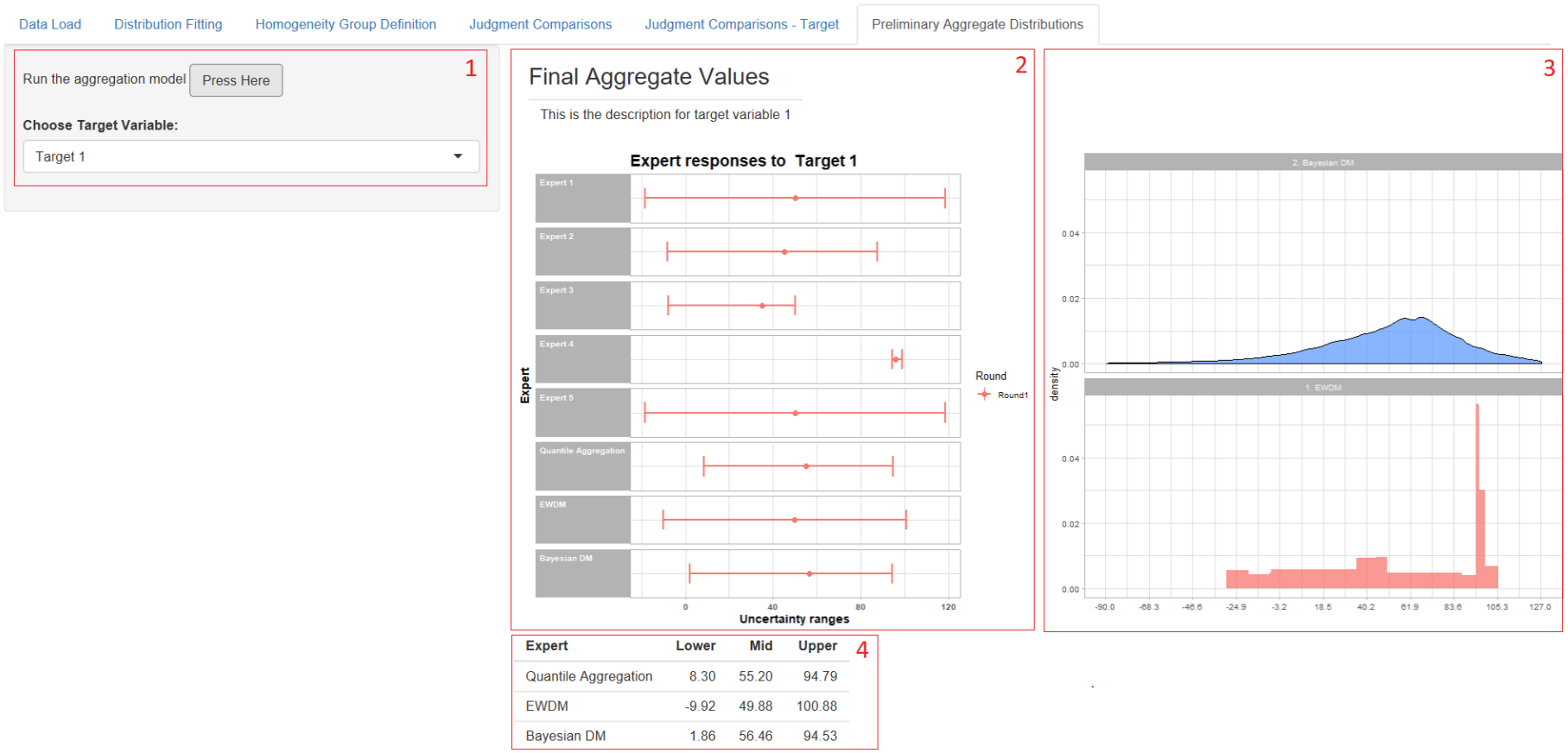


Figure 9.7: Screenshot of the BEAM aggregate distributions tab. The aggregation models are run through an action button (1). Unlike other tabs, as the aggregation process can take time, the algorithms are not run unless triggered. Aggregate standardised quantiles are shown relative to the elicited expert judgements (2). The distributional form of the Bayesian model and the equal weighted linear opinion pool are highlighted (3). Specific standardised quantile values for each model are displayed (4).

Aggregate distributions are generated in BEAM utilising three methods; Bayesian modelling, quantile aggregation and an equal weighted linear opinion pool. As standard, all three methods are displayed simultaneously to allow identification of the impact that different approaches have to the final aggregate numbers. To ensure consistency with other tabs within BEAM and to allow easy mental linkage between stages of the process, initially, identical plots are shared. After selecting a variable to review, the aggregate tab, Fig.9.7, shows a forest plot of the standardised quantiles for both experts and the three aggregation methods. This allows reviewers to identify the link between initially elicited judgements and the final aggregates.

Distributions are displayed simultaneously for the two methods which rely on their construction. The distributional structure, in addition to the standardised quantiles, is visualised for the linear opinion pool and the Bayesian model. Understanding the distribution of uncertainty implied by each aggregation method highlights the different elements of the expert judgements that are emphasised.

Distributional shapes are not shown for the quantiles aggregation method. Implicitly there is a set of underlying distributional shapes generated when utilising such a method but these are rarely reviewed. Quantile aggregation is often used when there is a desire not to deal with the complexity that distribution fitting brings, it can be generated directly from the standardised quantiles. As no distribution assumptions for the expert judgements have been confirmed as part of the preceding process, there are in principle an infinite number of distributional shapes that would generate the resulting aggregate values. A view of the quantile aggregation method distribution could also be included, if desired, by agreeing on a standard distributional form to map the experts' judgements with.

The second round of elicitation is visualised identically in BEAM to the first round. Elements of the first round that are already confirmed such as distributional fitting methods and homogeneity groups are removed and the tool automatically calculates these according to the decisions that were made in earlier tabs. To this extent, there are only three tabs in the second round section; data load, judgement comparison on the target variables and the final aggregate values. There are no selection elements required in this section, other than to choose the target variables to consider at any given time.

To help with discussions regarding the rationale for any changes in the aggregate numbers post the first round, the forest plot provided for each target variable includes both the first and second round values, both for the individual experts and the aggregate outputs, Fig.9.8.

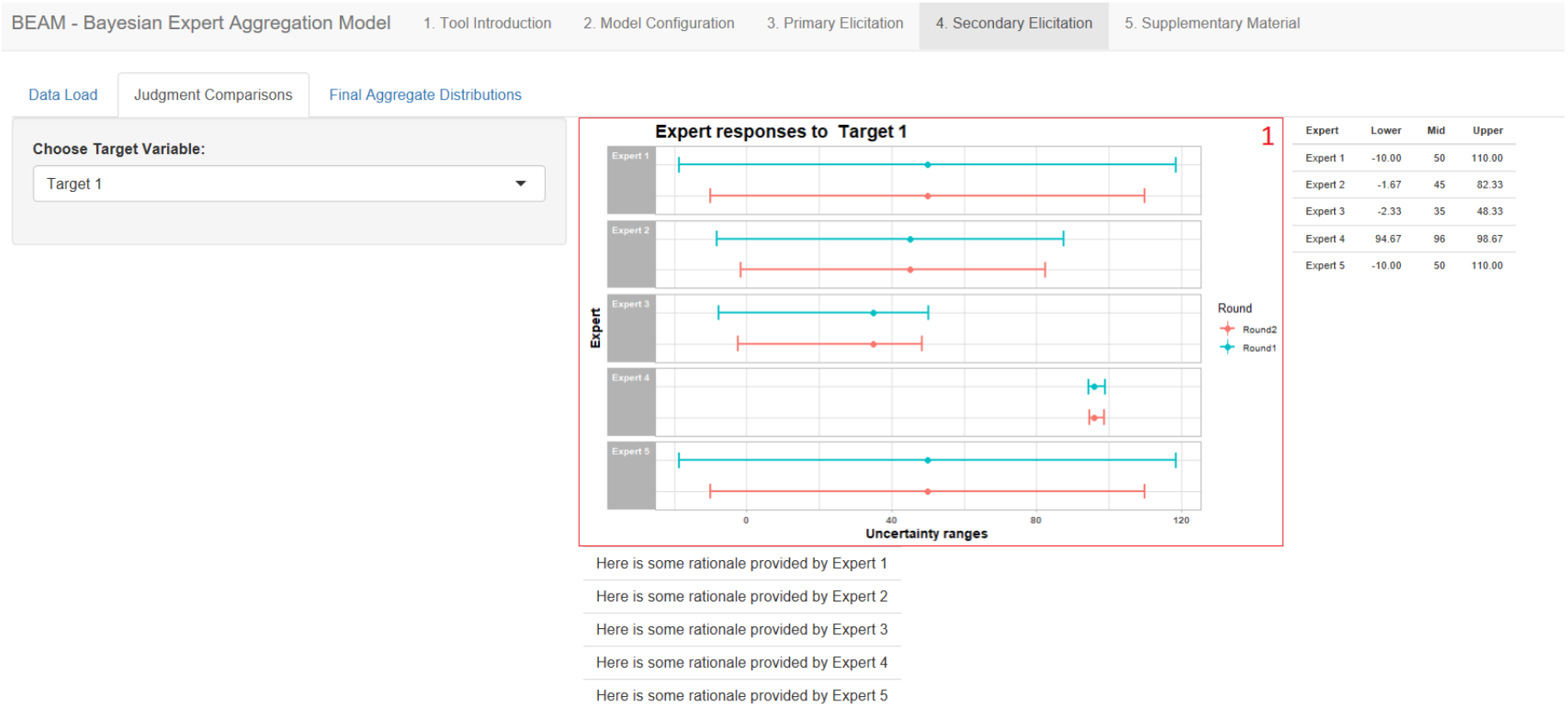


Figure 9.8: Screenshot of the BEAM judgement comparisons tab post the second elicitation process. Structure of the tab is identical to first round judgement comparisons. Round 1 and Round 2 values are shown simultaneously to help users identify where there has been changes to judgements following the group discussion (1).

Additional functionality for facilitators is provided in BEAM within the 'Supplementary Material tab'. Here, there are summaries of the information captured, overviews of individual expert calibration variables and homogeneity group values for each target variables. These are helpful for post elicitation analysis but are not on the critical path within the group setting and are therefore not included on preceding tabs.

The current version of BEAM represents a beta prototype of the tool and includes the most basic definition of the Bayesian model. Further modification could be implemented before putting the first version of the tool into production. Some of the proposed modifications include a mechanism to input Cooke's performance weighted calculations or a process for capturing the decision makers priors within the tool (currently uninformative priors are hardcoded into the underlying JAGS model). Currently the tool has been designed in isolation and so before making choices on further changes ideally the tool would be shared and discussed with the team who created the IDEA protocol. This would help with the generation of new ideas as well as providing a forum to prioritise possible changes. Significant validation of the tool is also required.

Across the landscape of SEJ, currently there are a number of pieces of software available, on a mixture of different platforms, with various levels of validation. One element that is critical to many decision making contexts is security and so for broad use it is probably important that software is not web-based. However, transparency and auditability in the software itself is paramount, in order to ensure adoption in many different contexts. Work needs to be done to harmonise the existing approaches, and any new models as they develop, into a single toolkit for analysts and decision makers alike. Integrating the software in this way could provide the support necessary to further enhance the procedural guidance, such as EFSA and IDEA, that is already available. BEAM attempts to do this for a small number of approaches, but there is significant opportunity to include a wide variety of methods. A comprehensive "meta-software" would further help to embed the use of structured expert judgement into currently untapped contexts.

Chapter 10

Discussion

10.1 Summary

The preceding text has outlined the application of a new Bayesian approach to aggregating expert judgements, and its ability to supplement existing models, by:

- Assessing the extent to which experts display systemic over or under confidence.
- Minimising potential overconfidence for the DM that arises from the impact of correlation between expert judgements driven by shared knowledge and common professional backgrounds.
- Emphasising the underlying consensus between experts whilst reflecting the diversity of judgements.
- Providing a fully parametrised posterior distribution that is easy to integrate into further analysis.

The framework has been assessed in detail against a small number of studies and then at a macro level across many studies within the Delft database.

This analysis has shown that such new Bayesian frameworks can be practical, unlike many preceding Bayesian approaches, and can be implemented without a significant overhead in defining complex priors. Utilising relatively diffuse priors (consistent across studies), has been shown to provide results on a similar order of magnitude to current approaches. This would also support the potential of applications of the Bayesian approach in contexts where the aggregate distribution is designed to emulate a rational scientists perspective in addition to those where

a specific DM, potentially with significant *a priori* belief and consequently tighter priors, exists.

The outputs of a Bayesian model of expert judgement have been compared across studies to the performance weighting approach of Cooke's Classical model. This comparison has shown that the resultant outputs of the Bayesian approach typically do not vary substantially from the performance weighted approach when only the median point is considered, however emphasise a different perspective of the uncertainty. Consistent with other analysis of the Bayesian approach (Hartley and French [2021]), the Bayesian model displays a unimodal posterior, with narrower shoulders than an equal weighted approach (as it emphasises underlying consensus) and has fatter tails than the performance weighted approach (as it usually highlights systemic overconfidence of experts).

Through cross validation we have shown that, as we might expect *a priori* given its structure, the Bayesian model demonstrates higher statistical accuracy than the performance weighted approach, but lower informativeness. This suggests that based on the decision making context the potential sensitivity to each of these metrics may impact the choice of model considered.

By considering the single combined score metric (the product of the information and statistical accuracy), we have seen that the performance weighted approach once again stands up to scrutiny and outperforms the Bayesian framework, when configured in this particular way, in the majority (circa 2/3) of cases. There are however, a substantial number of cases (circa 1/3) for which the Bayesian model outperforms the performance weighted approach lending credibility to the usage of the Bayesian model in general.

Finally, this thesis has made some recommendations with regards to how a Bayesian SEJ study should be conducted. Specifically commenting on the integration of the model into the IDEA protocol. Uniquely Bayesian considerations, such as prior ownership, have been outlined. The (B)ayesian (E)xpert (A)ggregation (M)odel R-Shiny app has been created to enable facilitators to deploy a Bayesian model within the IDEA framework easily.

Overall this research has demonstrated that the goal of a practical generic Bayesian framework for mathematical aggregation of expert judgement is feasible, and can produce reasonable results when compared to current best in class approaches even when considered broadly with a single set of parameterisations/priors. Much more work is required to assess:

- The impact of the number of seed variables/experts.

- Different parameterisations and priors within the generic framework.
- Approaches for dealing with variables on different scales.
- The drivers of out/under performance relative to performance weighted approaches.

However, we have now shown that there is sufficient evidence that the application of resources to assessing these areas is justified.

The performance weighted approach outlined by Cooke clearly remains the exemplar in this space for many applications, however, we now have a Bayesian approach which can provide a different perspective, add value for DM's with specific needs and which I hope will continue to evolve and challenge the performance weighted method.

10.2 Future work

There are many different ways the work in this thesis could be built upon. This section will describe a few recommended areas for further research.

10.2.1 Application to other datasets

The focus of research presented in this thesis has been on the application of a Bayesian approach to data from Cooke's database. Significant benefit would be generated by applying the model to SEJ studies that have been conducted in other contexts.

Utilising data generated as part of the IARPA competition would provide another mechanism by which to compare the Bayesian approach. Here, as many different elicitation protocols were already assessed as part of the competition, there would be opportunity to compare the Bayesian approach to a broader array of other methodologies, such as those outlined in Tetlock and Gardner [2016].

In addition to applying the method retroactively, the Bayesian approach would also benefit from being utilised for a bespoke SEJ problem. Designing a study from inception with a Bayesian framework would allow process integration elements to be better tested. In particular, utilising the framework in a setting with informative DM prior beliefs, would further highlight the differentiated outputs from the approach.

This Bayesian framework is shortly to be implemented within an SEJ study assessing the likelihood of reaching carbon neutrality in the UK across a selection of time horizons. The BEAM tool will be utilised within this process.

A Bayesian IDEA implementation is also being considered for assessments on the probability of technical and regulatory success within pharmaceutical, and vaccines, research and development.

10.2.2 Inconsistent scales

The Bayesian approach for recalibration outlined relies on consistent scales for variables. Multiplicative inflation factors could not be utilised on items measured with logarithmic scales. Transforming seed variables to allow for recalibration when mixed scales are used requires different methodologies.

When quantiles expressed are on a logarithmic, rather than uniform, scale one approach would be to test with experts whether they judge variables to be log-normally distributed. Here, if target variable $X \in \mathbf{X}$ is log-normally distributed then $\tilde{X} = ln(X)$ will have a normal distribution. In these circumstances, multiplicative inflation factors applied to quantiles of $\tilde{X}$, rather than X, would remain logically consistent. This could be handled relatively easily by transforming these variables before passing them through the remainder of the Bayesian model. This transformation would then need to be reversed before the final results are shared. This is an effective mechanism, in this unique case, but does not generalise.

Another approach postulated in Wiper and French [1995] is the most general method and the most exciting for application in an expert judgement setting. In Wiper and French [1995], the authors propose that to create a consistent scale for variables rather than using simple linear or log transformations, the DM's prior can be used. Here, elicited quantiles from experts are passed through the DM prior before being entered into the Bayesian model. Consequently all variables are transformed onto a single domain, the closed interval [0,1].

This approach is appealing for a number of reasons. Firstly, it means that no distributional parameterisations need to be discussed or agreed with experts. The only relevant parametrised distribution is the decision maker's. Expert's quantiles are all considered relative to the percentiles the DM would have ascribed them. This nicely changes the emphasis of the model, even more directly associating it with updating decision maker's belief. The second benefit is that in this context multiplicative inflation factors will still hold.

The challenge with this approach is that model parameterisation and hyper-parameter priors become much more conceptually complex. The easiest of these is the global prior. In the model outlined in this thesis, the global model prior, that associated with the highest hierarchical level, represents the prior belief of the decision maker on the variable domain. If variables were all transformed via

the decision maker's prior before modelling, the decision maker's prior in the DM domain, must be uniform on the interval [0,1].

Choices on the parameterisation of experts' beliefs within the decision maker domain, and the associated hierarchical model structure from here must be carefully considered. One approach would be to suggest that experts are parametrised uniformally within this space, more akin to (5.9), and recalibrated accordingly. Homogeneity groups could still then be structured by assuming that experts adjusted DM percentiles are drawn from a normal distribution representing the groups perspective and the highest level of the model represents the normal distribution from which the groups values are drawn.

Hyperparameters for these hierarchical layers must be carefully chosen as they, by definition, already encode significant DM belief. Ensuring that they remain theoretically robust is a challenge. Early attempts made to implement a model utilising this method had issues with convergence. There is significant scope for further research here however, and this remains the most exciting potential approach for addressing inconsistent scales.

10.2.3 Different parameterisations

The split normal parameterisation outlined within this thesis makes the model structure easily interpretable by most experts. It also creates a simple encoding when programming the model into JAGS. It was demonstrated, for a single study, that this parameterisation also allowed for the number of quantiles to be minimised without significant impact to the resulting distributions. As outlined early however, this parameterisation is ad-hoc and many other potential parameterisations could be used within the generic framework outlined.

One important aspect of future research is to identify the impact of different choices of parameterisation. Cross-validation of methodologies should be utilised to identify the efficacy of different approaches to ensure that poorly performing ones are not repeated within applications with real world consequences.

There will be no single version of the model appropriate for every context, as outlined in Chapter 8, the experts and DM should have a perspective on the way that their perception of the uncertainty is encoded. This should not be completely unstructured however. It is recommended that a small set of potential encodings is created and tested. This group of model parameterisations would represent the discrete options available at the start of each elicitation. The best choice for the specific context could then be made.

Once this small group of model approaches were created, this could then be

continuously assessed and updated. For a further parameterisation to be added to the list, the following principles should be adhered to;

- The parameterisation must be proven to be *robust*, i.e. provide comparable scores to other approaches, through cross-validation of historic studies.
- The approach must be *sufficiently differentiated* vs. other parameterisations currently available.
- The approach must be *explainable* to an expert panel.
- Once encoded, the approach must be *frugal* and not rely on an unreasonable amount of data or an unfeasible number of priors to be created in order to be effective.

These principles could also be applied when selecting the first set of discrete options.

Potential examples of other parameterisations would include; an approach which creates a more uniform distribution between elicited quantiles (akin to Equation 5.9), an adjusted beta based parameterisation which could be used for bounded variables and a logarithmic based approach. Each of these has a substantially different interpretation of the underlying uncertainty and would be more applicable than the split normal in specific circumstances.

More complex approaches would consider multimodal distributions. For some variables uni-modality will be inappropriate and the hierarchical framework outlined could be utilised to create a parametrised multimodal distribution. One particular example of this is in the use of expert judgement to assess clinical trial efficacy (Best et al. [2020], Dallow et al. [2018], Kinnersley and Day [2013]). Let us assume we are looking to judge the effect of a drug on a particular clinical trial endpoint. Often the expected outcome could be multimodal as there will be a probability that the drug does not promote the biological response anticipated i.e. it simply does not work, thereby having significant weight at zero. Conditional on success, the extent of the anticipated response will be distributed across some range, with a peak at the most likely value. This can be thought of as a multimodal distribution with a mode at zero representing the probability of failure and a mode at the most likely value conditional on success.

This could be encoded into the outlined framework by considering this as two separate elicitations, which are both unimodal, and then combining the results at the end, weighting appropriately, to create a single posterior. The extent to which these more complex structures may occur should be assessed, and then where

appropriate, further functionality added into the BEAM tool to allow these to be generated as standard.

10.2.4 Other correlation effects

Bayesian approaches to SEJ typically allow for more complex dynamics to be considered. This thesis has outlined how decision makers priors and overconfidence driven by inter-expert correlation can be incorporated, both of which are typically ignored within non-Bayesian methods. Other correlation effects have not been explicitly considered in the Bayesian model outlined. Two of these effects; Decision maker to expert correlation and target or calibration variable interdependency warrant further examination.

Similar to inter-expert dependence, decision maker to expert correlation could also be a key driver of overconfidence. SEJ studies should enable better decisions by enhancing decision maker understanding of uncertainty on the key variables of interest. As per standard Bayesian theory, when decision makers have strong priors, the final posterior will be determined by this. Experts who are directly correlated with decision makers may end up reinforcing this perspective without bringing any new information. Here, the rationale for the expert's perspective is already captured in the definition of the prior.

In the most extreme example of this, imagine a Decision maker going to a single expert to ask their opinion on target variable X. Let us further assume the DM's knowledge about X was entirely based on a piece of published literature, Article A. The Decision maker is unsure of the robustness of Article A however, and therefore approaches our expert to get their input. The DM's prior is thus encoded as the expected value for X outlined in A, with some broader uncertainty around it. If the expert also has knowledge of X entirely determined by Article A, their elicited values would reflect this. The decision maker, upon hearing the experts view, believes this additional perspective is aligned with what they read suggesting their reticence was unfounded. The DM therefore updates their prior accordingly, becoming more confident in their belief. In this case, the decision maker is reducing their perception of uncertainty without any further data being introduced. In many organisations, there can be an individual colloquially termed as a "yes-man/yes-person", who always agrees with their superior (French [1980]). This can be very dangerous in simply reflecting DM's views back to them, providing artificial reinforcement and increasing the risk of poor decisions being taken forward.

Decision maker to expert interdependence can either be considered as a modelling exercise or form part of expert selection. Further work is necessary to under-

stand the scale of impact and the best potential remedies.

Another potential dependence effect not captured within the current model formulation is inter-variable correlation. Dependence between calibration and target variables is fundamental for both creating homogeneity groups and to manage overconfidence. Dependence between target variables, however, has not been considered. Trying to elicit variable dependence structures is often very difficult. If entire correlation matrices are desired a vast number of values often need to be elicited. Elicitation of dependence structures will often consider copula based approaches, with vine-copulas recently being explicitly researched (Wilson [2018]). Further research is required to identify the best methods for dependence elicitation and how this can be integrated into a broader Bayesian model as outlined here.

10.2.5 Extending BEAM

The BEAM R-shiny tool, as currently structured, allows experts to easily run our Bayesian approach alongside quantile aggregation and equal weighted linear opinion pooling methods, integrated into the IDEA protocol. There exists significant opportunity to extend BEAM further, in isolation, and enhance the options available for study facilitators. There is then additional opportunity to integrate BEAM into a broader system architecture to provide a holistic SEJ expert elicitation and aggregation environment for decision makers.

The top priorities for extending BEAM as a tool are:

- *Further alignment to the* IDEA *protocol.* BEAM would benefit from a broader number of stakeholders inputting into the design. Currently tool flow has been designed based on personal experiences of common problems encountered in expert elicitation and in deploying IDEA. Significant benefit would be gained by having the original architects of the IDEA protocol influencing design choices made. A project to engage with this team would support prioritisation of areas for further development.

- *Dynamic creation of decision maker priors.* DM prior belief is a key component of the Bayesian approach, but currently within BEAM, these all need to be defined *a priori* and hard-coded into the underlying R-script. This is a barrier to use. Creating a mechanism within the user-interface by which priors over modelling hyper-parameters can be defined is key to ensuring that these are robustly considered as part of the tool workflow.

- *Saving studies.* The current version of BEAM creates static output tables within the summary section which can be kept for posterity, but requires all preceding

steps within the model to be conducted in a single pass. There is currently no ability to save progress within the tool and return at a later point with the same settings. As the tool increases in complexity, and more of the undertaken activities are conducted asynchronously, the need to save and return to work would increase. The best approach for doing this with BEAM requires some careful design.

There would also be advantage to integrating Cooke's Classical model into the tool, although this would start moving more towards the meta-tool approach. Attempts to build R versions of Cooke's method have been unsuccessful so far and the current two robust approaches are the original EXCALIBUR tool, and a MATLAB version. Rather than attempting to rebuild this functionality within R, the first step to integrating the Classical method into BEAM could be to provide the functionality to upload outputs of EXCALIBUR. The performance weighted decision maker values could then be displayed alongside the other aggregation methods, but the calculation engine would remain in the core system.

In the longer term further integration of SEJ processes, modelling approaches and tools is desired. No single SEJ approach is going to be appropriate in all contexts and so clearer delineation is required. This would extend much wider than BEAM/EXCALIBUR, or indeed IDEA/EFSA. Decision Makers and expert judgement practitioners would get considerable value with broader visibility to the different methods and processes available for SEJ, which currently would require a significant literature review to find.

My vision is that a *Structured Expert Judgement Cockpit* is built. The SEJ Cockpit would be a single place that would bring together the currently disparate methods and enable easy exploration of each. Examples applications for each process and method would be provided and the benefits/ limitations would be transparently outlined. Both behavioural and mathematical aggregation processes would be available. The Cockpit would have links to any existing tools both for elicitation, e.g. MATCH and aggregation e.g. EXCALIBUR/BEAM. In the long run these tools could be integrated more formally, similar to what is described for the three aggregation models in BEAM. In the short term having them all in the same place would be a significant step forward.

This meta-cockpit could also serve as a repository for all of the softer elements of SEJ required in order to ensure studies run effectively. Online training materials for facilitators could be integrated directly into the cockpit and links to class-room training, where available, provided. A practitioner database could also be included to link users directly to each other and to academic experts in order to ensure that

activities are deployed appropriately.

Validation would be a key component, both for the methods included and the tools that sit within the Cockpit. A governance body, cross academia and industry, could be formed in order to provide oversight to what is included.

Whilst the Cockpit would be online, the tools captured within should all be operated offline, or inside company IT firewalls. Expert judgement study data is often extremely sensitive and therefore enforcing study data to be entered into an online system would create significant security risk and limit adoption. There should however, be the opportunity within all of the tools underpinning the Cockpit to have the data (and if necessarily the study details) obfuscated and then submitted back to a central warehouse. This would ensure that no sensitive data is shared, but would create a body of applications that could then be used either to inform training activities or, more importantly to form the basis for further cross-validation work.

Integrating tools together and providing a mechanism for study data to be captured more broadly, expands the work underway in the DELFT database. Validation/cross-validation is key to building trust for SEJ as a scientific discipline in the longer term. Minimising barriers to validation being performed is fundamental. Cross-validation results should be visible.

Whilst there have been many applications of SEJ, broad awareness remains low. A holistic solution like the SEJ Cockpit would support in raising the profile of SEJ and hopefully create more traction for this exciting discipline.

10.2.6 Expert judgement and knowledge management

Much of the literature on expert knowledge today focuses on the fields of public policy definition or accidental risk analysis; (Cooke and Goossens [2000], Stiber [2004], Cárdenas et al. [2013], French [2012], Aspinall [2006], Drescher et al. [2013]). In each of these cases it is typically a complex problem, however, there are well defined experts and the problems are typically time bound in nature.

Conversely, in knowledge management (KM) literature, see (Alavi and Leidner [2001], Gold et al. [2001], Hedlund [1994], Petrash [1996]) for an excellent review, the problems usually addressed are those related to continuous enterprise, how do we take data and turn it into information and then correspondingly knowledge (here knowledge can be either procedural or human knowledge) in a systematic way.

It is proposed here, although these two disciplines are often kept separate, there is a lot of overlap between the two which is under researched. Whilst the knowledge capture in KM today is predominantly descriptive (for example the Mc-

Donalds operating manual (Alavi and Leidner [2001]), in the future, it is quite likely organisations will get advantages from enterprise wide, routine capture of expert probability distributions. One of the key challenges with this will be how to incentivise employees to give this knowledge systematically, how very large numbers of probability distributions can be assessed and calibrated and thirdly how to develop models and tools that can deal with the increased complexity of this new data. Expert judgement research will need to advance the tool sets for group decision making, develop procedures for systematic expert problems and do further research into expert identification this will assist organisations on the journey to evidence based management (Hewison [2004], Pfeffer and Sutton [2020], Walshe and Rundall [2001]). This in turn can impact how we think about decision support technology as it integrates with commercial organisations, (Carlsson [2002], Shim et al. [2002]).

Bibliography

Adams-Hosking, C., McBride, M.F., Baxter, G., Burgman, M., de Villiers, D., Kavanagh, R., and McAlpine, C.A., 2016. Use of expert knowledge to elicit population trends for the koala (Phascolarctos cinereus). *Diversity and Distributions*, 22: pp.249–262.

Alavi, M. and Leidner, D.E., 2001. Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly*, pp.107–136.

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K. and Rousseau, J., 2012. Combining expert opinions in prior elicitation. *Bayesian Analysis*, 7(3): pp.503–532.

Angner, E., 2006. Economists as experts: Overconfidence in theory and practice. *Journal of Economic Methodology*, 13(1): pp.1-24.

Arlot, S. and Celisse, A., 2010. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4: pp.40–79.

Aspinall, W.P., 2006. Structured elicitation of expert judgement for probabilistic hazard and risk assessment in volcanic eruptions. *Statistics in volcanology*, 1: pp.15–30.

Aspinall, W.P., 2021. Reminiscences of a Classical Model expert elicitation facilitator. *State of the Art in Structured Expert Judgment*, Preprint.

Bamber, J. L. and Aspinall, W.P., 2013. An expert judgement assessment of future sea level rise from the ice sheets *Nature Climate Change*, 3(4): p.424.

Bamber, J. L., Aspinall, W.P. and Cooke, R.M., 2016. A commentary on "how to interpret expert judgment assessments of twenty-first century sea-level rise" by Hylke de Vries and Roderik SW van de Wal. *Climatic Change*, 137(3-4): pp.321-328.

Barnett, V., 1999. *Comparative Statistical Inference*. Chichester: John Wiley and Sons.

Benedetti, R., 2010. Scoring rules for forecast verification. *Monthly Weather Review* 138(1): pp.203-211.

Best, N., Dallow, N. and Montague, T., 2020. Prior Elicitation. *Bayesian Methods in Pharmaceutical Research.* CRC Press.

Bhaskar, N.U., Naidu, P.P., Babu, S.R.C. and Govindarajulu, P., 2011. General principles of user interface design and websites. *International Journal of Software Engineering (IJSE)*, 2(3): pp.45-60.

Billari, F.C., Graziani, R., and Melilli, E., 2014. Stochastic population forecasting based on combinations of expert evaluations within the Bayesian paradigm. *Demography*, 51(5): pp.1933-1954.

Blair-Early, A. and Zender, M., 2008. User interface design principles for interaction design. *Design Issues*, 24(3): pp.85-107.

Bolger, F. and Rowe, G., 2015. The aggregation of expert judgment: Do good things come to those who weight?. *Risk Analysis*, 35(1): pp.5–11.

Bolger, F. and Rowe, G., 2015. There is data, and then there is data: only experimental evidence will determine the utility of differential weighting of expert judgment. *Risk analysis: an official publication of the Society for Risk Analysis*, 35(1): p.21.

Booker, J.M. and Meyer, M.A., 2014. Sources of and effects of interexpert correlation: an empirical study. *IEEE transactions on systems, man, and cybernetics*, 18(1): pp.135–142.

Bradley, P.S. and Fayyad, U.M., 1998. Refining Initial Points for K-Means Clustering. *ICML* , 98: pp.91–99.

Brenner, L., Griffin, D. and Koehler, D.J., 2005. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. Organizational Behavior and Human Decision Processes *Organizational Behavior and Human Decision Processes*, 97(1): pp.64-81.

Browne, M.W., 2000. Journal of mathematical psychology. *Conservation Letters*, 44(1): pp.108–132.

Burgman, M., Carr, A., Godden, L., Gregory, R., McBride, M., Flander, L. and Maguire, L., 2011. Redefining expertise and improving ecological judgment. *Conservation Letters*, 4(2): pp.81–87.

Burgman, M.A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. and Twardy, C., 2011. Expert status and performance. *PLoS One*, 6(7).

Burgman, M.A., 2016. *Trusting judgements: how to get the best out of experts.* Cambridge University Press.

Camerer, C.F. and Johnson, E.J., 1997. 10 the process-performance paradox in expert judgment: How can experts know so much and predict so badly? *Research on judgment and decision making: Currents, connections, and controversies*, p.342.

Cárdenas, I.C., Al-jibouri, S.S.H., Halman, J.I.M. and van Tol, F.A., 2013. Capturing and Integrating Knowledge for Managing Risks in Tunnel Works *Risk Analysis*, 33(1): pp.92–108.

Carlsson, C. and Turban, E., 2002. DSS: directions for the next decade *Decision Support Systems*, 33(2): pp.105–110.

Celisse, A. and Robin, S., 2008. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics & Data Analysis*, 52(5): pp.2350–2368.

Charniak, E., 1991. Bayesian networks without tears. *AI magazine*, 12(4): p.50.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A. and Charrad, M.M., 2014. Package 'nbclust'. *Journal of statistical software*, 61: pp.1-36.

Chen, H., Leung, C.-C., Xie, L., Ma, B., Li, H., (2015). Parallel inference of dirichlet process Gaussian mixture models for unsupervised acoustic modeling: a feasibility study *INTERSPEECH-2015*, pp.3189–3193.

Chesley, G.R., 1975. Elicitation of subjective probabilities: a review. *The Accounting Review*, 50(2): pp.325–337.

Clemen, R.T., 1999. Calibration and the aggregation of probabilities. *Management Science*, 32(3): pp.312–314.

Clemen, R.T., 2008. Comment on Cooke's classical method. *Reliability Engineering & System Safety*, 93(5): pp.760-765.

Clemen, R.T. and Lichtendahl, K.C., 2002. Debiasing expert overconfidence: A bayesian calibration model. Citeseer.

Clemen, R.T. and Winkler, R.L., 1999. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2): pp.187–203.

Colson, A.R. and Cooke, R.M., 2017. Cross validation for the classical model of structured expert judgment *Reliability Engineering & System Safety*, 163: pp.109–120.

Cooke, R.M., 1991. *Experts in uncertainty: opinion and subjective probability in science.* Oxford University Press on Demand.

Cooke, R.M., 2007. Expert Judgement Studies. *Reliability Engineering and System Safety.*, 93: pp.655-777.

Cooke, R.M., 2014. *Validating Expert Judgment with the Classical Model*, pp.191–212. Springer.

Cooke, R.M., 2008. Response to discussants. *Reliability Engineering & System Safety*, 93(5): pp.775-777.

Cooke, R.M., 2016. Supplementary Online Material for Cross Validation of Classical Model for Structured Expert Judgment.

Cooke, R.M. and Goossens, L.H.J., 2000. Procedures guide for structural expert judgement in accident consequence modelling. *Radiation Protection Dosimetry*, 90(3): pp.303–309.

Cooke, R.M. and Goossens, L.H.J., 2004. Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research*, 7(6): pp.643–656.

Cooke, R.M., and Goossens, L.H., 2008. TU Delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5): pp.657-674.

Cooke, R.M., ElSaadany, S. and Huang, X., 2008. On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering & System Safety*, 93(5): pp.745–756.

Cooke, R.M., Marti, D. and Mazzuchi, T., 2020. Expert forecasting with and without uncertainty quantification and weighting: What do the data say? *International Journal of Forecasting.*

Cooke, R.M. and Solomatine, D., 1992. EXCALIBUR Integrated System for Processing Expert Judgements version 3.0. *Delft University of Technology and SoLogic Delft, Delft.*

Cooke, R.M., Wittmann, M.E., Lodge, D.M., Rothlisberger, J.D., Rutherford, E.S., Zhang, H. and Mason, D.M., 2014. Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie. *Integrated Environmental Assessment and Management*, 10(4): pp.522-528.

Cox. D.R., 1958. Two further applications of a model for binary regression. *Biometrika*, pp.562–565.

Dalkey, N. and Helmer, O., 1963. An experimental application of the Delphi method to the use of experts. *Management science*, 9(3): pp.458–467.

Dallow, N., Best, N. and Montague, T.H., 2018. Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical Statistics*, 17(4): pp.301–316.

Dawid, A.P., 1982. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379): pp.605–610.

Dawid, A.P., DeGroot, M.H., Mortera, J., Cooke, R., French, S., Genest, C., Schervish, M.J., Lindley, D.V., McConway, K.J. and Winkler, R.L., 1995. Coherent combination of experts' opinions. *Test*, 4(2): pp.263–313.

De Finetti, B., 1974. *Theory of Probability.* Chichester: John Wiley and Sons.

De Finetti, B., 1975. *Theory of Probability.* Chichester: John Wiley and Sons.

De Groot, M.H., 2005. *Optimal statistical decisions.* John Wiley & Sons. Vol.82.

Delavande, A. and Rohwedder, S., 2008. Eliciting subjective probabilities in internet surveys. *Public Opinion Quarterly*, 72(5): pp.866–891.

de Vries, H. and van de Wal, R.SW., 2015. How to interpret expert judgment assessments of 21st century sea-level rise. *Climatic change*, 130(2): pp.87-100.

de Vries, H. and van de Wal, R.SW., 2016. Response to commentary by JL Bamber, WP Aspinall and RM Cooke (2016). *Climatic change*, 137(3-4): pp.329-332.

Dias, L.C., Morton, A. and Quigley, J., 2018. *Elicitation.* Springer International Publishing.

Drescher, M., Perera, A.H., Johnson, C.J., Buse, L.J., Drew, C.A. and Burgman, M.A., 2013. Toward rigorous use of expert knowledge in ecological research. *Ecosphere*, 4(7): p.83.

EFSA 2010. Statement of EFSA on the possible risks for public and animal health from the contamination of the feed and food chain due to possible ash-fall following the eruption of the Eyjafjallajökull volcano in Iceland. *EFSA Journal*, 8:1593.

EFSA., 2014. Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*.

Eggstaff, J.W., Mazzuchi,T.A. and Sarkani, S., 2014. The effect of the number of seed variables on the performance of cooke's classical model. *Reliability Engineering & System Safety*, 121: pp.72–82.

Fiske, S.T. and Berdahl, J., 2007. *Social power.*

Flandoli, F., Giorgi, E., Aspinall, W.P. and Neri, A. 2011. Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety*, 96(10): pp.1292-1310.

French, S., 1980. Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society. Series A (General)*, 143(1): pp.43–48.

French, S., 1985. Group consensus probability distributions: A critical survey. *Bayesian statistics*, 2: pp.183–202.

French, S., 1986. Calibration and the expert problem. *Management Science*, 32(3): pp.315–321.

French, S., 2011. Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 105(1): pp.181–206.

French, S., 2012. Expert judgment, meta-analysis, and participatory risk analysis. *Decision Analysis*, 9(2): pp.119–127.

French, S., 2013 Cynefin, Statistics and Decision Analysis. *Journal of the Operational Research Society*, 64(4): pp.547–561.

French, S. and Rios Insua, D., 2000 *Statistical Decision Theory.* London: Arnold.

French, S. and Argyris, N., 2018 Decision Analysis and Political Processes. *Decision Analysis*, 15(4): pp.208–222.

French, S., Bedford, T., Pollard, S.J.T. and Soane, E., 2011. Human reliability analysis: A critique and review for managers *Safety science*, 49(6): pp.753–763.

French, S., Maule, J. and Papamichail, N., 2009. *Decision behaviour, analysis and support.* Cambridge University Press.

French, S., Walmod-Larsen, O. and Sinkko, K., 1993. Decision conferencing on countermeasures after a large nuclear accident. *No. RISO-R–676 (EN)*, Risoe National Lab.

Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2): pp.137–146.

Galton, F., 1907. *Vox populi.*

Garthwaite, P.H., Kadane, J.B. and O'Hagan, A., 2005. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470): pp.680–701.

Gelman, A., 2004. Prior distributions for variance parameters in hierarchical models. *Report, EERI Research Paper Series.*

Genest, C., 1984. A conflict between two axioms for combining subjective distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.403–405.

Genest, C. and McConway, K.J., 1990. Allocating the weights in the linear opinion pool. *Journal of Forecasting*, 9(1): pp.53–73.

Genest, C. and Zidek, J.V., 1986. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, pp.114–135.

Gigerenzer, G., 2011. What are natural frequencies? *Bmj*, 343: p.d6386.

Gigerenzer, G. and Edwards, A., 2003. Simple tools for understanding risks: from innumeracy to insight. *Bmj*, 327(7417): pp.741-744.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): pp.243-268.

Gold, A.H., Malhotra, A. and Segars, A.H., 2001. Knowledge management: an organizational capabilities perspective. *Journal. of Management Information Systems*, 18(1): pp.185–214.

Goodwin, P. and Wright, G., 2010. The limits of forecasting methods in anticipating rare events. *Technological forecasting and social change*, 77(3): pp.355-368..

Gordon, K., 1924. Group Judgments in the Field of Lifted Weights. *Journal of Experimental Psychology*, 7(5): p.398.

Gordon, T. and Pease, A., 2006. RT Delphi: An efficient,"round-less" almost real time Delphi method. *Technological Forecasting and Social Change*, 73(4): pp.321–333.

Gosling, J.P., 2018. SHELF: the Sheffield elicitation framework. *Elicitation*, Springer, Cham., pp.61–93.

Gosling, J.P., Oakley,J.E., and O'Hagan, A., 2007. Nonparametric elicitation for heavy-tailed prior distributions. *Bayesian Analysis*, 2(4): pp.693–718.

Gosling, J.P., Hart, A., Mouat,D.C., Sabirovic, M., Scanlan, S. and Simmons, A., 2012. Quantifying experts' uncertainty about the future cost of exotic diseases. *Risk Analysis*, 32(5): pp.881–893.

Green, K.C., Armstrong, J.S. and Graefe, A., 2007. Methods to elicit forecasts from groups: Delphi and prediction markets compared.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E. and Nelson, C., 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1): p.19.

Hammitt, J.K. and Zhang, Y., 2013. Combining experts' judgments: Comparison of algorithmic methods using synthetic data. *Risk Analysis: An International Journal*, 33(1): pp.109–120.

Hanea, A.M., McBride, M.F., Burgman, M.A. and Wintle, B.C., 2016. Classical meets modern in the IDEA protocol for structured expert judgement. *Journal of Risk Research*, pp.1—17.

Hanea, A.M., McBride, M.F., Burgman, M.A., Wintle, B.C., Fidler, F., Flander, L., Twardy, C.R., Manning, B. and Mascaro, S., 2017. I nvestigate D iscuss E stimate A ggregate for structured expert judgement. *International journal of forecasting*, 33(1): pp.267–279.

Hanea, A.M., Burgman, M. and Hemming, V., 2018. IDEA for uncertainty quantification. *Elicitation*, Springer, Cham, pp.95–117.

Hartley, D.S. and French, S., 2018. Elicitation and calibration: a Bayesian perspective *Elicitation*, pp.119–140. Springer, Cham.

Hartley, D.S. and French, S. 2021. Bayesian modelling of dependence between experts: some comparisons with Cooke's Classical Model. *Expert Judgement in Risk and Decision Analysis*, Springer, Cham - In press.

Hartley, D.S. and French, S. 2020. A Bayesian method for calibration and aggregation of expert judgement. *In Submission* (TBC).

Hedlund, G., 1994. A model of knowledge management and the N-form corporation. *Strategic management journal*, 15(S2): pp.73–90.

Helmer-Hirschberg, O. Analysis of the future: The Delphi method. 1967.

Hemming, V., Burgman, M.A., Hanea, A.M., McBride, M.F. and Wintle, B.C., 2018. A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, 9(1): pp.169–180.

Hewison, A., 2004. Evidence-based management in the NHS: is it possible? *Journal of health organization and management*, 18(5): pp.336–348.

Hockey, G. R. J., Maule, A.J., Clough, P.J. and Bdzola, L., 2000. Effects of negative mood on risk in everyday decision making. *Cognition and Emotion*, 14: pp.823-856.

Hora, S., 2007. *Eliciting probabilities from experts. Advances in Decision Analysis: From Foundations to Applications.* Edwards, W. , Miles, R. F. and Von Winterfeldt, D. Cambridge: Cambridge University Press: pp.129-153.

Hsu, Y.L., Lee, C.H. and Kreng, V.B., 2010. The application of Fuzzy Delphi Method and Fuzzy AHP in lubricant regenerative technology selection. *Expert Systems with Applications*, 37(1): pp.419–425.

Jeffreys, H., 1961. *Theory of Probability.* Oxford: Oxford University Press.

Jouini, M.N. and Clemen, R.T., 1996. Copula models for aggregating expert opinions. *Operations Research*, 44(3): pp.444–457.

Kadane, J.B., 1982. The Well-Calibrated Bayesian: Comment. *Journal of the American Statistical Association*, 77(379): pp.610–611.

Kadane, J.B. and Fischhoff, B., 2013. A cautionary note on global recalibration. *Judgment and Decision Making*, 8(1): p.25.

Kahneman, D., 2011. *Thinking, fast and slow.* Macmillan.

Kahneman, D., Slovic, P. and Tversky, A., 1982. *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press.

Kahneman, D. and Tversky, A., 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pp.263–291.

Kallen, M. and Cooke, R., 2002 Expert aggregation with dependence. *Probabilistic Safety Assessment and Management*, Elsevier.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence*, 24(7): pp.881–892.

Kaplan, S., 2000. 'combining probability distributions from experts in risk analysis'. *Risk Analysis*, 20(2): pp.155–156.

Kaustia, M. and Perttula, M., 2012. Overconfidence and debiasing in the financial industry. *Review of Behavioural Finance*

Keren, G., 1991. Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3): pp.217-273.

Kinnersley, N. and Day, S., 2013. Structured approach to the elicitation of expert beliefs for a bayesian-designed clinical trial: a case study. *Pharmaceutical statistics*, 12(2): pp.104–113.

Koehler, D.J., Brenner, L. and Griffin, D., 2002. The calibration of expert judgment: Heuristics and biases beyond the laboratory. *Heuristics and biases: The psychology of intuitive judgment*, pp.686-715.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 14(2): pp.1137–1145.

Kumar, V., Chhabra, J.K. and Kumar, D., 2014. Performance evaluation of distance metrics in the clustering algorithms. *INFOCOMP Journal of Computer Science*, 13(1): pp.38-52.

Kynn, M., 2008. The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1): pp.239–264.

Lambert, J., Bessière, V. and N'Goala, G., 2012. Does expertise influence the impact of overconfidence on judgment, valuation and investment decision?. *Journal of Economic Psychology*, 33(6): pp.1115-1128.

Lichtendahl, K. C., 2005. Bayesian Models of Expert Forecasts. *PhD Thesis*.

Lichtendahl, K.C. and Winkler, R.L., 2007. Probability Elicitation, Scoring Rules, and Competition among Forecasters. *Management Science*, 53(11): pp.1745 - 1755.

Lichtendahl Jr, K.C., Grushka-Cockayne, Y. and Winkler, R.L., 2013. Is it better to average probabilities or quantiles? *Management Science*, 59(7): pp.1594–1611.

Lichtenstein, S. and Fischhoff, B., 1980. Training for calibration. *Organizational Behavior and Human Performance*, 26(2): pp.149–171.

Lichtenstein, S., Fischhoff, B. and Phillips, L.D., 1977. Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pp.275–324, Springer, Dordrecht.

Lichtenstein, S., Fischhoff, B. and Phillips, L.D., 1982. Calibration of probabilities: the state of the art to 1980. *Judgement under Uncertainty. D. Kahneman, P. Slovic and A. Tversky*. Cambridge: Cambridge University Press: pp.306-334.

Lin J., 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1): pp.145–151.

Lin, S.-W. and Bier, V.M., 2008. A study of expert overconfidence. *Reliability Engineering and System Safety*, 93: pp.711-721.

Lin, S.-W and Cheng, C.-H., 2009. The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management*, 4(2): pp.149-161.

Lindley, D.V., 1983. Reconciliation of probability distributions. *Operations Research*, 31(5): pp.866-880.

Lindley, D.V., Tversky, A. and Brown, R.V., 1979. On the reconciliation of probability judgements (with discussion). *Journal of the Royal Statistical Society*, A142: pp.146–180.

Linstone, H.A. and Turoff, M. eds.,1975. The delphi method. *Reading, MA: Addison-Wesley*, pp.3–12.

Looney, R.E.,2004. DARPA's policy analysis market for intelligence: Outside the box or off the wall? *International Journal of Intelligence and Counterintelligence*, 17(3): pp.405–419.

Loughlin, S.C., Aspinall, W.P., Vye-Brown, C., Baxter, P.J., Braban, C., Hort, M., Schmidt, A., Thordarson, T., Witham, C., 2012. Large-magnitude fissure eruptions in Iceland: source characterisation. *BGS Open File Report, OR/12/098;* 231pp, available at: http://www.bgs.ac.uk/research/volcanoes/LakiEruptionScenarioPlanning.html.

Madansky, A., 1964. Externally Bayesian groups. *RM-4141-PR: RAND.*

Madhulatha, T.S., 2012. An overview on clustering methods. *arXiv preprint arXiv:1205.1117.*

Mather, P.M., 1976. *Computational methods of multivariate analysis in physical geography.*, John Wiley & Sons.

Mazur, A., 1973. Disputes between experts. *Minerva*, pp.243-262.

McGill. (1). http://www.medicine.mcgill.ca/epidemiology/Joseph/courses/common/Tricks.html

McKenzie, C.R., Liersch, M.J. and Yaniv, I., 2008. Overconfidence in interval estimates: What does expertise buy you?. *Organizational Behavior and Human Decision Processes*, 107(2): pp.179-191.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B. Fincher, K., Swift, S.A., 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25: pp.1106–1115.

Mérigot, B., Durbec, J.P. and Gaertner, J.C., 2010. On goodness-of-fit measure for dendrogram-based analyses. *Ecology*, 91(6): pp.1850-1859.

Meyer, M.A. and Booker, J.M., 2001. *Eliciting and analyzing expert judgment: a practical guide.* Society for Industrial and Applied Mathematics.

Miller, R.G., 1974. The jackknife-a review. *Biometrika*, 61(1): pp.1-15.

Mosleh, A. and Apostolakis, G., 1986. The assessment of probability distributions from expert opinions with an application to seismic fragility curves. *Risk Analysis*, 6(4): pp.447–461.

Morris, P.A., 1974. Decision analysis expert use. *Management Science*, 20(9): pp.1233–1241.

Morris, P.A., 1977. Combining expert judgments: A bayesian approach. *Management Science*, 23(7): pp.679–693.

Morris, D.E., Oakley, J.E. and Crowe, J.A., 2014. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52: pp.1–4.

Morton, A., Bird, D., Jones, A. and White, M., 2011. Decision conferencing for science prioritisation in the UK public sector: a dual case study. *Journal of the Operational Research Society*, 62(1): pp.50–59.

Mumpower, J.L. and Stewart, T.R., 1996. Expert judgement and expert disagreement. *Thinking and Reasoning*, 2(2/3): pp.191–212.

Nielsen, J., 1995. 10 usability heuristics for user interface design. *Nielsen Norman Group*, 1(1).

Norouzi, M., Fleet, D.J. and Salakhutdinov, R.R., 2012. Hamming distance metric learning. *Advances in neural information processing systems*, pp. 1061-1069.

Oakes, D., 1985. Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390): p.339–339.

O'Hagan, A. and Oakley, J.E., 2004. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety*, 85(1): pp.239–248.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. and Rakow, T., 2006. *Uncertain judgements: eliciting experts' probabilities.* Chichester: John Wiley & Sons.

O'Hagan, A., 2019. Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73(1): pp.69–81.

O'Leary, D.E., 1998. Knowledge acquisition from multiple experts: an empirical study *Management science*, 44(8): pp.1049-1058.

Parnell, G.S., Terry Bresnick, M.B.A., Tani, S.N. and Johnson, E.R., 2013. Decision Conferencing *Handbook of decision analysis (Vol. 6)*, John Wiley & Sons, pp.381–391.

Petrash, G., 1996. Dow's journey to a knowledge value management culture *European Management Journal*, 14(4): pp.365–373.

Perälä, T., Vanhatalo, J. and Chrysafi, A., 2019. Calibrating expert assessments using hierarchical Gaussian process models *Bayesian Analysis*, Advance Publication.

Petrolia, D.R., Nyanzu, F., Cebrian, J., Harri, A., Amato, J. and Walton, W.C., 2020. Eliciting expert judgment to inform management of diverse oyster resources for multiple ecosystem services. *Journal of Environmental Management*, 268: p.110676.

Pfeffer, J. and Sutton, R.I., 2006. Evidence-based management. *Harvard business review*, 84(1): p.62.

Phillips, L.D., 1991. Decision conferencing. *IEE Colloquium on CSCW: Some Fundamental Issue*, pp.6–1.

Picard, R.R. and Cook, R.D., 1984. Cross-validation of regression models. *Journal of the American Statistical Association*, 79:(387): pp.575–583.

Poses, R.M., Bekes, C., Winkler, R.L., Scott, W.E. and Copare, F.J., 1990. Are Two (Inexperienced) Heads Better Than One (Experienced) Head?: Averaging House Officers' Prognostic Judgments for Critically Ill Patients. *Archives of Internal Medicine*, 150:(9), pp.1874–1878.

Ramsey, F.P., 1926. *Truth and Probability. The Foundations of Mathematics and Other Logical Essays.* R. B. Braithwaite: Harcourt, Brace and Co.

Rausch, E., Cassidy, M.F. and Buede, D., 2009. Does the accuracy of expert judgment comply with common sense: caveat emptor. *Management Decision.*

Rohlf, F.J. and Fisher, D.R., 1968. Tests for hierarchical structure in random data sets. *Systematic Biology*, 17(4): pp.407-412.

Ronen, A. and Wahrmann, L., 2005. Prediction games. *International Workshop on Internet and Network Economics*, Springer, Berlin pp.129–140.

Rowe, G. and Wright, G., 2001 Expert opinions in forecasting: the role of the Delphi technique. *Principles of forecasting*, Springer, Boston, MA. pp.125–144.

Sagi, D., 2015. The concept of power in decision making process: a cross cultural perspective. *International Journal of Management and Social Science Research Review*, 1(9).

Sasirekha, K. and Baby, P., 2013. Agglomerative hierarchical clustering algorithm-a. *International Journal of Scientific and Research Publications*, 83, p.83.

Savage, L.J., 1972 *The Foundations of Statistics.* New York: Dover.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. *Statistica sinica*, pp.639-650.

Shao, J., 1993. Linear model selection by cross-validation. *Journal of the American statistical Association.*, 88:(422): pp.486–494.

Shanteau, J., 1992. How much information does an expert use? Is it relevant? *Acta psychologica*, 81(1): pp.75-86.

Shanteau, J., 1995. *Expert Judgment and Financial Decision Making. Risky Business: Risk Behavior and Risk Management.* B. Green: Stockholm, Stockholm University.

Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R. and Carlsson, C., 2002. Past, present, and future of decision support technology *Decision support systems*, 33(2): pp.111-126.

Skjong, R. and Wentworth, B.H., 2001. *Expert judgement and Risk Perception. Proceedings of the Eleventh (2001) International Offshore and Polar Engineering Conference*, Stavanger, Norway: The International Society of Offshore and Polar Engineers.

Skulmoski, G.J., Hartman, F.T. and Krahn, J., 2007. The Delphi method for graduate research. *Journal of Information Technology Education: Research.*, 6(1): pp.1–21.

Skronska, A., 2014. Meta-Analysis of Expert Judgement - The Bayesian Hierarchical model *(unpublished).*

Slovic, P. and Weber, E.U., 2002. Perception of risk posed by extreme events *Regulation of Toxic Substances and Hazardous Waste (2nd edition)(Applegate, Gabba, Laitos, and Sachs, Editors)* Foundation Press.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., and Burgman, M., 2010. Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, 30: pp.512–523.

Steurer, J., 2011. The Delphi method: an efficient procedure to generate knowledge. *Skeletal radiology*, 40(8): pp.959–961.

Surowiecki, J., 2005. *The wisdom of crowds* Anchor.

Smith, J.Q., 2010. *Bayesian decision analysis: principles and practice*. Cambridge University Press.

Stiber, N.A., Small, M.J. and Pantazidou, M., 2004. Site-Specific Updating and Aggregation of Bayesian Belief Network Models for Multiple Experts *Risk Analysis*, 24(6): pp.1529–1538.

Taylor, R.E. and Judd, L.L., 1989 Delphi method applied to tourism. *Delphi method applied to tourism.* pp.95–98.

Tetlock, P.E. and Gardner, D., 2016. *Superforecasting: The art and science of prediction*. Random House.

Tyshenko, M.G., ElSaadany, S., Oraby, T., Darshan, S., Aspinall, W., Cooke, R., Catford, A., and Krewski, D., 2011. Expert elicitation for the judgment of prion disease risk uncertainties. *Journal of Toxicology and Environmental Health, Part A*, 74(2-4): pp.261–285.

Tyshenko, M.G., ElSaadany, S., Oraby, T., Darshan, S., Catford, A., Aspinall, W., Cooke, R., and Krewski, D., 2012. Expert judgement and re-elicitation for prion disease risk uncertainties. *International Journal of Risk Assessment and Management*, 16(1-3): pp.48–77.

Van der Fels-Klerx, H.J., Cooke, R.M., Nauta, M.N., Goossens, L.H. and Havelaar, A.H., 2005. A structured expert judgment study for a model of Campylobacter transmission during broiler-chicken processing. *Risk Analysis: An International Journal*, 25(1): pp.109-124.

Wagstaff, K., Cardie, C., Rogers, S. and Schrödl, S., 2001. Constrained k-means clustering with background knowledge *Icml*, 1: pp.577–584.

Walker, K.D., Catalano, P., Hammitt, J.K. and Evans, J.S., 2003. Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene. *Journal of Exposure Science & Environmental Epidemiology*, 13(1): pp.1–16.

Walshe, K. and Rundall, T.G., 2001. Evidence-based Management: From Theory to Practice in Health Care. *Milbank Quarterly*, 79(3): pp.429–457.

Walton, D., 2010. *Appeal to expert opinion: Arguments from authority.*, Penn State Press.

Wang, X., Gao, Z. and Guo, H., 2012. Delphi method for estimating uncertainty distributions. *Information: An International Interdisciplinary Journal*, 15(2): pp.449–460.

Williams, C.J., Wilson, K.J. and Wilson, N., 2020. A Comparison of Prior Elicitation Aggregation using the Classical Method and SHELF. *arXiv preprint arXiv:2001.11365*.

Wilks, D.S., 2010. Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 136(653): pp.2109–2118.

Wilson, K.J., 2017. An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, 33(1): pp.325–336.

Wilson, K.J., 2018. Specification of informative prior distributions for multinomial models using vine copulas. *Bayesian Analysis*, 13(3): pp.749–766.

Wilson, K.J. and Farrow, M., 2018. Combining judgements from correlated experts. *Elicitation*, Springer, Cham, pp.211–240.

Wiper, M.P. and French, S., 1995. Combining experts' opinions using a normal-wishart model. *Journal of Forecasting*, 14(1): pp.25–34.

Wiper, M.P., French, S. and Cooke, R., 1994. Hypothesis-based calibration scores *The Statistician*, pp.231–236.

Winkler, R.L., Grushka-Cockayne, Y., Lichtendahl Jr, K.C. and Jose, V.R.R.,2019. Probability Forecasts and Their Combination: A Research Perspective. *Decision Analysis*, 16(4): pp.239–260.

Winkler, R.L., 1981. Combining probability distributions from dependent information sources. *Management Science*, 27(4): pp.479–488.

Winkler, R.L., Grushka-Cockayne, Y., Lichtendahl, K.C. and Jose, V.R.R., 2018. *Averaging Probability Forecasts: Back to the Future.* Harvard Business School.

Wittmann, M.E., Cooke, R.M., Rothlisberger, J.D. and Lodge, D.M., 2014. Using Structured Expert Judgment to Assess Invasive Species Prevention: Asian Carp and the Mississippi- Great Lakes Hydrologic Connection. *Environmental science & technology*, 48(4): pp.2150-2156.

Wittmann, M.E., Cooke, R.M., Rothlisberger, J.D., Rutherford, E.S., Zhang, H., Mason, D.M., and Lodge, D.M., 2015. Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. *Conservation Biology*, 29(1): pp.187-197.

Wright, G., Rowe, G., Bolger, F. and Gammack, J., 1994. Coherence, calibration, and expertise in judgmental probability forecasting. *Organizational Behavior and Human Decision Processes*, 57(1): pp.1–25.

Zhang, H., Rutherford, E. S., Mason, D.M., Breck, J.T., Wittmann, M.E., Cooke, R.M., Lodge, D.M., Rothlisberger, J.D., Zhu, X. and Johnson, T.B., 2016. Forecasting the impacts of silver and bighead carp on the Lake Erie food web. *Transactions of the American Fisheries Society*, 145(1): pp.136-162.

Chapter 11

Glossary

11.1 Abbreviations

BDM - A Bayesian decision maker.

BUGS - Bayesian inference using Gibbs sampling, a program for analysing Bayesian hierarchical models using Markov chain Monte Carlo.

CEBRA - Centre of Excellence for Biosecurity Risk Analysis, at the University of Melbourne.

CHIPRA - US Children's Health Insurance Program Reauthorization Act of 2009.

CWD - Chronic wasting disease.

DM - A decision maker.

DPMM - Dirichlet process mixture model.

EFSA - European food standards agency.

EWDM - Equal weighted decision maker (based on a linear opinion pool).

EXCALIBUR - EXpert CALIBration, the tool for deploying Cooke's Classical model.

IDEA - Investigate, Discuss, Estimate and Aggregate. An expert judgement elicitation and aggregation protocol.

JAGS - Just Another Gibbs Sampler, a program for analysing Bayesian hierarchical models using Markov chain Monte Carlo, integrated into the R coding language.

KM - Knowledge management.

MCMC - Markov chain Monte Carlo.

OOS - Out of sample validation.

PCA - Principal component analysis.

PWDM - Performance weighted decision maker (with weightings based on Cooke's Classical model).

ROAT - Remove one at a time validation.

SEJ - Structured expert judgement.

SHELF - Sheffield Elicitation Framework, a behavioural expert elicitation protocol.

11.2 Mathematical Symbols

11.2.1 Study meta-data

e - An individual expert.

$\mathbf{E}$ - A group of experts.

$|\mathbf{E}|$ - The number of experts within the group.

x_X - The outcomes of a target variable $X \in \mathbf{X}$.

X - A random variable denoting a single target variable.

$\mathbf{X}$ - The set of target variables within an SEJ study.

$|\mathbf{X}|$ - The number of target variables considered within an SEJ study.

y_Y - The true realisation, with value known *a priori* of a study, of a seed variable $Y \in \mathbf{Y}$.

Y - A random variable denoting a single seed variable.

$\mathbf{Y}$ - The set of seed variables within an SEJ study.

$|\mathbf{Y}|$ - The number of seed variables considered within an SEJ study.

P_L - The percentile that the lower value is going to be elicited against.

P_M - The percentile that the middle value is going to be elicited against.

P_U - The percentile that the upper value is going to be elicited against

11.2.2 Elicited variables

L_e - The lower quantile elicited from expert $e \in \mathbf{E}$.

L_{eh} - The lower quantile elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$.

L_{Xeh}/L_{Yeh} - The lower quantile elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$ for either target variable $X \in \mathbf{X}$ or seed variable $Y \in \mathbf{Y}$.

M_e - The middle quantile elicited from expert $e \in \mathbf{E}$.

M_{eh} - The middle quantile elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$.

M_{Xeh}/M_{Yeh} - The middle quantile elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$ for either target variable $X \in \mathbf{X}$ or seed variable $Y \in \mathbf{Y}$.

U_e - The upper quantile elicited from expert $e \in \mathbf{E}$.

U_{eh} - The upper quantile elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$.

U_{Xeh}/U_{Yeh} - The upper quantile elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$ for either target variable $X \in \mathbf{X}$ or seed variable $Y \in \mathbf{Y}$.

11.2.3 Distribution fitting

g_e - the probability density function of the distribution that will be fit to the elicited quantiles for expert $e \in \mathbf{E}$.

G_e - the cumulative density function of the distribution that will be fit to the elicited quantiles for expert $e \in \mathbf{E}$.

11.2.4 Homogeneity group calculation

C - The set of clusters within the seed variable space.

eh - The pair, expert e and the homogeneity group h to which they belong.

h - An individual homogeneity group.

H - The set of homogeneity groups.

$|\mathbf{H}|$ - The number of homogeneity groups calculated for a study.

rM_{Yeh} - The middle quantile for expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$ for seed variable $Y \in \mathbf{Y}$ rescaled onto the unit interval.

$\mathbf{Y}_e$ - The $|\mathbf{Y}|$ dimensional tuple of all of the mid quantile estimates for expert e within a study, rescaled to the interval [0,1].

θ_h - The parameters within a Dirichlet process mixture model for the mixture within which a given expert e sits.

θ_k - The set of parameters for all potential mixtures in the DPMM.

θ_0 - The hyperparameter of the mixture parameters in a DPMM.

ν - The mixing weights generated during a stick breaking process dependent on tuning parameter o.

ν_k - The kth mixing weight in a DPMM.

o - The mixing weight tuning parameter in a DPMM.

11.2.5 Calibration variables

L^*_{Xeh}/L^*_{Yeh} - The unbiased (post calibration) lower quantile estimates from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$ for either target variable $X \in \mathbf{X}$ or seed variable $Y \in \mathbf{Y}$.

M^*_{Xeh}/M^*_{Yeh} - The unbiased (post calibration) middle quantile estimates from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$ for either target variable $X \in \mathbf{X}$ or seed variable $Y \in \mathbf{Y}$.

U^*_{Xeh}/U^*_{Yeh} - The unbiased (post calibration) upper quantile estimates from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$ for either target variable $X \in \mathbf{X}$ or seed variable $Y \in \mathbf{Y}$.

α_{le} - The inflation factor for expert $e \in E$ applied to calculate the lower unbiased estimators.

α_{ue} - The inflation factor for expert $e \in E$ applied to calculate the upper unbiased estimators.

$A_{lh}/B_{lh}/A_{uh}/B_{uh}$ - Hierarchical parameters of the calibration model used to infer inter-expert calibration dependence.

a_l/b_l - Hyperparameters of the calibration model.

11.2.6 Aggregation variables

γ_{Xeh} - Generic parameters of the fitted distribution for target variable $X \in \mathbf{X}$ elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$.

γ^*_{Xeh} - Unbiased (post calibration) generic parameters of the fitted distribution for target variable $X \in \mathbf{X}$ elicited from expert $e \in \mathbf{E}$ who belongs to homogeneity group $h \in \mathbf{H}$.

γ_{Xh} - Generic parameters for homogeneity group $h \in \mathbf{H}$ for target variable $X \in \mathbf{X}$.

γ_X - Generic parameters at the global level for target variable $X \in \mathbf{X}$. These represent the output of the model, the consensus of the group, and the parameters of the (Supra-)Bayesian decision maker.

π_{DM_X} - The decision maker's prior on the target variable $X \in \mathbf{X}$.

ρ_{Xh} - Dispersion parameter for the homogeneity group $h \in \mathbf{H}$ for target variable $X \in \mathbf{X}$.

ρ_X - Global dispersion parameter for target variable $X \in \mathbf{X}$.

11.2.7 Split normal parameterisation

M_{Xh} - Median of the homogeneity group for target variable $X \in \mathbf{X}$.

M_X - Global median for target variable $X \in \mathbf{X}$. Represents the consensus median and the final output for the (Supra-)Bayesian decision maker.

M_{DM_X} - Decision maker prior on the median for the target variable $X \in \mathbf{X}$.

ξ_{Xlh}/ξ_{Xuh} - Parameter of the gamma distribution governing the ratio between the precision of the homogeneity group and the unbiased precision of an expert within that group.

ξ_{Xl}/ξ_{Xu} - Parameter of the gamma distribution governing the ratio between the global precision, representing that of the (Supra-)Bayesian decision maker and the precision of each homogeneity group.

ρ_{Xh} - Dispersion parameter of the normal distribution representing the uncertainty on the median of the homogeneity group $h \in \mathbf{H}$ for target variable $X \in \mathbf{X}$.

ρ_{Xh} - Dispersion parameter of the normal distribution representing the uncertainty on the global median for target variable $X \in \mathbf{X}$.

ρ_{X0} - Dispersion hyper-parameter representing the prior belief on the global median for target variable $X \in \mathbf{X}$.

ξ_{Xl}/ξ_{Xu} - Parameter of the gamma distribution governing the ratio between the global precision, representing that of the (Supra-)Bayesian decision maker and the precision of each homogeneity group.

$\sigma^*_{Xueh}/\sigma^*_{Xleh}$ - Unbiased standard deviation on the upper/lower portion of a split normal parameterisation of the beliefs of expert $e \in \mathbf{E}$ about target variable $X \in X$, where expert e belongs to homogoneity group $h \in \mathbf{H}$.

$\tau^*_{Xueh}/\tau^*_{Xleh}$ - Unbiased precision on the upper/lower portion of a split normal parameterisation of the beliefs of expert $e \in \mathbf{E}$ about target variable $X \in \mathbf{X}$, where expert e belongs to homogoneity group $h \in \mathbf{H}$.

τ_{Xuh}/τ_{Xlh} - Precision of the upper/lower portion of a split normal parameterisation of homogoneity group $h \in \mathbf{H}$ for target variable $X \in \mathbf{X}$.

τ_{Xl}/τ_{Xu} - Global precision variable for the upper/lower portion of a split normal parameterisation of target variable $X \in \mathbf{X}$.

τ_{Xl0}/τ_{Xu0} - Global hyperparameters for the precision variable for the upper/lower portion of a split normal parameterisation of target variable $X \in \mathbf{X}$.

Appendix A

Appendix

A.1 Additional Arkansas study analysis and figures

A.1.1 Dendogram of expert homogeneity groups

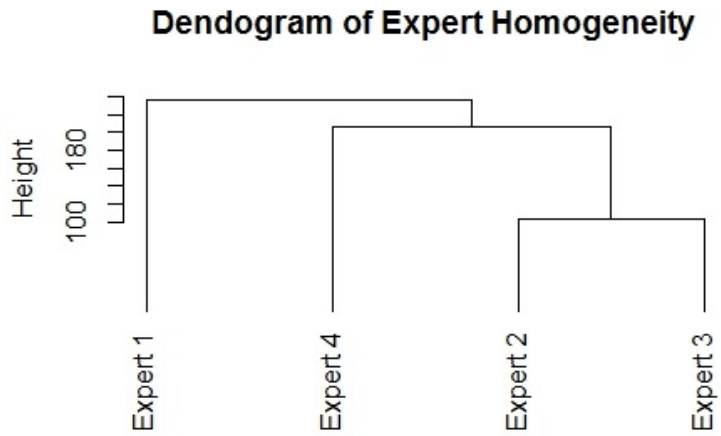


Figure A.1: Hierarchical clustering dendogram for the identification of expert homogeneity groups within the Arkansas study. Expert 2 and 3 form a single homogeneity group.

A.1.2 Distributions for all target variables

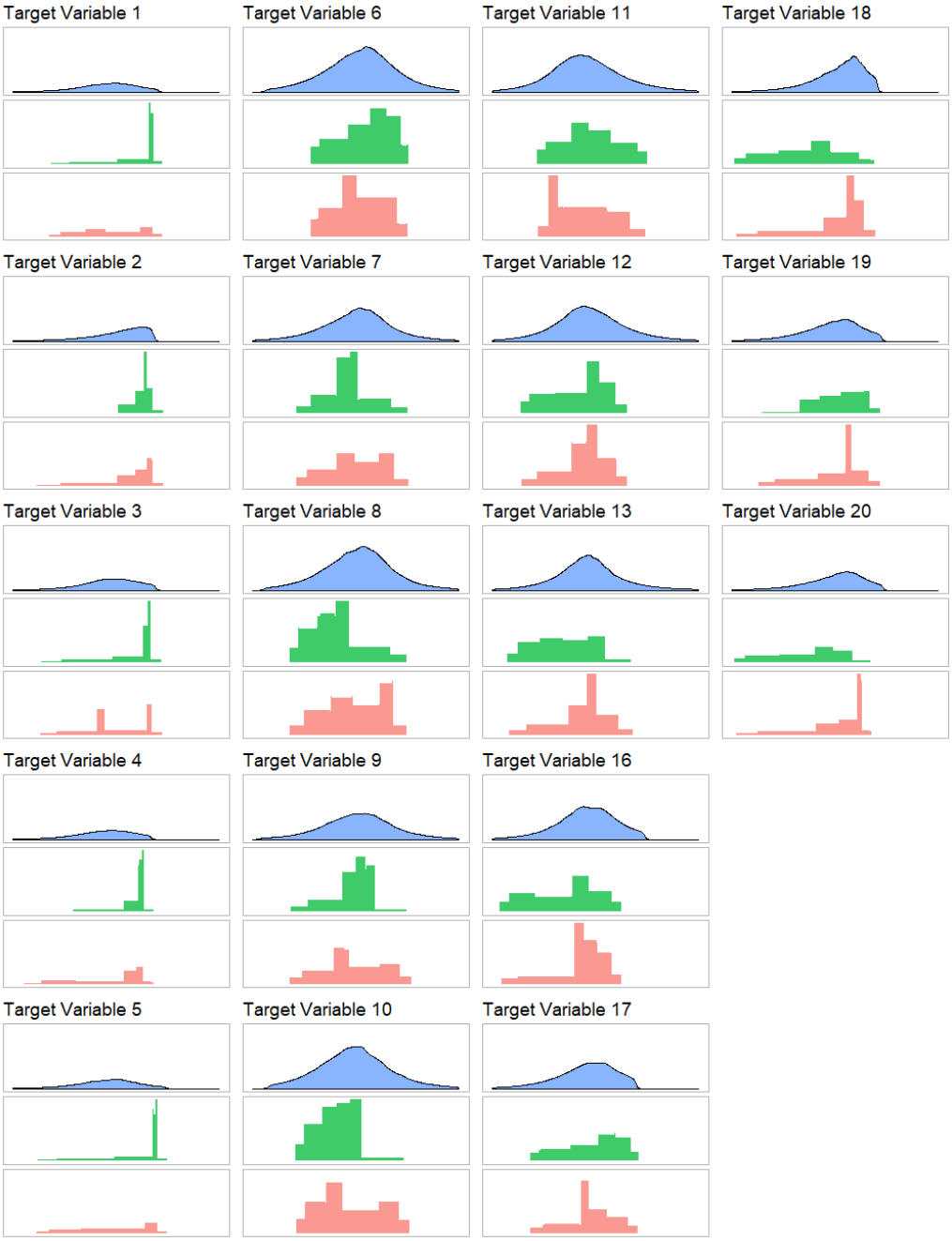


Figure A.2: Comparison of final distributions across all target variables within the Arkansas study. The Bayesian model (blue) demonstrates a larger support, aligned to the overconfidence demonstrated by experts in the seed variables.

A.1.3 Cumulative density functions for different parameterisations of the calibration and aggregation model

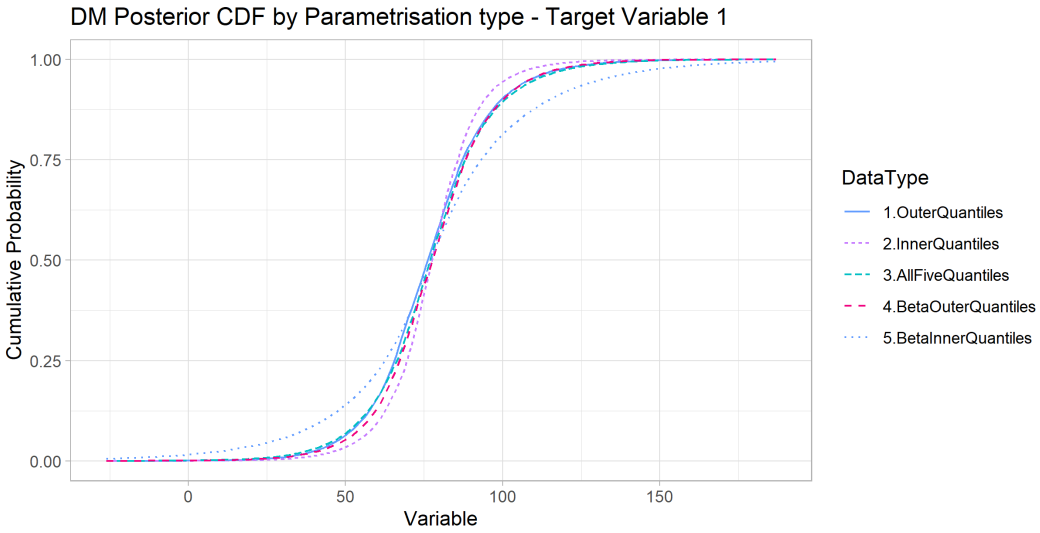


Figure A.3: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 1 in the Arkansas study.

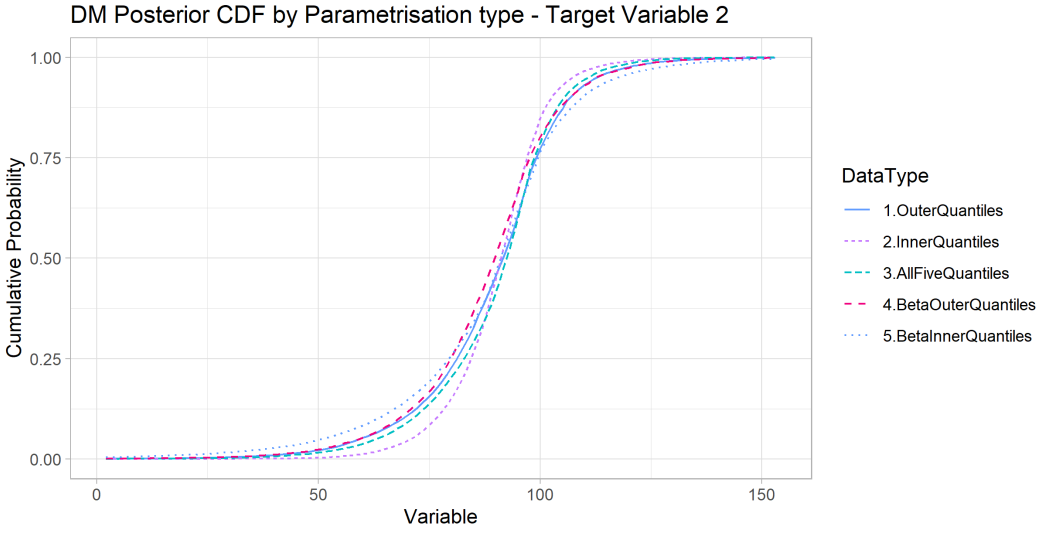


Figure A.4: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 2 in the Arkansas study.

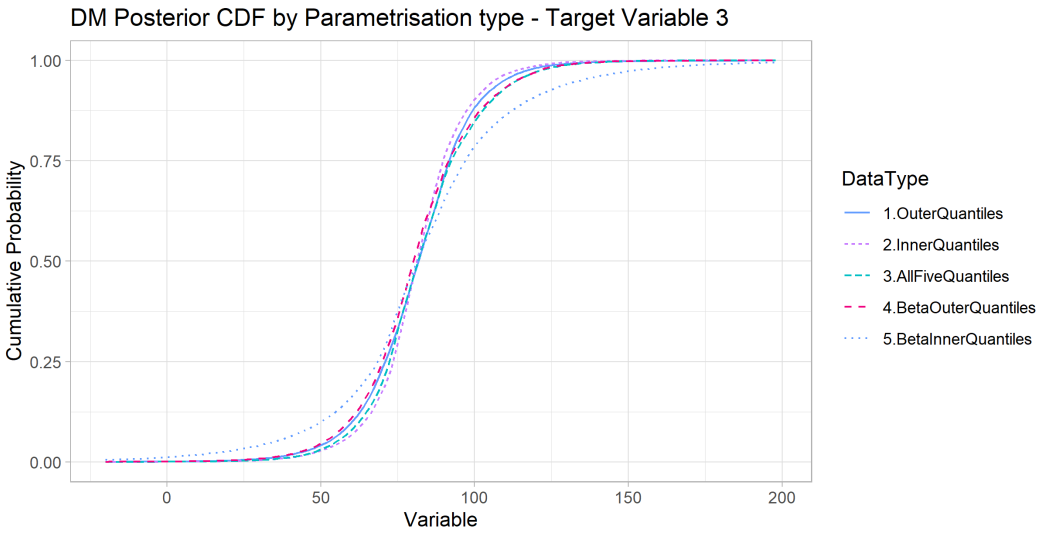


Figure A.5: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 3 in the Arkansas study.

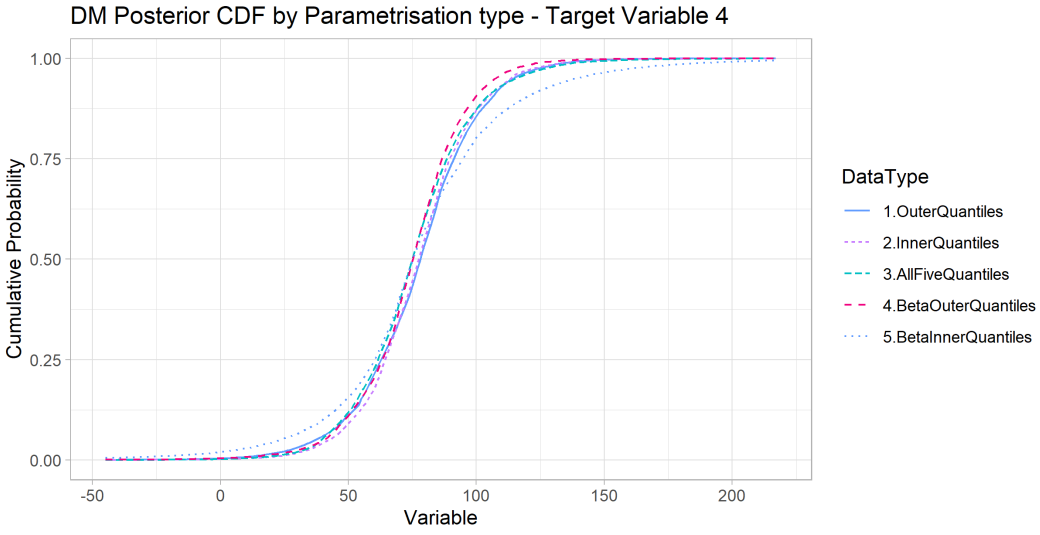


Figure A.6: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 4 in the Arkansas study.

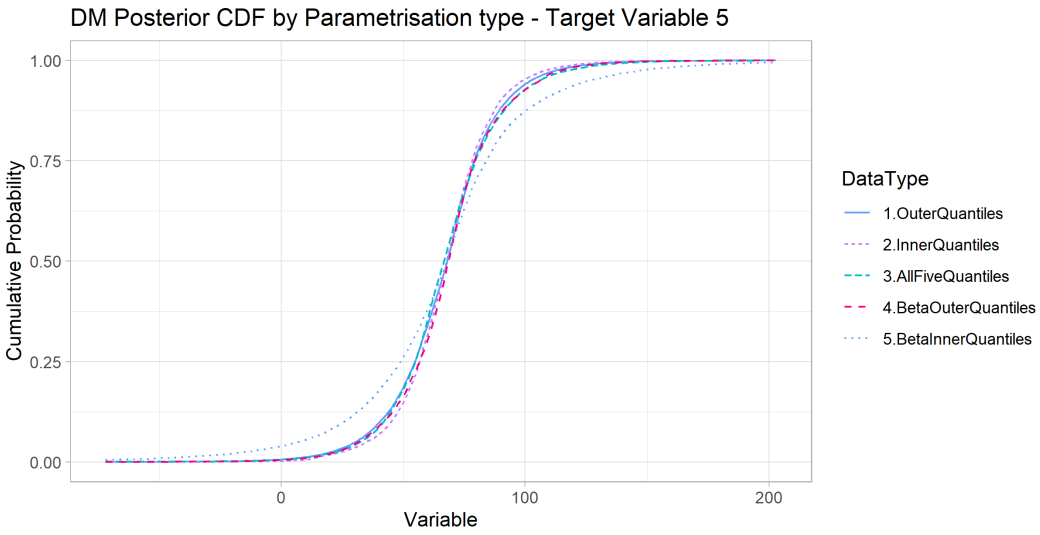


Figure A.7: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 5 in the Arkansas study.

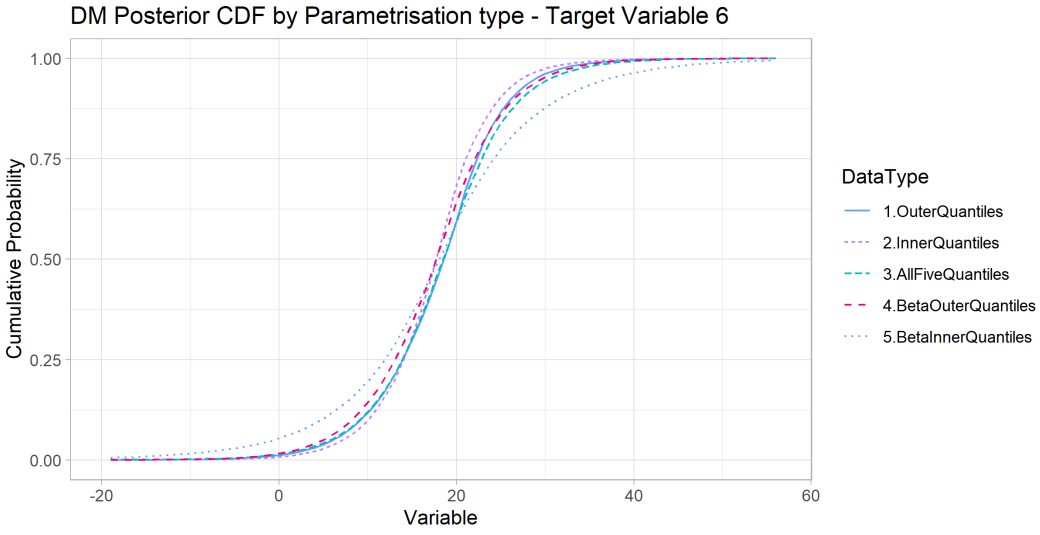


Figure A.8: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 6 in the Arkansas study.

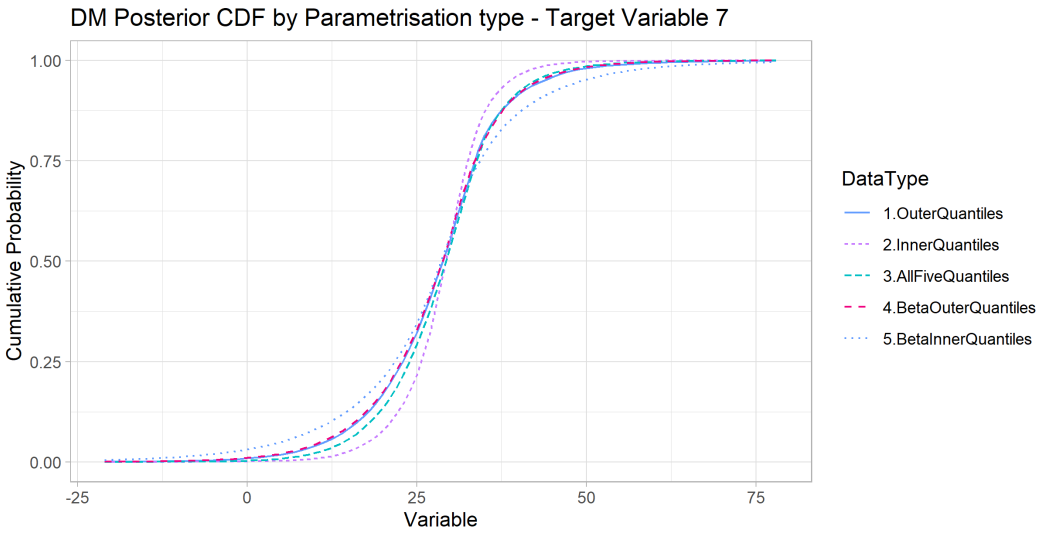


Figure A.9: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 7 in the Arkansas study.

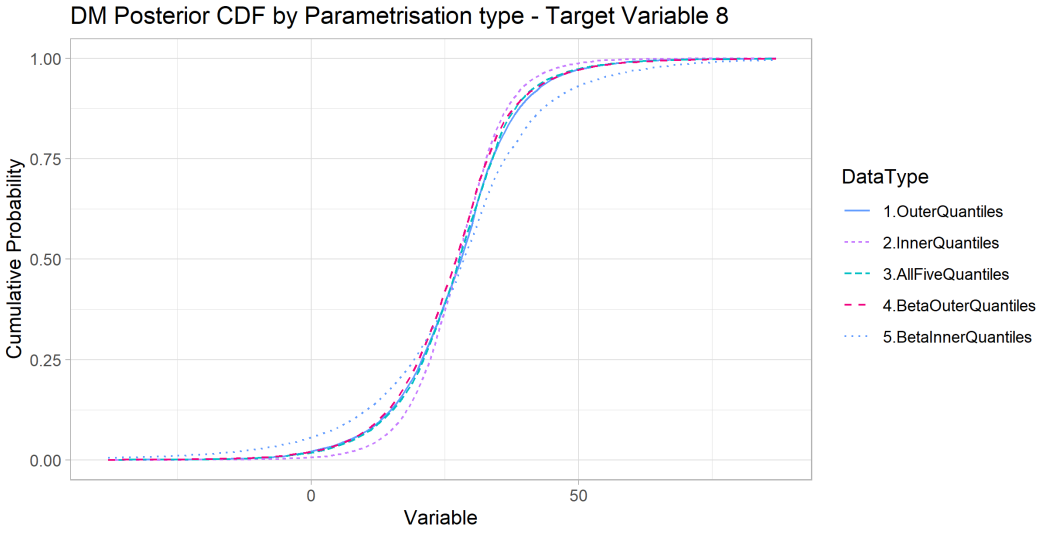


Figure A.10: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 8 in the Arkansas study.

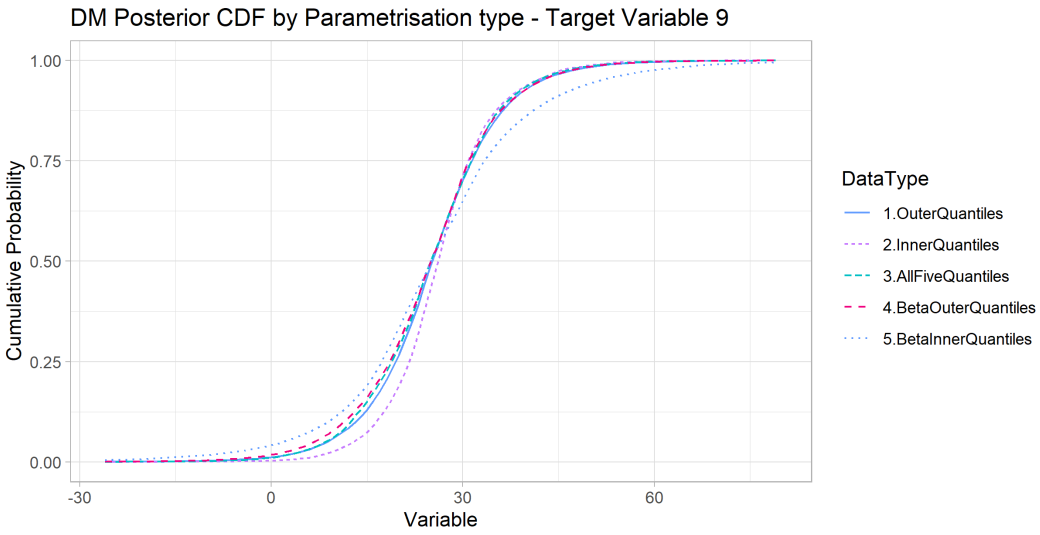


Figure A.11: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 9 in the Arkansas study.

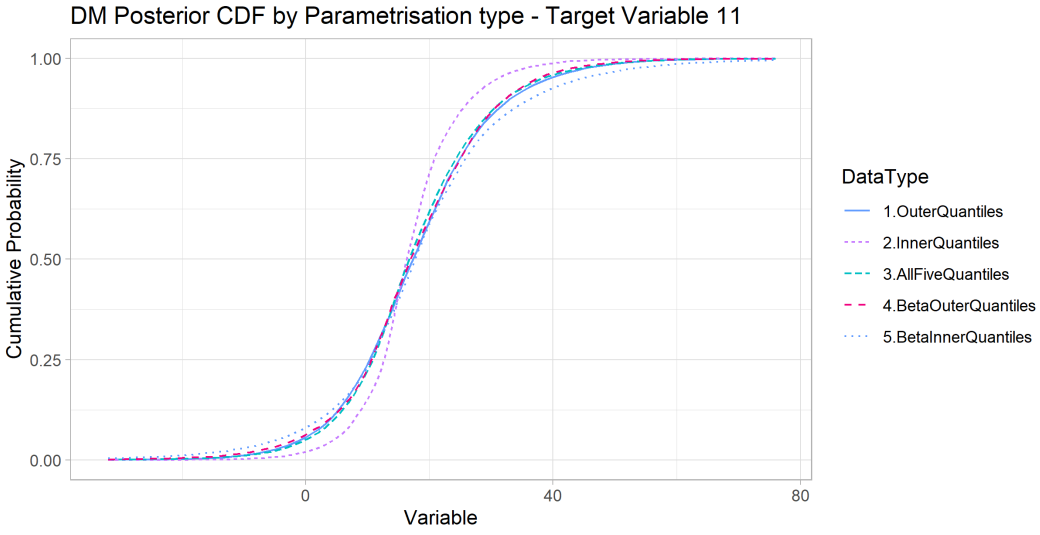


Figure A.12: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 11 in the Arkansas study.

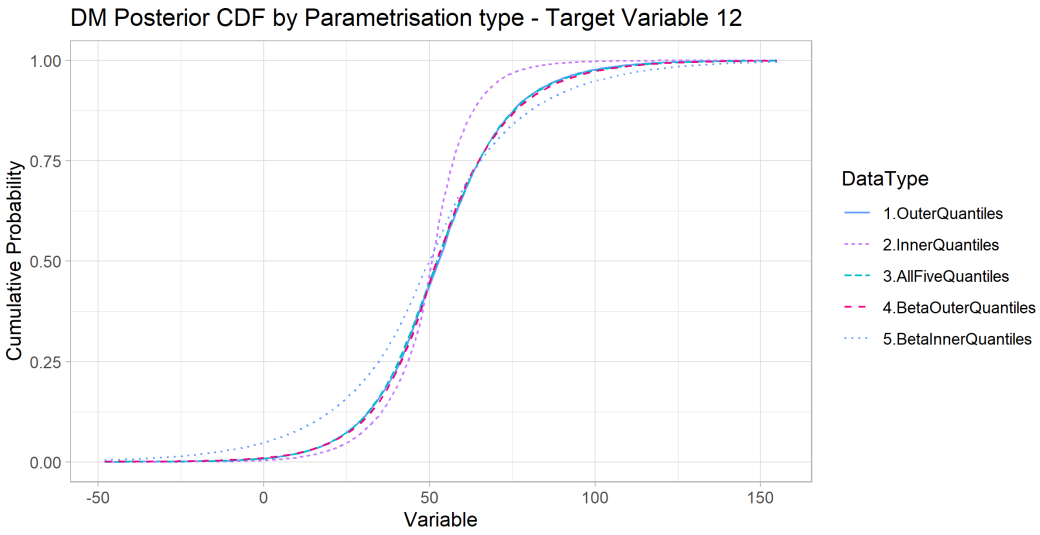


Figure A.13: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 12 in the Arkansas study.

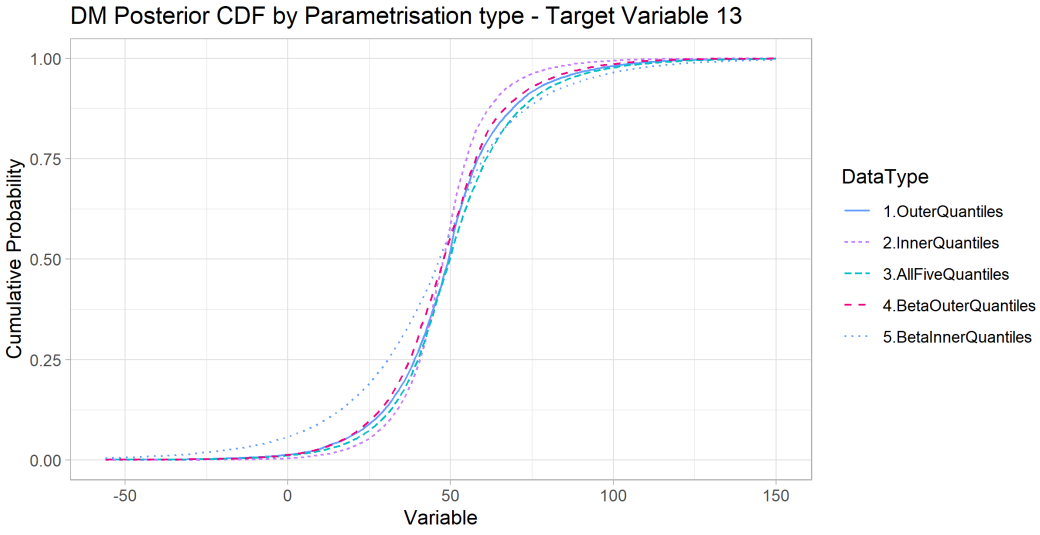


Figure A.14: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 13 in the Arkansas study.

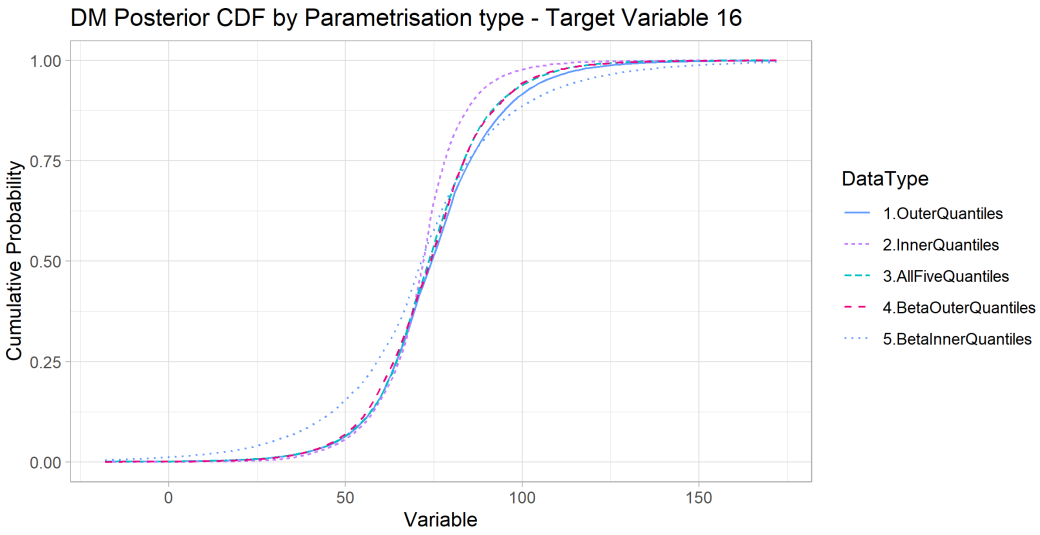


Figure A.15: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 16 in the Arkansas study.

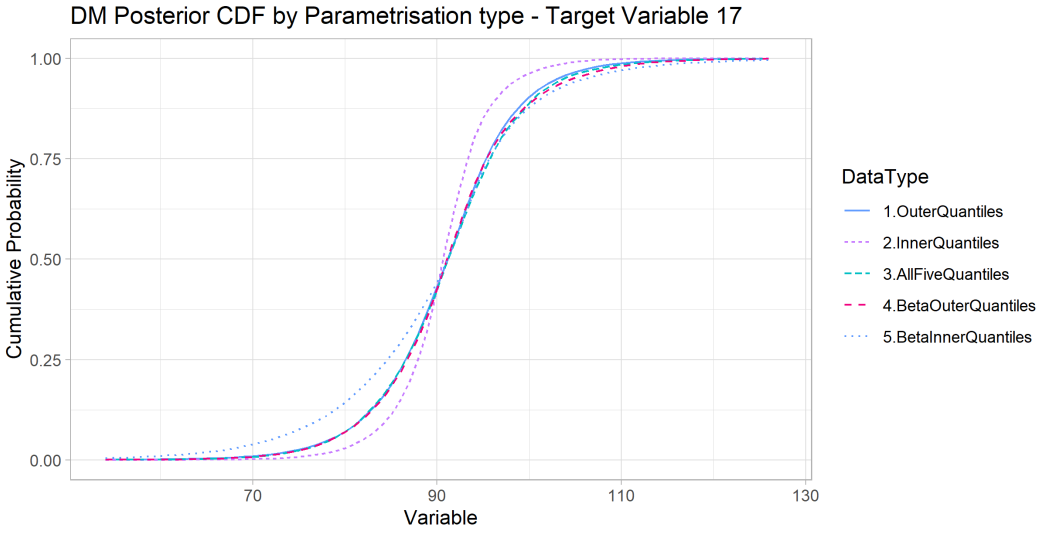


Figure A.16: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 17 in the Arkansas study.

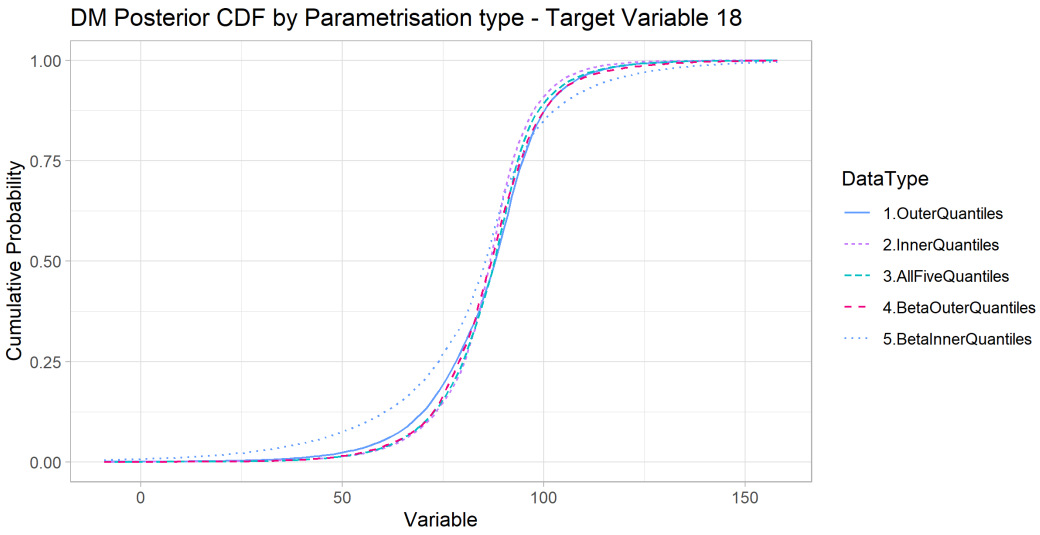


Figure A.17: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 18 in the Arkansas study.

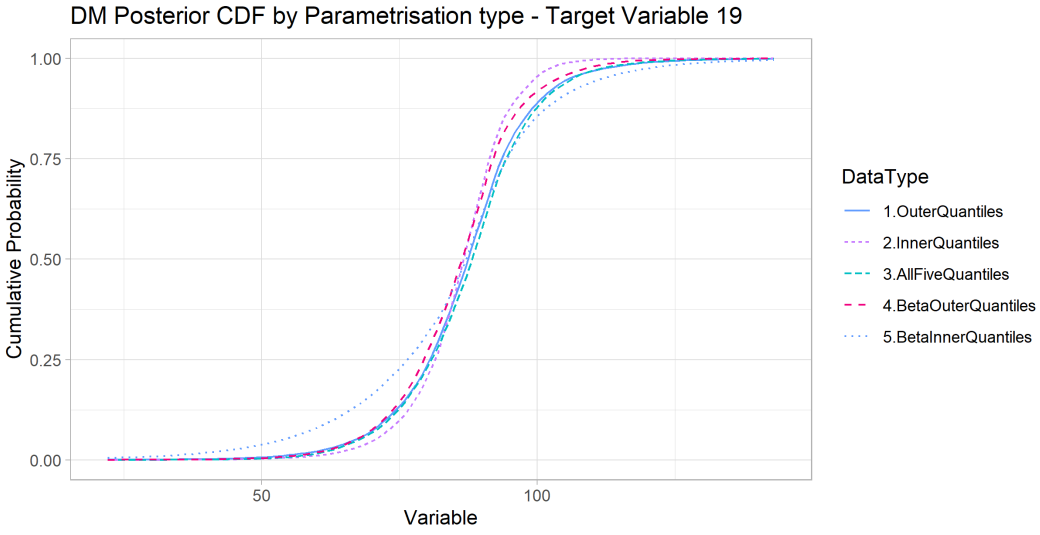


Figure A.18: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 19 in the Arkansas study.

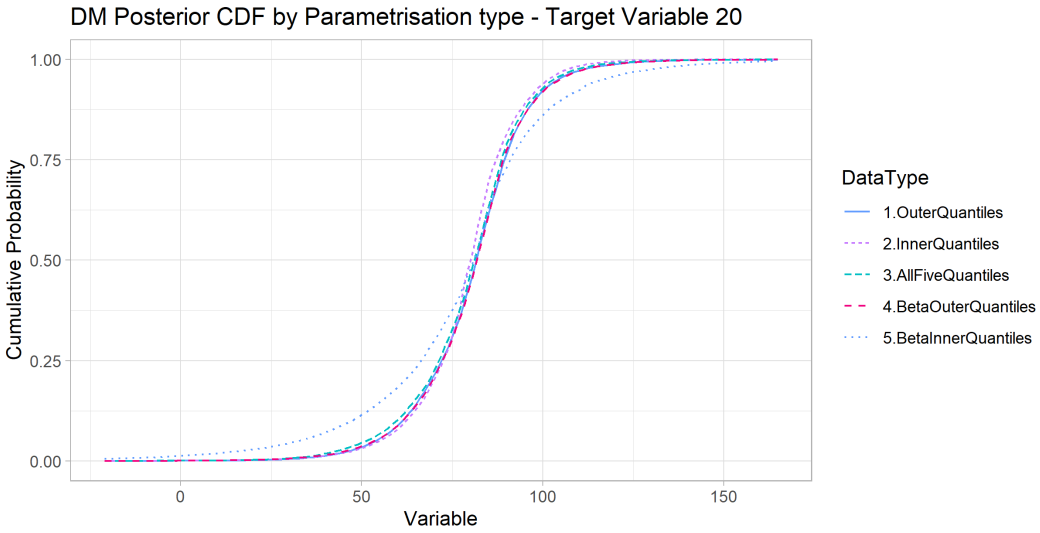


Figure A.19: Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 20 in the Arkansas study.

A.2 Additional CWD study analysis

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Variable 1	0.0	9	90.5	0.0	1.7	19.8	-12.3	8.3	47.3
Variable 2	0.2	29.7	95.8	0.0	5	58.7	-37.9	30.8	114.8
Variable 3	0.6	14.3	454.4	0.5	4.8	40	-13.4	11	457
Variable 4	0.0	9.1	238.7	0.0	15.6	193.4	-8	7.1	345.6
Variable 5	0.1	65.3	100	0.0	13.3	92.1	-24.7	62.9	122.4
Variable 6	12.1	66.4	95.1	13.9	69.5	89.7	-18.9	68.6	121.8
Variable 7	11.7	56.5	91.8	12	62.1	84.8	-17	59.2	119.1
Variable 8	7.8	49.7	89.6	10	46.6	89	-36.6	52.8	123.4
Variable 9	0.0	8.7	83	0.0	3.4	19.8	-14.6	8.4	69.7
Variable 10	41.2	98.7	100	90.1	98.4	100	75.3	99.8	102.7
Variable 11	0.0	1.3	461.7	0.1	0.9	2	-2	1.1	38.2
Variable 12	0.0	14.8	87.8	0.0	0.7	19.7	-24.6	12.8	87.5
Variable 13	2.1	39.5	95.3	1.9	39.6	97.4	-46	44.7	120.2

Table A.1: Comparison of DM quantiles for different modelling approaches to the CWD Study.

A.3 Additional effusive eruption study analysis

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
10x SO2 Probability	0.02	2.76	32.19	0.24	5.03	19.2	-0.41	0.99	39.44
Column Height	5.24	13.61	22.68	6.52	12.87	22.22	-2.96	13.47	31.67
Avg Plume Height	0.6	3.79	11.92	0.58	4.05	11.26	-2.95	4.22	13.06
Max Plume Height	0.53	3.93	13.2	0.85	3.87	15.78	-3.63	3.98	16.74
Max% SO2 Emissions	53.04	86.18	99.74	50	75.63	97.45	32.63	86.15	116.77
Min% SO2 Emissions	25.09	70.89	89.96	40.81	68.2	84.61	10.27	71.21	118.77
Max No. Fissures	3.88	26.15	472.4	6.08	18.45	98.43	-25.05	28.77	174.32
Min No. Fissures	0.12	2.95	13.6	1.19	6.6	16.83	-4.31	3.55	20.27
Duration explosive phase	0.14	2.88	15.84	0.34	5.64	23.79	-4.3	2.86	21.44
Gap between outbursts	0.12	7.17	183.4	0.51	5.71	29.87	-2.17	4.37	163.99

Table A.2: Target variable predicted quantiles for the effusive eruption study across models.

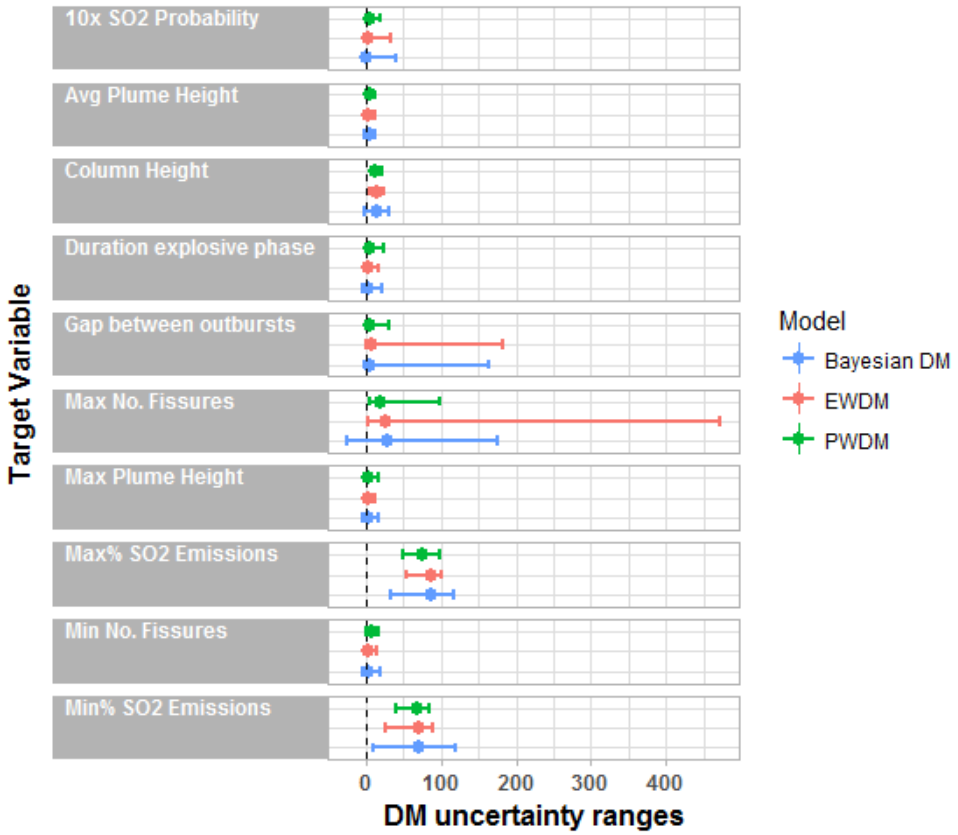


Figure A.20: Replication of Fig.7.13 including the EWDM. The Bayesian model displays posterior uncertainty ranges consistently broader than the PWDM, however, displays uncertainty bounds both broader and narrower than the EWDM.