ABSTRACT

Title of Dissertation: DATA-INFORMED CALIBRATION AND AGGREGATION OF EXPERT JUDGMENT IN A BAYESIAN FRAMEWORK

Calvin Homayoon Shirazi, Doctor of Philosophy, 2009

Dissertation Directed by: Professor Ali Mosleh, Reliability Engineering Program, Department of Mechanical Engineering

Historically, decision-makers have used expert opinion to supplement lack of data. Expert opinion, however, is applied with much caution. This is because judgment is subjective and contains estimation error with some degree of uncertainty. The purpose of this study is to quantify the uncertainty surrounding the unknown of interest, given an expert opinion, in order to reduce the error of the estimate. This task is carried out by data-informed calibration and aggregation of expert opinion in a Bayesian framework. Additionally, this study evaluates the impact of the number of experts on the accuracy of aggregated estimate. The objective is to determine the correlation between the number of experts and the accuracy of the combined estimate in order to recommend an expert panel size.

DATA-INFORMED CALIBRATION AND AGGREGATION OF EXPERT OPINION IN A BAYESIAN FRAMEWORK

By

Calvin Homayoon Shirazi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Ali Mosleh, Chair
Professor Mohammad Modarres
Professor Jeffrey Hermann
Professor Michael Cukier
Professor Gregory Baecher (Dean Representative)

Dedication

To my parents…

To my wife and son…

Who waited patiently for me to reach higher!

Acknowledgement

I express my sincere appreciation to Professor Ali Mosleh for his guidance and support in completion of this dissertation. I also extend my gratitude to all who contributed to this study.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Historically, decision-makers have utilized expert judgment to supplement insufficient data or carry out a task proficiently. A major source of information in estimating parameters of risk and reliability models is expert knowledge. Cases involving new design, very rare events, and proceedings that are beyond our direct experience, call for the use of expert opinion as a surrogate source of information. Experts can extensively influence key decisions in the political, financial, legal, and social arenas.

Although, expert estimate is treated as scientific data, it is applied with much caution. This is because an opinion is not a fact, verified by an experiment; it is a person's assessment or judgment about a specific subject.

According to the RAND Corporation, opinion is a blend of knowledge and speculation (Forrester, 2005). In the Oxford English dictionary, speculation denotes assumptions with minimum or no supporting evidence and knowledge is defined as the theoretical or practical understanding of a subject. Considering these definitions, uncertainty in judgment simply translates into a range of possible outcomes, given the current state of expert knowledge. Though, it can be argued that other types of data can also be uncertain, the human psyche introduces a unique category of complications by itself. This means that there are degrees of inherent variation in the expert judgment. Problems in expert judgment studies begin with the identification of attributes by which one can qualify an individual as an ‘expert’. There is no established intra- or interdisciplinary taxonomy based on the relation between the expert qualifications and the accuracy of judgment.

Expert selection is often founded on uncorroborated ideas or subjective criteria, such as sufficient knowledge or experience in a discipline. Of course, this kind of general approach is subject to interpretation, which in turn, results in inconsistencies across the board.

Additionally, the majority of the developed models used for the assessment of expert accuracy are based on historical performance of the individual expert. Therefore, decision makers need to be aware of the prior performance of the expert. When such information is not available, analysts are puzzled about the quality of opinion or the degree of confidence to place on the judgment. In practice, decision makers remain uncertain about the proper procedure to evaluate the expert judgment accuracy.

In contrast to many studies revealing deficiencies in the expert judgment, this research study assesses how well experts are able to make predictions. This task is carried out by data-informed calibration of experts in a Bayesian framework.

Bayesian method begins with the analyst prior belief of an unknown. Once the expert estimate is obtained, this prior belief is renewed using Bayes' method to establish a posterior, describing the analyst updated knowledge of unknown of interest.

The main problem in applying Bayesian method is the complications associated with the development of a proper likelihood function. This distribution is a probabilistic model for data and must capture the interrelationships among estimates and the unknown. The first part of this research is dedicated to development and validation of proper likelihood functions.

In the beginning, a comprehensive database of observed relative errors of experts in various fields is assembled to determine the distribution of errors. Realizing the norm and the spread of errors, a totally unique generic likelihood is developed, independent of discipline, capable of improving the expert estimate. The generic likelihood along with case-specific likelihood distributions developed by Droguette and Mosleh (2003) is then tested using empirical data to reveal their ability in reducing future error of prediction. To the author's best knowledge, there has not been any study conducted, in comparison with this comprehensive research, employing such sizeable empirical data from various fields.

This study also considers the impact of the number of experts on the accuracy of aggregated estimate in a Bayesian framework. Because expert opinion is considered uncertain, it seems logical to consult multiple experts in an attempt to have a more inclusive database or at least gather more information. Speculations about the positive correlation between the prediction accuracy and the number of experts, assert that the more experts are elicited, the higher the accuracy of the combined estimates achieved. Question still remains whether empirical data actually support this assertion, and if so, to what extend this link has an impact on practical cases. The second part of this study answers this question.

Collected expert judgments are combined in a Bayesian framework using likelihood distributions developed in the first part of the research study. Total number of estimates with reduced errors is depicted against corresponding expert panel size. The objective is to determine the correlation between the number of experts and the accuracy of the combined estimate to recommend an expert panel size.

The material presented in this research begins with a comprehensive literature review in eliciting and aggregating of expert opinion in Chapter 2. Chapter 3 characterizes the collected empirical data and explains the rationale of selection of the forecast accuracy measure. An introduction of Bayesian methodology as well as detail mathematical formulation of likelihood functions and posterior distributions are presented in Chapter 4. Chapter 5 is dedicated to the result of calibration studies as well as performance evaluation of the developed generic likelihood function. Chapter 6 presents the result of aggregation analysis via empirical data. In this chapter, Bayesian mathematical aggregation method is evaluated and compared with representative models of axiomatic methods. Additionally, expert panel size is suggested based on the accuracy of aggregated estimate achieved using likelihood functions formulated. The last chapter, Chapter 7, wraps up the topics discussed in this research and summarizes the results of the study for a quick reference.

Chapter 2: Literature Review

In the absence of complete scientific information, decision-makers have to rely on their own intuition or on expert opinion (Baldwin, 1975). Expert judgment represents the expert state of knowledge at the time of response to a question (Keeney and von Winterfeldt, 1991).

According to Booker and Meyer (1996), expert opinion is used in the structuring of technical problems including the determination of relevant information for analysis. It is also used in direct qualitative or quantitative estimates of uncertainties and probabilities.

Lannoy and Procaccia (2001) assert that recourse to expert judgment is required in the completing, validating, interpreting and integrating the existing data as well as predicting the rate of future events and the consequences of a decision. Other situations requiring expert judgment include determining the present state of knowledge in one field and providing the basis for decision-making in the presence of several options.

Issues surrounding the use of expert opinion fall into two broad categories of eliciting and utilizing the opinion, which includes selection of experts, determination of expert panel size, ascertain calibration and aggregation methods, and so on. In line with the scope of this research, a brief review of the literature related to eliciting and aggregating of expert opinion is presented in this chapter.

2.1 Eliciting Expert Opinion

DeGroot (1988) believed that the range of people who can be considered as expert includes “anyone or any system that will give you a prediction” to “someone whose prediction you will simply adopt as your own posterior probability without modification”. Nevertheless, expert judgments should be used with caution, not to replace ‘‘hard science” (Apostolakis, 1990).

The poor quality of expert judgment can be broadly classified as those associated with the individual expert (i.e. attributes, expert definition or distinction), the actual estimates or judgments as well as the elicitation process (formal vs. informal elicitation), aggregation or combining estimates, calibration (performance measures of experts and expertise), and available technical documents (Mosleh and Forrester, 2005). According to Garthwaite et al. (2005), the quality of expert judgments can be controlled by a formal procedure of expert elicitation and documentation.

Application of formal elicitation processes have been recommended by Hora and Iman (1989), Keeney and von Winterfeldt (1991), among many others. The formal elicitation of expert judgment started with the establishment of the RAND Corporation in the United States after Word War II (Cooke, 1991). RAND developed two formal methods for eliciting expert opinion, Delphi and Scenario Analysis through the collaborative project with U.S. Air Force and Douglas Aircraft in 1946 (Ayyub, 2001).

Herman Kahn is regarded as the father of scenario analysis (Cooke, 1991). In this method, scenarios or hypothetical sequences of events are set forth to concentrate on decision-making processes (Kahn and Wiener, 1967).

Helmer and Dalkey were founders of Delphi method (Günaydin, 2009). According to Helmer (1977), Delphi method facilitates level communication among experts and therefore assists the formation of a group judgment. Wissema (1982) states Delphi procedure is developed in order to make discussion between experts possible without permitting a certain social interactive behavior. By 1974, the Delphi study count exceeded 10,000 (Linstone and Turoff, 1975). Delphi method has been widely used to generate forecasts in technology, education, and other fields (Cornish, 1977).

Delphi is based on a structured process for collecting and refining data from a group of experts by means of a series of questionnaires interspersed with controlled opinion feedback (Adler and Ziglio, 1996). Many researchers have suggested that performance feedback is a particularly effective method for improving calibration (e.g., Fischhoff, 1982). Perhaps the most intensive study using performance feedback was conducted by Lichtenstein and Fischhoff (1980). Subjects completed 11 training sessions of 200 general knowledge questions. At the completion of each training session, they were given personalized feedback, including performance measures in calibration and overconfidence. This feedback was then discussed with all the subjects for about 5 to 10 minutes. There result of the training was clear improvement in calibration (Stone, 2000).

In some fields, experts have shown relatively well-calibrated judgments. The typical example is meteorology, where forecasts of precipitation and of maximum and minimum daily temperatures have been shown to be well calibrated (Murphy and Winkler, 1977). In contrast, financial analysts have been shown to significantly overestimate corporate earnings growth (Chatfield et al., 1989; Dechow and Sloan, 1997).

In the context of environmental risk analysis, Hawkins and Evans (1989) found that industrial hygienists provided reasonably accurate estimates of the mean and 90th percentile of a distribution of personal exposure to chemical-industry workers. Walker et al. (2003) found that experts provided reasonably well calibrated estimates of mean and 90th percentile ambient, indoor, and personal exposures to benzene.

Human decision is a function of heuristics and biases (Tversky and Kahneman, 1974). An important point to consider is when eliciting from an expert who has some sort of personal interest in the prediction outcome (Kadane and Winkler, 1988). Also, experts and novices may experience the same biases in decision-making (Ericsson and Staszewski, 1989).

Perhaps the most widely used heuristic is judgment by anchoring and adjustment (Tversky and Kahneman, 1974). With this strategy, an expert estimates an unknown with an initial value. This estimate is then adjusted to obtain a nominal value. The adjustment of the initial value (which is named the anchor) is usually too small (Slovic, 1972), a phenomenon called anchoring.

An experiment conducted by Tversky and Kahneman (1974) demonstrated this problem. Subjects were asked to estimate various quantities, stated in percentages (e.g. the percentage of African countries in the United Nations). They were given randomly chosen starting values and had to adjust it to their best estimate. Subjects whose starting values were high ended up with substantially higher estimates than those who started with low values. For example, the median estimates of the percentage of African countries in the U.N. were 25% for subjects who received 10% as their starting point and 45% for those who received 65%.

Another aspect of using expert judgment is the problem of adjusting for the overconfidence (Alpert and Raiffa, 1982; Morgan and Henrion, 1990). Shlyakhter et al. (1994) has developed an empirical model for adjusting individual expert distributions to account for overconfidence. The model uses a single parameter to calibrate the spread of an expert distribution. Hammitt and Shlyakhter (1999) use this model in their study of expert assessments related to global climate change. Other situations to consider include convergence and conflict among experts (Hynes and Vanmarke, 1977).

Expert elicitation has been criticized in many ways as well, such as selection method of experts and accurate expression of expert knowledge (O'Hagan and Oakley, 2004).

Simon and Chase (1973) suggest that for most domains it takes a minimum of ten years of experience to gain expertise. According to Ericsson, Krampe, and Tesch-Römer (1993), expert knowledge is only achieved through continuing involvement in the subject matter. Wilson (1994) states that expert knowledge is more coherent and structured than novice knowledge. Although there are certainly instances of positive correlations between experience and expertise, there is little reason to expect this relation to apply universally (Shanteau, 2002). Vegelin (2003) states that experience significantly influences accuracy.

In the context of Bayesian analysis, elicitation arises often as a method for specifying the prior distribution for an unknown of interest (O'Hagan et al, 2004). Eliciting a prior distribution is difficult due to the subjectivity nature of the prior (O'Hagan, 1998). An excellent literature review of the elicitation of prior beliefs in the Bayesian framework is presented by Kadane and Wolfson (1998).

The expert elicitation has been applied to many studies, such as future climate change (Arnell et al., 2005; Miklas et al., 1995), performance assessment of proposed nuclear waste repositories (Hora and Jensen, 2005; McKenna et al., 2003; Draper et al., 1999; Hora and von Winterfeldt, 1997; Zio and Apostolakis, 1996; Morgan and Keith, 1995; DeWispelare et al., 1995; Bonano and Apostolakis, 1991; Bonano et al., 1990), estimation of parameter distributions (Parent and Bernier, 2003; Geomatrix Consultants, 1998; O'Hagan, 1998), development of Bayesian network (Pike, 2004; Stiber et al., 1999, 2004; Ghabayen et al., 2006), and interpretation of seismic images (Bond et al., 2007).

Another question in elicitation is to determine number of experts needed. Ashton and Ashton (1985) studied judgmental forecasts of the number of advertising pages in Time magazine. The conclusion was that by combining the forecasts of four experts, error of estimates is reduced by 3.5%. Study reported that accuracy improved by increasing the panel size up to 13 experts. Hogarth model (1978) showed using at least six experts but no more than 20. Libby and Blashfield (1978) showed improvement in accuracy of forecasts when increasing the size of the expert panel from one to three, but recommended the optimum size between five and nine. Batchelor and Dua (1995) showed increase in accuracy from 10 to 22 economists. Their study also revealed a small improvement from 22 to the remaining 12.

2.2 Utilizing Expert Opinion

In uncertain situation, combining data can reduce error (Armstrong, 2001). For example, Klugman (1945) found that combining judgments led to greater improvements for estimates of heterogeneous items (irregularly-shaped lima beans in a jar) than of homogeneous items (identically-sized marbles in a jar). Krishnamurti et al. (1999), in a study of short-term weather forecasts, concluded that accurate predictions are needed from combining of six or seven estimates. Winkler and Poses (1993) examined physician's predictions of survival for 231 patients who were admitted to an intensive care unit. Physicians sometimes received unambiguous and timely feedback, so those with more experience were more accurate. They grouped the physicians into four classes based on their experience, 23 interns, four fellows, four attending physicians, and four primary care physicians. The group averages were then averaged. Accuracy improved substantially as they included two, three, and then all four groups. The error measure dropped by 12% when they averaged all four groups across the 231 patients (compared to that of just one group).

The two well-established mathematical approaches to aggregate opinions are axiomatic and Bayesian models (Boring, 2007; Clemen and Winkler, 1997). Many different methodologies have been developed for axiomatic aggregation. Previous research has considered simple averaging as a mental model of the aggregation process (Anderson, 1981; Dawes, 1979; Einhorn and Hogarth, 1975; Einhorn, Hogarth, and Klempner, 1977; Hastie, 1986; Sniezek and Henry, 1989). Many studies have suggested simple averaging of individual opinions as a method for improving the accuracy of predictions (Armstrong, 1985; Ashton, 1986; Hill, 1982; Hogarth, 1978; Zajonc, 1962; Zarnowitz, 1984).

Stone (1961) proposed a linear opinion pools in which the aggregation result is expressed as a linear combination of estimates. A linear opinion pool provides a very simple mechanism for representing unequal degrees of expertise. The determination of expertise (weight) can be a subjective matter and prone to numerous assumptions and interpretations (Genest and McConway, 1990). Cooke's classical method is a linear opinion pool, applied widely in Europe (Clemen and Winkler, 1993), including major studies of nuclear-power risks, among others (Cooke, 1994; Goossens and Harper, 1998; Jones et al., 2001). Morris (1983, 1986) introduced an axiomatic approach to expert aggregation. French (1985) and Genest and Zidek (1986) provide critical reviews of axiomatic aggregation literature.

The first formal proposal to apply the Bayesian method in expert judgment study was offered by Morris (1974, 1977). Since original research by Morris, many forms of Bayesian procedures have been introduced in various papers. Mendel and Sheridan (1989) developed a Bayesian model that allows for the aggregation of non-normal probability distributions. Clemen and Winkler (1993) proposed subjective aggregation of point estimates using 'influence diagram'. Bayesian hierarchical model (where prior depends on parameters not addressed in the likelihood) was presented by Lipscomb, Parmigiani, and Hasselblad (1998). Wisse, Bedford and Quigley (2005) introduced 'moment method' to avoid the computational complications of continuous probability distributions. In addition, Genest and Schervish (1986) consider the problem of aggregating expert judgments when the decision maker does not provide complete probabilistic assessments of the required distributions, but instead offer certain moments of the distributions.

A major issue in aggregation is the problem of dependence among experts. Judgments of multiple experts about a parameter can be extremely informative when experts are probabilistically independent, conditional on the "true" value. Clemen and Winkler (1985) reveal the number of independent experts whose combined data is equivalent to that of a larger number of dependent experts. Dependence is both central to proper combination of expert judgments and difficult to evaluate (Kallen and Cooke, 2002).

Jouini and Clemen (1996) propose a copula-based approach to combining distributions. This approach provides a flexible method for representing dependence among experts. A copula function (e.g., Nelsen, 1999) provides a way to write a joint distribution function as a function of its marginal distributions. Hammitt and Shlyakhter (1999) and Lacke (1998) use the copula aggregation models in the contexts of global climate change and colon cancer risk modeling, respectively. Clemen and Reilly (1999) suggest using the multivariate normal copula, which does not require that experts be treated symmetrically and so permits greater flexibility in modeling dependence.

Overall, identifying a likelihood function for expert probability assessments is considered as one of the actual difficulties in using Bayesian. Some of the recent research studies such as Mosleh and Forrester (2005) indicate multiple attempts to tackle the problem of developing proper likelihood functions. The appropriate likelihood model in which each expert provides a normal distribution for the target parameter developed by Winkler (1981) and studied by Winkler and Makridakis (1983), Clemen and Winkler (1985), Schmittlein et al. (1990), Chhibber and Apostolakis (1993), and Chandrasekharan et al. (1994).

Difficulties with the axioms themselves are discussed by French (1985) and Genest and Zidek (1986). Lindley (1985) gives an example of the failure of both axioms. Genest and Zidek (1986), Winkler (1968), French (1985), and Lindley (1985) all ruled for Bayesian approach. The limited available evidence on relative performance of combination methods suggests that simple averages often perform nearly as well as the theoretically superior Bayesian methods (Clemen and Winkler, 1999; Kallen and Cooke, 2002). A comprehensive review of aggregation literature, including dependence, can be found in French (1985), Ouchi (2004), Genest and Zidek (1986), French and Ríos Insua (2000).

Chapter 3: Data Collection and Characterization

3.1 Data Collection

Generally, in assessing uncertainty about an unknown of interest, information can come in form of existing evidence about the unknown, evidence on the credibility of the expert's estimate, evidence on the applicability and relevance of judgment, and data provided by the expert (Droguette and Mosleh). Experts provide qualitative information or quantitative estimates in form of a probability distribution, point estimate, range, statement or partial evidence of the unknown.

In classical mathematics, data refers to a collection of organized information, which is often the result of experience, observation or experiment. In this research, data is subjective information and refers to expert point estimate in discrete or continuous form. Estimates are generated by experts or produced by forecasting models using expert input, review or final adjustment.

A data collection plan is first established to populate a database with large number of expert estimate with corresponding seed (calibration), target (acceptance criterion or specification), true (real), or observed (as a result of experiment) values in different disciplines.

The search for evidence on expert accuracy began with a general survey of the literature, internet publications, books, refereed and non-referred sources. Additionally, a broad exploration of the relevant Dissertation Abstracts database was performed to identify work across expert judgment studies and disciplines.

The wide literature search included databases such as Econpapers, Elsvier, PubMed, IEEE Digital Library, University of Maryland Digital Library, Medline, TU Delft Database, DOE's Information Bridge, ACM Digital Library, WorldCat, CE Database, and Waste Management Research Abstracts.

Over 2000 sources and publications since 1930s were initially flagged for general relevance. Of these sources, approximately 500 were selected. Each source was examined for significance to the elicitation and aggregation of expert judgment. Additionally, TU Delft expert judgment database was used, which reports the assessment of over 800 experts on over 4000 variables, representing 80,000 elicited questions. From the selected sources in this stockpile, over 1900 point estimates were collected in more than 60 different disciplines. In the next section, data sources utilized in this research are introduced.

3.2 Description of Case Studies

In this section, a brief description of case studies used as data source is presented. An attempt is made to echo the objective of each case and convey any explanations or rationale offered by the authors to address the expert error.

3.2.1 Case #1

This study was conducted by National Human Exposure Assessment Survey (NHEXAS) using the estimates of seven experts to obtain exposure assessment in residential ambient, residential indoor and personal air Benzene concentrations ($\mu g/m^3$) in United State Environmental Protection Agency (U.S. EPA's Region V), experienced by the nonsmoking, non-occupationally exposed population. These experts were selected by a peer nomination process. Individually elicited judgments were gathered from the experts during a 2-day workshop. (Walker, K. et al. Use of expert judgment in exposure assessment - Part 1. Characterization of personal exposure to benzene. Journal of Exposure Analysis and Environmental Epidemiology, 2003 (11):308-322 and Part 2. Calibration of expert judgments about personal exposures to benzene. Journal of Exposure Analysis and Environmental Epidemiology, 2003 (13):1-16)

3.2.2 Case #2

This study focus on value-added forecasting. It claims that due to internal politics, personal agendas, and financial performance requirements that skew the process, much of the management effort directed toward forecasting actually makes the forecast worse. (Gilliland, M. Is Forecasting a Waste of Time? Supply Chain Management Review, 2002)

3.2.3 Case #3

This article examins weather trends for eight locations in Kansas to determine the relationship between rainfall, yields, and farm income. Wheat, grain sorghum, corn and soybean yields are predicted using the yield prediction formulas and historical monthly precipitation. The predicted yields are then compared to the actual county average yield for a given crop and year. Data is obtained from Kansas Agricultural Statistics for the years 1970-2001 in Colby, Tribune, Garden City, Hays, Hutchinson, Manhattan, Ottawa, and Parsons Counties. (Dumler, T. J. Rainfall and Farm Income. Risk and Profit Conference, 2003)

3.2.4 Case #4

This study lists the criteria for selecting an appropriate error measure in forecast of hotel occupancy. The reported data are taken from a 166-room hotel in the mid-west of United State. It contains two sets of figures, the predicted and the actual daily occupancies for the month of September 1996. The predicted figures are the combined product expert predictions and input of hotel managers based on their experience and expectations. (Schwartz, Z. Monitoring the Accuracy of Multiple Occupancy Forecasts)

3.2.5 Case #5

The objective of this study is to compare the clinical acumen of paediatric cardiovascular examination between various hospital paediatrician grades. Pre-echocardiography clinical diagnoses are compared with echocardiography results according to grade of referring hospital doctor (ranging from houseman to consultant). The results show that Echocardiographers had the highest clinical accuracy and the highest attempts at reaching a clinical diagnosis. Accuracy and attempts at diagnosis decreased as doctor's hospital grade decreased, from consultant to houseman. It is reported that the echocardiographers are the most accurate in the clinical detection of cardiac pathology, or its absence due to the fact that echocardiographers have the greatest experience. It is stated that Doctors with less paediatric cardiology exposure naturally experience more difficulty and housemen or senior house officers attempted the least diagnoses. Study concludes that experienced doctors are more likely to differentiate between normal and abnormal hearts. (Spiteri, A. Torpiano, J. Bailey, M. Mercieca, V. & Grech, V. A comparison of clinical paediatric murmur assessment with echocardiography. Malta Medical Journal, November 2004, (16):4)

3.2.6 Case #6

A weather precipitation case study among expert meteorologists at the University of Maryland, College Park was performed. The objectives of the study were to predict the APE of experts given their estimates and to determined the effect of expertise on expert performance. The study involved four experts who were asked to make 48-hour precipitation forecasts projections. In the field of meteorology, a 48-hour forecast of precipitation is considered moderately difficult, and requires specialized skills. The forecast were conducted on three different days for cities of Orlando, Seattle, San Francisco, New Orleans and Detroit. (Forrester, Y. 2005. The Quality of Expert Judgment: An Interdisciplinary Investigation. Weather precipitation research study among expert meteorologists at UMCP)

3.2.7 Case #7, 8, 9, 10

This study describes an evaluation of forecasting model accuracy and induced demand representation over a 10-year period in the integrated land use and transportation model, the 2000 Sacramento MEPLAN model. It is reported that error may be due to a developer model with limited sensitivity to process set too low or large zones in the outer regions which tend to underestimate the travel time. (Rodier, C. J. 2005. Verify the accuracy of land use model used in transportation and air quality planning: a case study in Sacramento, California region, MTI Report 05-02)

3.2.8 Case #11

This article evaluates the labor force, employment by industry, and occupation projections that BLS made in 1989 for the year 2000. The different causes of forecast errors, such as participation rate, are reported. The results show that in most cases, the accuracy of the BLS projections is comparable to estimates obtained from naïve extrapolative models, and hence, are of low accuracy. (Stekler, H. O. & Thomas, R. Evaluating BLS Labor Force, Employment and Occupation Projection for 2000)

3.2.9 Case #12

The Bureau of Labor Statistic (BLS) has made labor force projections since the late 1950s. Beginning in 1968, the Bureau of Labor Statistics has not considered the projection process complete until it assesses the accuracy of its projections. This article examines the errors in the labor force projections to 1995 and the sources of the errors. The analysis compares projected and actual (most recent Current Population Survey estimate) levels of the labor force. The different causes of error are reported which includes immigration, projection period, or participation by age, sex, and race. The analysis also shows that gradual improvement in the accuracy of projections occurs over time. (Fullerton, H. N. BLS. Evaluation the 1995 BLS labor force projection)

3.2.10 Case #13

This study analyzes the accuracy of the United Nations' (UN) population forecasts in the past, based on six Southeast Asian countries: Indonesia, Malaysia, Singapore, Philippines, Thailand, and Vietnam. The study uses available projected and estimated age-structured data published by the UN from 1950 onwards. The study reveals that there is inconsistancies in the accuracy of the UN projections for different countries and the errors are age specific. The analysis also shows that gradual improvement in the accuracy of projections occurs over time. The fluctuation in error amount is reported to be due to the wrong assumptions made in various past projections. (Abdullah Khan, H. T. A Comparative Analysis of the Accuracy of the United Nations' Population. Projections for Six Southeast Asian Countries. IR-03-015)

3.2.11 Case #14 & 15

In this study, census 2000 counts are used to measure forecast error in projections for April 1, 2000. The different causes of error are reported includes up and down swings in population growth, projection outliners, or forecast evaluation of the detailed demographic components. The analysis also shows that gradual improvement in the accuracy of projections occurs over time. (Campbell R. Evaluating Forecast Error in State Population Projections Using Census 2000 Counts. U.S. Bureau of Census, Population Division Working Paper Series No. 57, 2002)

3.2.12 Case #16

In this article, a number of forecasts as well as actual data are provided for a monthly electric bill from January, 1991, through December 2000 for educational purposes. Paper claims that the values provide a real dataset to use for applications ranging from simple graphical analysis through a variety of time series forecasting methods. (McLaren, C. H. & McLaren, B. J. 2003. Electric Bill Data. Journal of Statistics, Ed. [Online], 11, 1)

3.2.13 Case #17

This work involves forecasting the number of domestic and international airline passengers in Saudi Arabia. Annual data from 1975 to 1986 was used and categorized into 16 variables. The forecast was obtained using the Model Quest Miner package, using some historical data for developing the model then proceeds to an evaluation phase. The period used for developing the model for the number of passengers was 18 years, while the period used for evaluation was 6 years for the five cities of Cities of Dhahran, Madina, Riyadh, Jeddah and Taif in Saudi Arabia. (BaFail, A. O. Applying Data Mining Techniques to Forecast Number of Airline Passengers in Saudi Arabia, Domestic and International Travels. King Abdul Aziz University, 2004)

3.2.14 Case #18

These data are obtained from Dr. Ali Mosleh from the University of Maryland, Mechanical Engineering Department, Reliability Engineering Program, reporting repair time for mechanical and electrical equipment. (Forrester, Y. The Quality of Expert Judgment: An Interdisciplinary Investigation, 2005)

3.2.15 Case #19

The case study contains experts' responses to 11 questions on Adult Weight Management, and the completion of a brief inquiry about experts' expertise. The entirety of experts attributes is used to predict the performance of experts. A weight management survey instrument is administered to registered dieticians with varying degrees of expertise. Experts are given a clinical nutrition diagnostic problem regarding the recommended "very low calorie diet" for an obese girl. Experts wereasked to make a judgment about maximum recommended Kcal per day. (Forrester, Y. The Quality of Expert Judgment: An Interdisciplinary Investigation, 2005)

3.2.16 Case #20

A) In this study, the Foodborne Illness Risk Ranking Model (FIRRM) is developed, which is a decision-making tool that quantifies and compares the relative burden to society of 28 food-borne pathogens. An expert elicitation survey was designed and implemented, in which experts were asked to estimate, for each pathogen, the percentage of illnesses attributable to each food vehicle. The survey was developed, with the aid of Dr. Paul Fischbeck, Carnegie Mellon University, a recognized authority in the field of expert elicitation, using standard methodologies found in the literature (Morgan et al. 1990; Cooke 1991). The survey included 11 major pathogens and elicited uncertainty bounds around responses. The survey was sent to a peer-reviewed list of 101 scientists, public health officials, and food safety policy experts; and received 45 responses. The data include experts' best judgment estimates of attribution percentages for Campylobacter and Listeria and outbreak data.

(Batz, M. B., et al. Identifying the Most Significant Microbiological Food-borne Hazards to Public Health: A New Risk Ranking Model, Food Safety Research Consortium. Discussion Paper Series Number (1) - FIRRM Food Attribution Percentages for Illnesses from Foodborne Campylobacter and Listeria monocytogenes, 2004)

B) Hoffmann et al. develops a formal protocol for expert elicitation with large, cross-functional expert panels and uses formal survey methods to take advantage of variation in individual expert uncertainty and inconsistancy among experts as a means of quantifying and comparing sources of uncertainty about parameters of interest. The pool of respondents represent a broad range of workplaces; three respondents reported having significant work experience in multiple institutional settings and the remainder were evenly distributed among government, academia, and industry. It is reported that experts' backgrounds and experiences as well as self-reported pathogen expertise help explain variation in individual experts' ranges. Respondents who identify government as their primary career setting have tighter ranges than those whose careers have been primarily in academia, industry, or multiple sectors. Those with significant career experience in multiple sectors have the largest ranges, followed by those in industry, and followed by academia. Highest degree also explains variation in range. Those with master's degrees have the least confidence in their best estimates, and Doctors of Veterinary Medicne or DVMs have the most.

(Hoffmann, S., et al. Eliciting Information on Uncertainty from Heterogeneous Expert Panels: Attributing U.S. Foodborne Pathogen Illness to Food Consumption. RFFDP6-17, April of 2006)

3.2.17 Case #21

This research, employing 11 experts who estimated an exposure parameter (the percentages of four nickel species) in 12 workplaces in a nickel primary production industry, providing a large dataset from which useful inferences can be drawn about the quality of expert judgments and the variability among the experts. It describes the application of Bayesian ideas to the comparison of expert opinions, mathematically combining expert opinions and refining these combined expert opinions with actual workplace measurements. The study reports that expertise does not necessarily require intimate familiarity with the workplace, however, the expert judgment knowledge has indeed enhanced the quality of the combination of expert judgment. (Ramachandran, G. et al. Expert Judgment and Occupational Hygiene: Application to Aerosol Speciation in the Nickel Primary Production Industry)

3.2.18 Case #22

The accuracy of cause-specific mortality by physician review is reported in this article. Data is drawn from a multi-center validation study of 796 adult deaths that occurred in hospitals in Tanzania, Ethiopia, and Ghana. Study reveals that the physician review shows a high diagnostics accuracy. (Quigley, M. A., et al. Diagnosis accuracy of physician review expert algorithms and data-derived algorithms in adult verbal autopsies International epidemiological Association, International Journal of Epidemiology, 1999(28): 1081-1087)

3.2.19 Case #23

In this article, four forecasts are evaluated for relative forecast accuracy by examining their performance over specified period of time. The reported actual price data and individual forecast series extracted are quarterly observations on and forecasts of the USDA seven-market-average hog price for barrows and gilts (200-220 lb.) from the third quarter of 1973 through the second quarter of 1986. According to this article, the individual forecast data are an expert's forecast and the expert's forecasts are for one-quarter-ahead cash prices made by Glen Grimes, professor of Agricultural Economics. The futures forecast prices would correspond directly to the expert forecasts. The futures forecasts for each period are the closing price quoted in the annual Yearbook of the Chicago Mercantile Exchange for the day Grimes' forecast is published and for the contract that would expire as close as possible to the end of the one-quarter lead time. The results of this study reveals that the it would have been better for analyst to use a composite forecast rather than tempting to identify a "best" individual value obtained from each of the forecast. (McIntosh, S. & Bessler, A. Forecasting Agricultural Prices Using a Bayesian Composite Approach. Southern Journal of Agricultural Economics, December of 1988)

3.2.20 Case #24

In this article, AEPCO and the University of Arizona, Department of Agriculture and Resource Economics (AREC) collaborate during the fall semester 2005 on a project to improve forecasts of next-day electricity load reported in Mega-Watt. The project is conducted as part of an AREC graduate class in applied econometrics. Mr. Cathers of AEPCO developed a detailed proposal outlining specific objectives for improving forecast accuracy. Dr. Gary Thompson of the University of Arizona, AREC, agreed to coordinate the department's efforts and conduct the project in connection with his graduate course, Advance Applied Econometrics. Students developed econometric models for forecasting next-day hourly load profiles. The particular econometric models developed are known as ARIMA (autoregressive, integrated, moving average) models.
It is concluded in this paper that existing methods using expert judgment appear to have been sufficiently accurate for AEPCO's current load levels and thus it is suggested that AEPCO may continue to employ expert judgment methods while comparing their daily forecasts to those derived from statistical models. (Cathers, C. A. & Thompson, G. D. 2006. Forecasting Short-Term Electricity Load Profiles. Sierra Southwest Cooperative Services, Inc. The University of Arizona, Cardon Research Papers)

3.2.21 Case # 25 & 26

Tennessee Valley Authority produces its own forecasts of regional economic activity based on forecasts of the national economy developed by a forecasting service, Global Insight. These forecasts are publicly distributed throughout the Tennessee Valley. The reported data are TVA Economic Forecast Five-Year Forecast Gross Product in Billions of Dollars from 1980 to 1995. It is stated in the study appendix that the regional economic forecast performance improvement can be attributed, in part, to the better performance of the national forecasts and to improvements in the TVA economic forecasting process, including validation procedures. (Tennessee Valley Authority (TVA). Appendix B – Methodology and Results from Socioeconomic Modeling. Final Environmental Assessment)

3.2.22 Case #27

This paper considers a dilemma an analyst faces as influential forecaster. It states that clients request an unbiased forecast but pressures sometimes exist to provide a bias forecast. The impact of these pressures on the quality of forecasts is evaluated and the different causes of error are reported such as the difference between forecasting and decision-making or lack of control on new product launches. (Ehrman, C. M. & Shugan, S. M. 1995. The Forcaster's Dilemma. Marketing Science, 14(2): 123-127, Springer)

3.2.23 Case #28

Over the last fifteen years, the Delft University of Technology (both the Safety Science Group and the Department of Mathematics of TU Delft) has developed methods and tools to support the formal application of expert judgment. Over 800 experts assessed over 4000 variables, in total representing more than 80,000 elicited questions. Applications were made in a variety of sectors, such as nuclear, chemical and gas industries, toxicity of chemicals, external effects (pollution, waste disposal sites, inundation, volcano eruptions), aerospace and aviation sector, occupational sector, health sector, and the banking sector. Expert judgment data provided by Dr. R. M. Cooke on 2009. (Goossens, L. H. J.; Cooke, R. M.; Hale, A. R. & Rodic´-Wiersma, Lj. Fifteen years of expert judgement at TU Delft. Safety Science 46 (2008) 234–244)

3.3 Data Characterization

In the expert judgment case studies, where empirical data is collected, there is a range of reasons explaining the expert error. This includes, but not limited to, career affiliation, academic degree, field of expertise, or years of experience. The errors of expert estimates vary by subject matters as well. For example, the study conducted by Hoffmann, Fischbeck, Krupnick, and McWilliams (2006) show that variability in best estimates does differ by professional background and discipline as well as expert characteristics. Respondents who identify government as their primary career setting have smaller ranges than those whose careers have been primarily in academia or industry; individuals with significant career experience in multiple sectors have the largest ranges, followed by those in industry and academia.

For the forecasts obtained by model, in addition to model inputs and assumptions, there is a range of reasons listed to explain the error of forecasts such as model types, forecast period and projection horizon, forecast accuracy measures used, additional information that becomes available, the size of the error, seasonal and geographical errors, and so on.

Overall, there are many factors affecting the estimate accuracy such as expert attributes, calibration method, decision processes, aggregation procedure, and so on. Inconsistencies caused by these elements are accepted as inherent variation in the modeling and assessment processes in this research study. The purpose is to capture actual errors (though the sources of these fallacies remain unknown) and examine the formulated likelihood functions in dealing with these variations. This is especially true for generic likelihood function which is domain independent, but is made from a pool of data from different fields.

A question may arise as how to draw a conclusion from information without any boundaries. It should be noted that there are circumstances that expert previous performance is not entirely realized. There are also events that are beyond our direct experience. In these cases, decision makers are indeed puzzled about the quality level of the opinion, or in other words, the degree of confidence to place in the judgment. The generic likelihood function developed in this study can justly be used to update the expert estimate when facing with lack of such information.

3.4 Selection of Forecast Accuracy Measure

According to Armstrong and Fildes (1995), the objective of a forecast accuracy measure is to provide information about the error distribution. It has been shown by Chen and Yang (2004) that Mean Square Error (MSE) is the optimal selection when the errors are normally distributed. However, MSE and similar measures are not suitable for this study since they are not unit-free. Absolute performance measures such as simple difference between the estimate and true value may produce very big numbers due to outliers, which can make the comparison of different estimates not feasible.

It is generally accepted that there is no single best accuracy measure, and selecting an assessment method is essentially a subjective decision. Figure 1 depicts the logic of selecting the forecast accuracy measure in this research. As reflected in this diagram, general and specific provisions were first defined. Among the most popular measures listed, relative error measure is chosen since it is scale-independent, interpretable, minimally impacted by outlier observations or errors and can eliminate the bias introduced by possible trends, and seasonal components. Amongst the relative error candidates, the simplest form was selected since it seemed to be able to satisfy the majority of established requirements, while being easy enough for numerical calculations:

$$E = \frac{u'}{u} \qquad \text{Equation 1}$$

u : is the quantity of interest,

u' : is the expert opinion and

E : is the relative error.

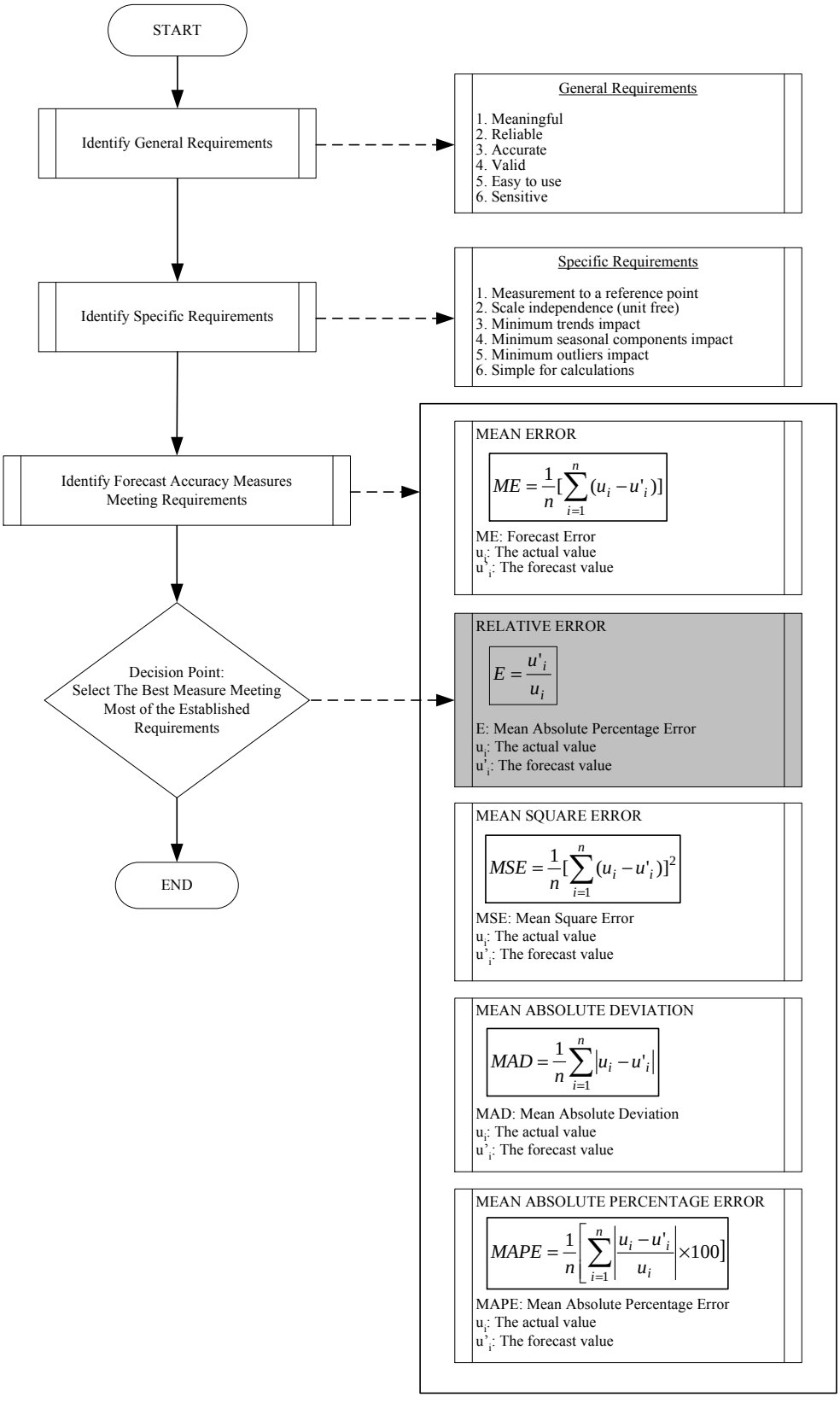


Figure 1. Process of Selecting Forecast Accuracy Measure

The unit of measurement has minimum impact on the development of the domain-independent (generic) likelihood function. For illustration purpose, consider an expert predict, i.e., tomorrow's temperature to be 78ºF. If the observed temperature turns to be 85ºF, the relative error of estimates becomes 0.88. One may argue that the prediction under same conditions results in relative error of 0.92 in Celsius scale. Therefore, unit of measurement still plays a role in calculations despite the fact that the dimensionless relative error is selected as the accuracy measure of estimate. However, it should be noted that first, the impact of this types of errors on the desired outcome where they are used in the research (distribution identification of expert relative errors) phases out as the population of data gets larger. Additionally, an expert should have the same accuracy in predicting a same unknown in different measuring systems (i.e. in temperature estimation example, expert relative error should be 0.88 in Fahrenheit and Celsius systems). This is because expert knowledge or expertise (or any other attributes qualified one as an expert) does not change from one measuring system to another. Even if estimates do change in various measuring systems, these kinds of inconsistencies are accepted as inherent variation in modeling and assessment processes o test the capability and robustness of formulated likelihood functions in tolerating variations.

Chapter 4: Bayesian Formalism

4.1 Introduction

Conceptually, the formulation of the Bayesian method for use of expert opinion is quite simple. The expert estimate is treated as a piece of evidence about the unknown quantity of interest. This evidence is then used to update the analyst's or decision maker's own (prior) knowledge through Bayes'.

$$\pi(u|u') = \frac{L(u'|u)\pi_o(u)}{\int L(u'|u)\pi_o(u)du} \qquad \text{Equation 2}$$

u : is the quantity of interest,

u' : is the set of the experts' opinions,

$\pi_o(u)$: is the decision maker's prior or initial state of knowledge about the unknown quantity u (prior to obtaining the opinion of the experts). Prior distributions are used to describe the uncertainty surrounding the unknown.

$L(u'|u)$: is the likelihood of the evidence u' given that the true value of the unknown quantity is u. The likelihood function asks this question: If the true value is u, what is the probability that the expert estimates it as u'? As such, the likelihood function is a statement on the accuracy and credibility of the expert as viewed by the decision maker.

$\pi(u|u')$: is posterior distribution representing the decision maker's updated state of knowledge about the unknown quantity, u'. After observing the data (in this case expert opinion), the posterior distribution provides a coherent post data summary of the remaining uncertainties.

The first formal framework of the Bayesian methods for use of expert opinion was presented by Morris (1974, 1977). Morris's work fully establishes the foundations for the Bayesian paradigm in the analysis of expert judgment. Building on Morris's method, Mosleh and Apostolakis (1986) proposed the use of 'Additive' and 'Multiplicative' error models for constructing the likelihood functions, expressing the experts' assessments as the sum (or ratio) of the true value of unknown and an 'error' term. Mathematically speaking:

1) Additive error model: $u = u' + E$ Equation 3

2) Multiplicative error model: $E = \frac{u'}{u}$ (refer to Equation 1)

Still, the main problem in applying the Bayesian technique remains as complications associated with the development of a suitable likelihood function. This distribution is a probabilistic model for data and must capture the interrelationships among estimates and the unknown of interest. Particularly, it must account for the bias of the individual estimate, represent expert expertise and be able to model dependencies among experts.

4.2 Governing Model

The prior knowledge of u is updated using the likelihood function developed by relative errors. The error distribution can be marginalized in terms of a finite set of parameters ($\underline{\theta}$) or epistemic uncertainty, which by itself is a variable symbolized by a population variability distribution of g($\underline{\theta}$) or aleatory uncertainty. Using likelihood averaging technique:

$$L\left(u'|u,\underline{E}\right)=\int_{\underline{\theta}} L\left(u'|u,\underline{\theta}\right)g\left(\underline{\theta}|\underline{E}\right)d\underline{\theta} \qquad \text{Equation 4}$$

Applying Equation 2:

$$\pi\left(u|u',\underline{E}\right)=\frac{\int_{\underline{\theta}} L\left(u'|u,\underline{\theta}\right)g\left(\underline{\theta}|\underline{E}\right)d\underline{\theta}\pi_0(u)}{\int_{u}\int_{\underline{\theta}} L\left(u'|u,\underline{\theta}\right)g\left(\underline{\theta}|\underline{E}\right)d\underline{\theta}\pi_0(u)du} \qquad \text{Equation 5}$$

Where,

u : is the quantity of interest

u' : is expert estimate

$\underline{E}=\left(E_1...E_n\right)$: is evidence or relative error of estimates

$\underline{\theta}=\left(\theta_1...\theta_n\right)$: reflects that parameters of error distribution

In the next sections, the likelihood functions and posterior distributions are constructed for homogenous, nonhomogenous and hybrid pools of data. The hybrid or mixed case has been formulated in this research only.

4.3 Construction of the Likelihood and Posterior: Homogenous Pool

As represented in Table 1 and illustrated in Figure 2, the available information regarding the quantity of interest (u) is comprised of expert's estimates ($u'_1 ... u'_n$) and evidence in form of error of estimates ($E_1 ... E_n$). The overall distribution of errors of estimates, *f(E)*, can be characterized in terms of a finite set of parameters. Postulating a lognormal distribution:

$\underline{\theta} = (E_{50}, \sigma_E)$

(E_{50}): is the median of the error distribution

(σ_E): is the standard deviation of the error distribution

$$f(E) = \frac{1}{\sqrt{2\pi}\sigma_E E} e^{-\frac{1}{2}\left(\frac{\ln E - \ln E_{50}}{\sigma_E}\right)^2} \qquad \text{Equation 6}$$

The probability distribution of errors also represents the likelihood of errors given the distribution parameters. Assuming independence among experts:

$$L(E_1 ... E_n | E_{50}, \sigma_E) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_E E_i} e^{-\frac{1}{2}\left(\frac{\ln E_i - \ln E_{50}}{\sigma_E}\right)^2} \qquad \text{Equation 7}$$

Estimating the set of likelihood parameters:

$$\pi(E_{50}, \sigma_E | E_1 ... E_n) = \frac{L(E_1 ... E_n | E_{50}, \sigma_E)\pi_0(E_{50}, \sigma_E)}{\int_{E_{50}} \int_{\sigma_E} L(E_1 ... E_n | E_{50}, \sigma_E)\pi_0(E_{50}, \sigma_E) dE_{50} d\sigma_E} \qquad \text{Equation 8}$$

The term $\pi_0(E_{50}, \sigma_E)$ is the prior which is assumed to be a lognormal distribution as well. A generic likelihood function, $\overline{f}(E)$, can be formulated by de-conditioning the posterior (Equation 8) using:

$$\overline{f}(E) = \int_{E_{50}} \int_{\sigma_E} f(E | E_{50}, \sigma_E)\pi(E_{50}, \sigma_E | E_1 ... E_n) dE_{50} d\sigma_E \qquad \text{Equation 9}$$

To construct the likelihood function, $L(u'|u)$, based on the likelihood of relative errors $L(E|u)$, the relation between the distribution of relative errors, *f*(E), and the distribution of estimates, *f*(u'), must be established.

$$E = \frac{u'}{u} \Rightarrow udE = du' \Rightarrow \frac{dE}{du'} = \frac{1}{u} \qquad \text{Equation 10}$$

$$f(u')du' = f(E)dE \Rightarrow f(u') = \frac{dE}{du'} f(E) \qquad \text{Equation 11}$$

$$f(u') = \frac{1}{u} f(E) \qquad \text{Equation 12}$$

Therefore the likelihood function $L(u'|u)$ can be linked to $L(E|u)$ as:

$$L(u'|u) = \left(\frac{1}{u}\right) L(E|u) = \left(\frac{1}{u}\right) \frac{1}{\sqrt{2\pi}\sigma_E E} e^{-\frac{1}{2}\left(\frac{\ln E - \ln E_{50}}{\sigma_E}\right)^2} \Rightarrow$$

$$L(u'|u,\underline{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_E u'} e^{-\frac{1}{2}\left(\frac{\ln u' - \ln u - \ln E_{50}}{\sigma_E}\right)^2} \qquad \text{Equation 13}$$

The above equation is the first term in Equation 4. Estimating the epistemic uncertainty of $\underline{\theta}$:

$$g(\underline{\theta}|\underline{E}) = \frac{L(\underline{E}|\underline{\theta})\pi_0(\underline{\theta})}{\int_{\underline{\theta}} L(\underline{E}|\underline{\theta})\pi_0(\underline{\theta}) d\underline{\theta}} \qquad \text{Equation 14}$$

Where,

$$L(\underline{E}|\underline{\theta}) = \prod_{i=1}^{n} L(E_i|\theta_i) \qquad \text{Equation 15}$$

The new expert estimate can now be updated using Equation 5. The mean or median of the posterior, both shown with symbol (μ) in figures, as the distribution marker, is compared with the true value (μ/u) in order to determine if and how much the formulated likelihood function has been able to reduced the error of estimates. This process is depicted in Figure 3.

Table 1. Representation of Homogenous Data

Estimate (i = 1...n)	True Value	Expert's Error ($E_i = \frac{u'_i}{u}$)
u'_1	u	E_1
u'_2		E_2
…		…
…		…
…		…
u'_n		E_n

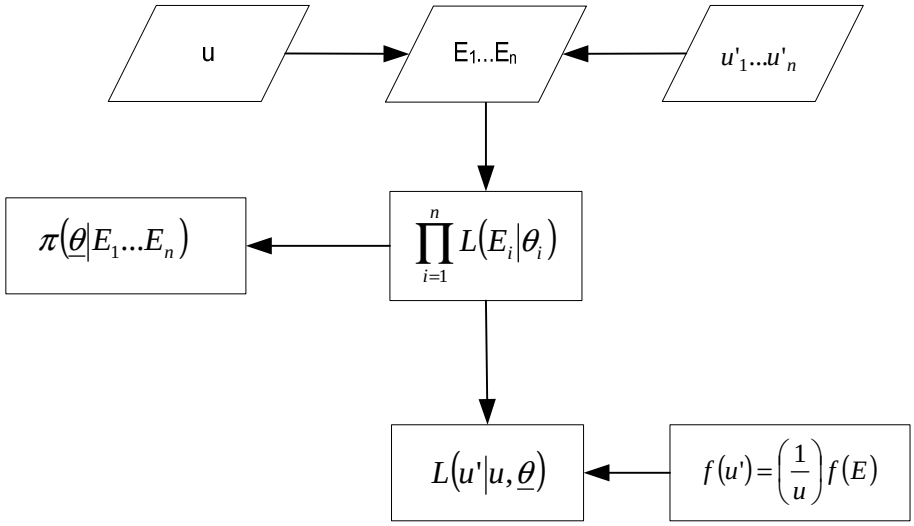


Figure 2. Construction of Likelihood Functions for Homogenous Data

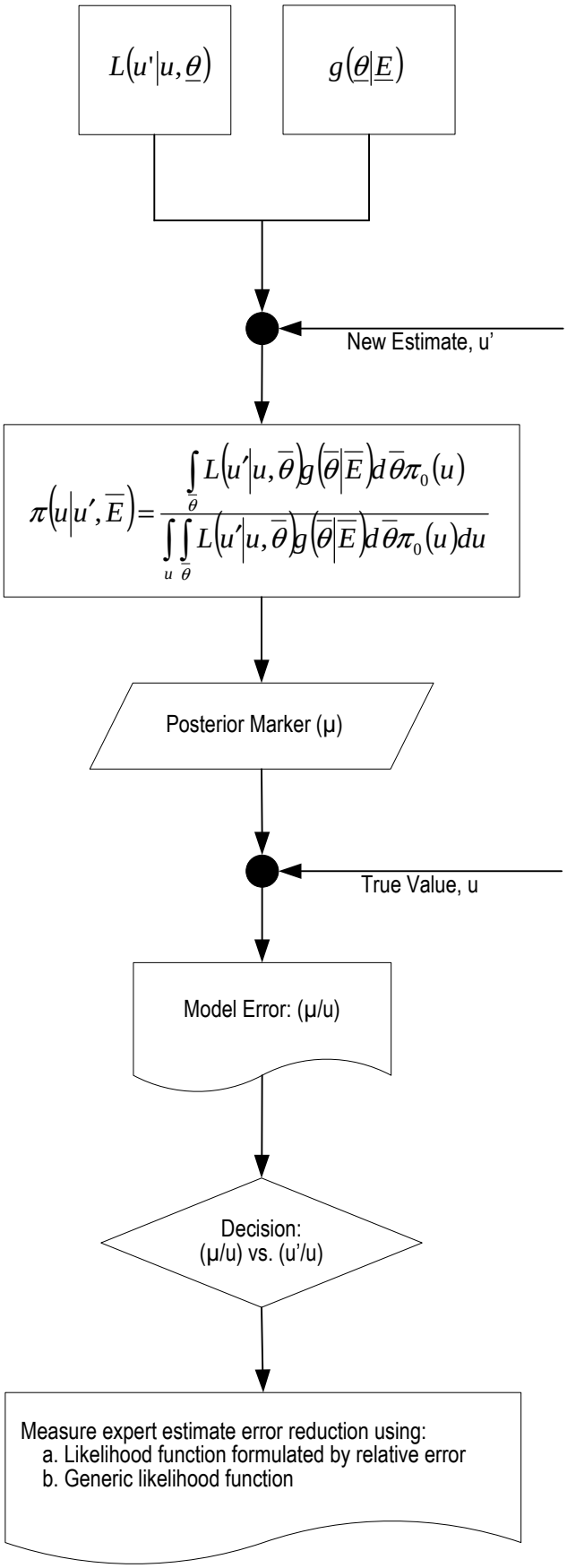


Figure 3. Treatment of Homogenous Data

4.4 Construction of Likelihood and Posterior: Nonhomogenous Pool

As represented in Table 2 and Figure 4, the information regarding true values of $(u_1 \ldots u_n)$ is comprised of expert's estimates $(u'_1 \ldots u'_n)$. The error distribution can be marginalized in terms of a finite set of parameters ($\underline{\theta}$), which by itself is a variable symbolized by a population variability distribution of g($\underline{\theta}$). This 'hyper' distribution can be characterized by a set of 'hyper-parameters' $(\underline{\omega})$:

$$\underline{\omega} = (\omega_1 \ldots \omega_n) \qquad \text{Equation 16}$$

$$g(\underline{\theta}) = g(\underline{\theta}|\underline{\omega}) \qquad \text{Equation 17}$$

The likelihood function for the data point (u'_i, u_i) and therefore E_i is estimated by eliminating the epistemic uncertainty over $\underline{\theta}$:

$$L(E_i|\underline{\omega}) = \int_{\underline{\theta}} L(E_i|\underline{\theta}) g(\underline{\theta}|\underline{\omega}) d\underline{\theta} \qquad \text{Equation 18}$$

Under the assumption of independence among experts:

$$L(\underline{E}|\underline{\omega}) = \prod_{i=1}^{n} \int_{\underline{\theta}} L(E_i|\underline{\theta}) g(\underline{\theta}|\underline{\omega}) d\underline{\theta} \qquad \text{Equation 19}$$

Estimating the 'hyper-parameters' using likelihood function $L(E|\underline{\omega})$:

$$\pi(\underline{\omega}|\underline{E}) = \frac{\left(\prod_{i=1}^{n} \int_{\underline{\theta}} L(E_i|\underline{\theta}) g(\underline{\theta}|\underline{\omega}) d\underline{\theta} \right) \pi_0(\underline{\omega})}{\int_{\underline{\omega}} \left(\prod_{i=1}^{n} \int_{\underline{\theta}} L(E_i|\underline{\theta}) g(\underline{\theta}|\underline{\omega}) d\underline{\theta} \right) \pi_0(\underline{\omega}) d\underline{\omega}} \qquad \text{Equation 20}$$

The posterior expected distribution as $\bar{g}(\underline{\theta}|\underline{E})$ is estimated by eliminating the aleatory uncertainty over $\underline{\omega}$:

$$\bar{g}(\underline{\theta}|\underline{E}) = \int_{\underline{\omega}} g(\underline{\theta}|\underline{\omega}) \pi(\underline{\omega}|E) d\underline{\omega} \qquad \text{Equation 21}$$

The new expert estimate can be updated using general Bayesian procedure:

$$\pi\left(u\middle|u',\underline{E}\right)=\frac{\int_{\underline{\theta}}\bar{g}\left(\underline{\theta}\middle|\underline{E}\right)L\left(u'\middle|u,\underline{\theta}\right)\pi_0(u)}{\int_u\left(\int_{\underline{\theta}}\bar{g}\left(\underline{\theta}\middle|\underline{E}\right)L\left(u'\middle|u,\underline{\theta}\right)\right)\pi_0(u)du} \qquad \text{Equation 22}$$

Table 2. Representation of Non-Homogenous Data

Estimate (i = 1...n)	True Value (i = 1...n)	Expert's Error (i = 1...n)
u'_1	u_1	E_1
u'_2	u_2	E_2
…	…	…
u'_n	u_n	E_n

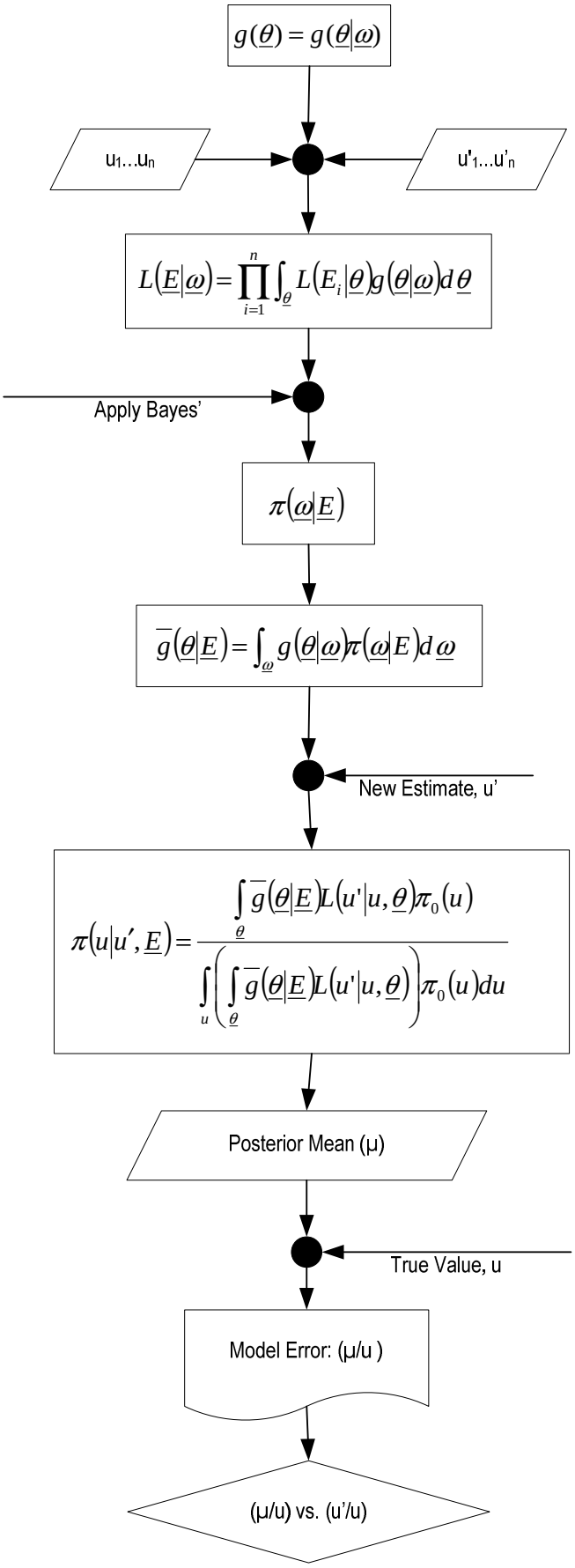


Figure 4. Treatment of Non-Homogenous Data

It can be shown that homogenous is an especial case of nonhomogenous, when evidence provides perfect knowledge of the parameter set $\underline{\theta}$. Additionally, error distribution parameters ($\underline{\theta}$) have no aleatory variability. The distribution $g(\underline{\theta}|\underline{\omega})$ turns to a Dirac Delta function and hence in Equation 20:

$$\pi(\underline{\omega}|\underline{E}) = \frac{\left(\prod_{i=1}^{n}\int_{\underline{\theta}} L(E_i|\underline{\theta})\delta(\underline{\theta}-\underline{\omega})d\underline{\theta}\right)\pi_0(\underline{\omega})}{\int_{\underline{\omega}}\left(\prod_{i=1}^{n}\int_{\underline{\theta}} L(E_i|\underline{\theta})\delta(\underline{\theta}-\underline{\omega})d\underline{\theta}\right)\pi_0(\underline{\omega})d\underline{\omega}} \qquad \text{Equation 23}$$

Since for Dirac Delta function we have

$$f(x_0) = \int f(x)\delta(x-x_0)dx \qquad \text{Equation 24}$$

Then Equation 23 changes to:

$$\pi(\underline{\omega}|\underline{E}) = \frac{\prod_{i=1}^{n} L(E_i|\underline{\omega})\pi_0(\underline{\omega})}{\int_{\underline{\omega}}\prod_{i=1}^{n} L(E_i|\underline{\omega})\pi_0(\underline{\omega})d\underline{\omega}} \qquad \text{Equation 25}$$

From Equation 21, we have:

$$\overline{g}(\underline{\theta}|\underline{E}) = \int_{\underline{\omega}} \delta(\underline{\theta}-\underline{\omega})\left(\frac{\prod_{i=1}^{n} L(E_i|\underline{\omega})\pi_0(\underline{\omega})}{\int_{\underline{\omega}}\prod_{i=1}^{n} L(E_i|\underline{\omega})\pi_0(\underline{\omega})d\underline{\omega}}\right)d\underline{\omega} \qquad \text{Equation 26}$$

Applying Equation 24, which is the same equation as for homogenous data:

$$g(\underline{\theta}|\underline{E}) = \frac{\prod_{i=1}^{n} L(E_i|\theta_i)\pi_0(\underline{\theta})}{\int_{\underline{\theta}}\prod_{i=1}^{n} L(E_i|\theta_i)\pi_0(\underline{\theta})d\underline{\theta}} \qquad \text{Equation 27}$$

4.5 Construction of Likelihood and Posterior: Hybrid Pool

In the case of mixed or hybrid data, for each instance of (k =1...N), the estimate $i=(1...M_k)$ of (u_k) is (u'_{ki}), representing evidence (E_{ki}). Therefore, as represented in Table 3, the relative error term has two dimensions of (i, k) to cover all k instances:

$$\pi(\underline{\omega}|\underline{E})=\frac{\prod_{k=1}^{N}\left(\prod_{i=1}^{M_k}\int_{\underline{\theta}}L(E_{ik}|\underline{\theta})g(\underline{\theta}|\underline{\omega})d\underline{\theta}\right)\pi_0(\underline{\omega})}{\int_{\underline{\omega}}\prod_{k=1}^{N}\left(\prod_{i=1}^{M_k}\int_{\underline{\theta}}L(E_{ik}|\underline{\theta})g(\underline{\theta}|\underline{\omega})d\underline{\theta}\right)\pi_0(\underline{\omega})d\underline{\omega}} \qquad \text{Equation 28}$$

$$\bar{g}(\underline{\theta}|\underline{E})=\int_{\underline{\omega}}g(\underline{\theta}|\underline{\omega})\frac{\prod_{k=1}^{N}\left(\prod_{i=1}^{M_k}\int_{\underline{\theta}}L(E_{ik}|\underline{\theta})g(\underline{\theta}|\underline{\omega})d\underline{\theta}\right)\pi_0(\underline{\omega})}{\int_{\underline{\omega}}\prod_{k=1}^{N}\left(\prod_{i=1}^{M_k}\int_{\underline{\theta}}L(E_{ik}|\underline{\theta})g(\underline{\theta}|\underline{\omega})d\underline{\theta}\right)\pi_0(\underline{\omega})d\underline{\omega}}d\underline{\omega} \qquad \text{Equation 29}$$

As we can see the homogenous and non-homogeneous cases are special case of the mixed pool. For example Equation 28 is reduced to Equation 20 when for each true value we have only one estimate, that is when $M_K = 1$ for all k.

Table 3. Representation of Hybrid Data

Cases (k = 1...N)	Estimate (i = 1…M_k)	True Value (k = 1...N)	Expert's Error ($E_{ki}=\frac{u'_{ki}}{u_k}$)
1	$[u'_{11} \quad u'_{12}]$	u_1	$[E_{11} \quad E_{12}]$
2	$[u'_{23} \quad u'_{24} \quad u'_{25} \quad u'_{26}]$	u_2	$[E_{23} \quad E_{24} \quad E_{25} \quad E_{26}]$
3	u'_{37}	u_3	E_{37}
...	…	…	…
N	$\lfloor u'_{N,M_k-1} \quad u'_{N,M_k} \rfloor$	u_N	$\lfloor E_{N,M_k-1} \quad E_{N,M_k} \rfloor$

Chapter 5: Data-Informed Calibration of Expert Opinions

5.1 Introduction

The objective of this section is data-driven expert calibration within the Bayesian formalism. Calibration is defined as the degree of agreement between the estimates of an event compared to its actual occurrence value.

In some fields, experts have been shown to make relatively well-calibrated judgments. The typical example is meteorology (Murphy and Winkler, 1977). In contrast, financial analysts have been shown to significantly overestimate corporate earnings growth (Chatfield et al., 1989; Dechow and Sloan, 1997). Hawkins and Evans (1989) found that industrial hygienists provided reasonably accurate estimates of the mean and 90^{th} percentile of a distribution of personal exposure to chemical-industry workers.

An investigation of several practical questions is conducted regarding the calibration of expert judgment using empirical data. The objectives are:

[1] Measuring the uncertainty surrounding the unknown of interest in the Bayesian framework, given an expert estimate

[2] Formulate a 'generic' likelihood function based on large numbers of observed expert relative errors in different domains, and

[3] To explore whether use of generic likelihood would reduce future prediction errors

[4] Performance comparison between posterior mean and median in reducing the overall errors of experts when using generic likelihood distribution

5.2 Methodology

Likelihood functions for homogenous, nonhomogenous, and hybrid data have been developed in Chapter 4. The further steps taken to conduct the study in this chapter include:

I. Descriptive statistics of empirical errors are produced to quantitatively summarize the data.

II. Relative errors are fitted into matching probability distributions to select the form of the likelihood function.

III. A generic error likelihood distribution for use in Bayesian assessment of expert opinion is developed using empirical data.

IV. Bayesian method is employed to update the expert estimate using:

   i. Case-specific likelihood function

   ii. Domain-independent or generic likelihood function

To perform the analyses flat or noninformative priors are used. This approach can provide a basis for defining knowledge or expertise of information sources (in the matter of estimating true value) relative to the analyst. Additionally, if the decision maker or analyst believes, as would normally be the case in consulting experts, that prior information should have little or no impact on the posterior, a noninformative prior of true value would be a proper modeling choice (Edwards, 1963).

In the Bayesian method, the posterior marker or estimator is compared with the true value to assess the error of updated estimate. According to Christensen and Huffman (1985), the most often used posterior markers have been the mean, median, and mode, with no consensus among experts on which is the most appropriate.

Barnett (1982) believes that there would be no useful criterion for choosing a single value than to use the most likely value, unless further information on the consequences of incorrect choice is incorporated. Berger (1980) states the mean and median are often better values than the mode. According to Cox and Hinkley (1974), if it is required to summarize the posterior distribution in a single quantity, mean is frequently the most sensible. In particular, if the prior density is exactly or approximately constant, the use of the mean of the likelihood function with respect to the parameter is indicated. For illustration purposes, step by step numerical calculations of posterior markers for an example of each data type are presented. The steps to numerical execution of the first part are depicted in Figure 5.

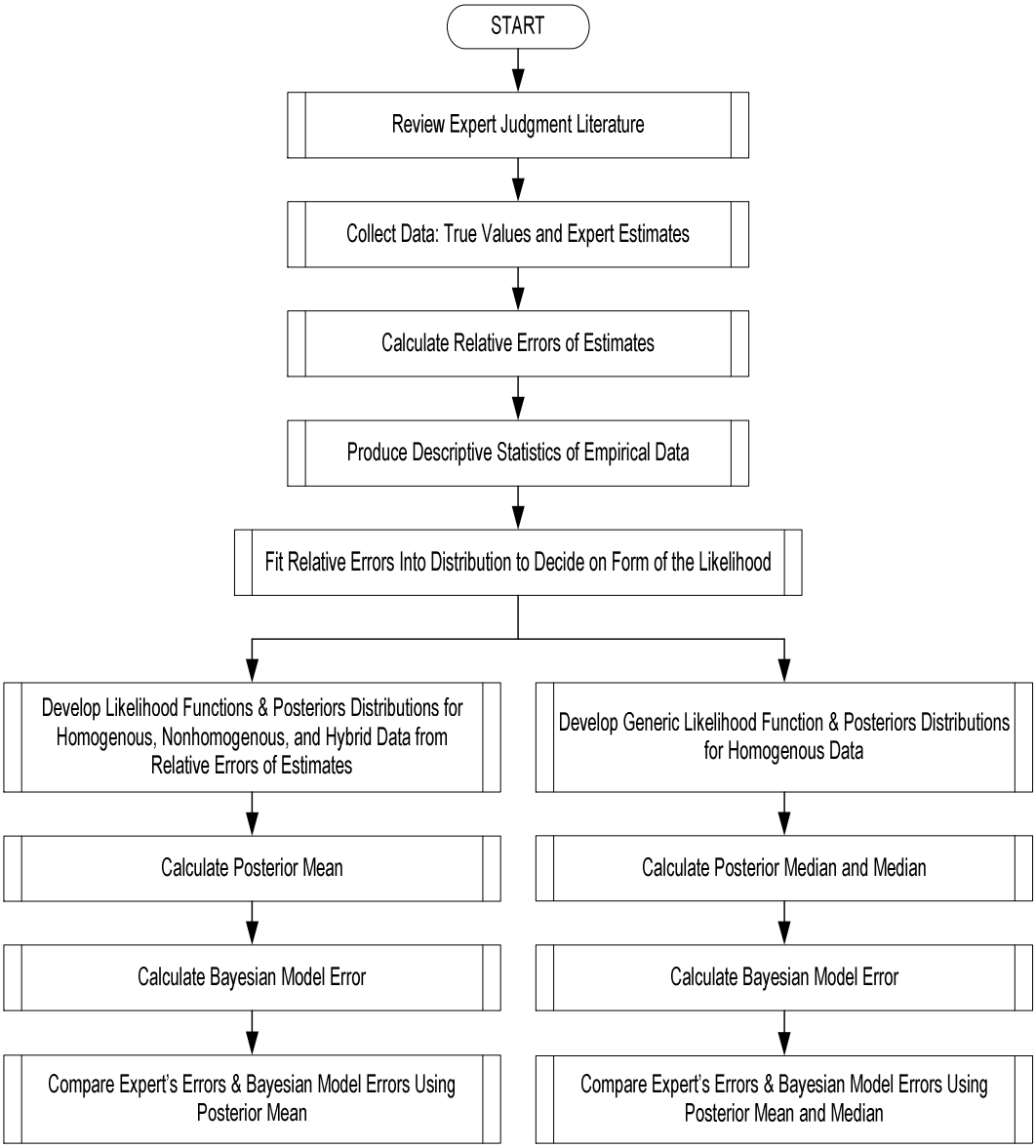


Figure 5. Process Flow of Bayesian Treatment

5.3 Performance Assessment of Case-Specific Likelihood Functions

Assessment in Bayesian framework was performed by 'Uncertainty Modeling' software released and validated by 'The Center for Risk and Reliability (CRR)', Droguette and Mosleh (2003). Evaluation of the data included descriptive data generation and distribution analysis by Mathwave Easyfit™ and MINITAB®.

Table 4 shows empirical data reported in the Benzene concentration case study (case #1) used as an example of a homogenous pool. It is shown that Bayesian treatment of the homogenous data improves 62% of the estimates on average. For nonhomogenous data, an example of study can be found in Table 5 (case #1). For nonhomogenous pool, the percentage of improved estimates increases to 71%. Case #1 is also used for an example of hybrid data. The percentage of improved expert estimates is 71%, as shown in Table 6.

The histogram of relative errors of two homogenous and nonhomogenous cases, Figure 6, shows that over 57% of relative errors of estimates are between (0.5 – 0.8), and about 71% of data points fall between (0.5 – 1.0). The average of relative errors is 1.3 with standard deviation of 0.5. Figure 7 shows best-fitted distributions to all relative errors. Considering the producer risk of 5% ($\alpha = 0.05$), lognormal is among the top three fitted distributions.

Table 4. Bayesian Treatment of Homogenous Pool Using Case-Specific Likelihood Function

True Value	Expert Estimate	Expert Relative Error	Bayesian Mean	Bayesian Mean Relative Error	Error Reduction: + Error Increase: - No change: 0
3.6	3.9	1.083	3.3	0.917	0
3.6	3.2	0.889	3.1	0.861	-
3.6	4.6	1.278	3.3	0.917	+
3.6	7.8	2.167	3.8	1.056	+
3.6	5.8	1.611	4.8	1.333	+
3.6	3.2	0.889	3.1	0.861	-
3.6	3.7	1.028	3.5	0.972	0
7.2	5.5	0.764	5.1	0.708	-
7.2	6.2	0.861	5.4	0.750	-
7.2	6.5	0.903	6.7	0.931	+
7.2	16.2	2.250	8.7	1.208	+
7.2	15.6	2.167	9.5	1.319	+
7.2	11.2	1.556	10.8	1.500	+
7.2	6.0	0.833	7.5	1.042	+
7.5	13.9	1.853	11.5	1.533	+
7.5	7.0	0.933	6.4	0.853	-
7.5	8.6	1.147	6.5	0.867	+
7.5	11.2	1.493	5.8	0.773	+
7.5	21.7	2.893	7.9	1.053	+
7.5	12.1	1.613	8.3	1.107	+
7.5	7.9	1.053	9.2	1.227	-
Average		1.394		1.038	
Standard Deviation		0.587		0.237	
% of Estimates Improved		62% (13 out of 21)			

Table 5. Bayesian Treatment of Non-Homogenous Pool Using Case-Specific Likelihood Function

True Value	Expert Estimate	Expert Relative Error	Bayesian Mean	Bayesian Relative Error	Error Reduction: + Error Increase: - No change: 0
90	115	1.278	111	1.233	+
110	95	0.864	93	0.845	-
90	95	1.056	91	1.011	+
105	110	1.048	93	0.886	-
100	115	1.150	101	1.010	+
115	125	1.087	120	1.043	+
130	145	1.115	134	1.031	+
Average		1.085		1.008	
Standard Deviation		0.125		0.125	
% of Estimates Improved		71% (5 out of 7)			

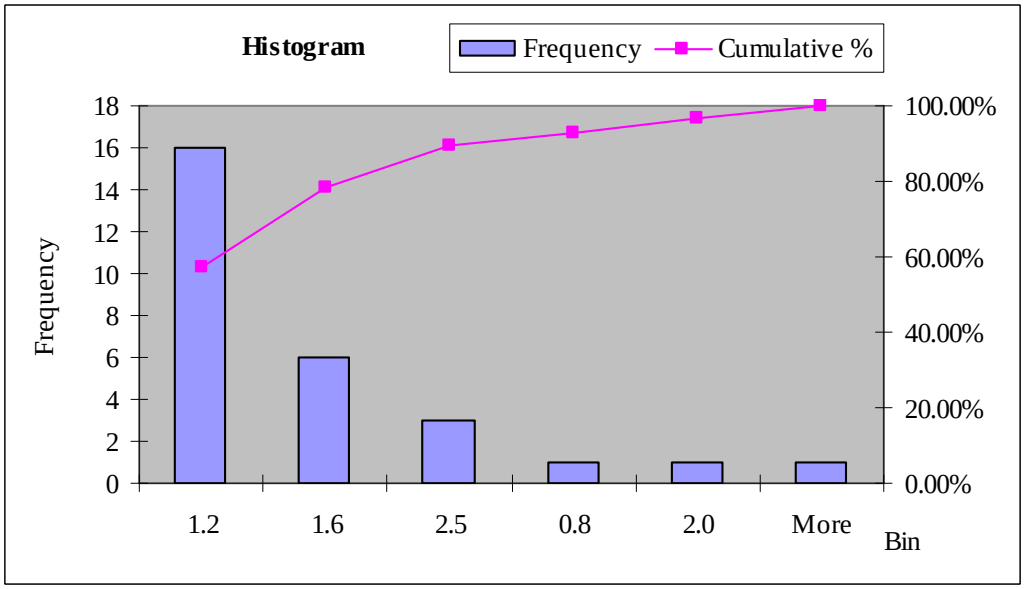
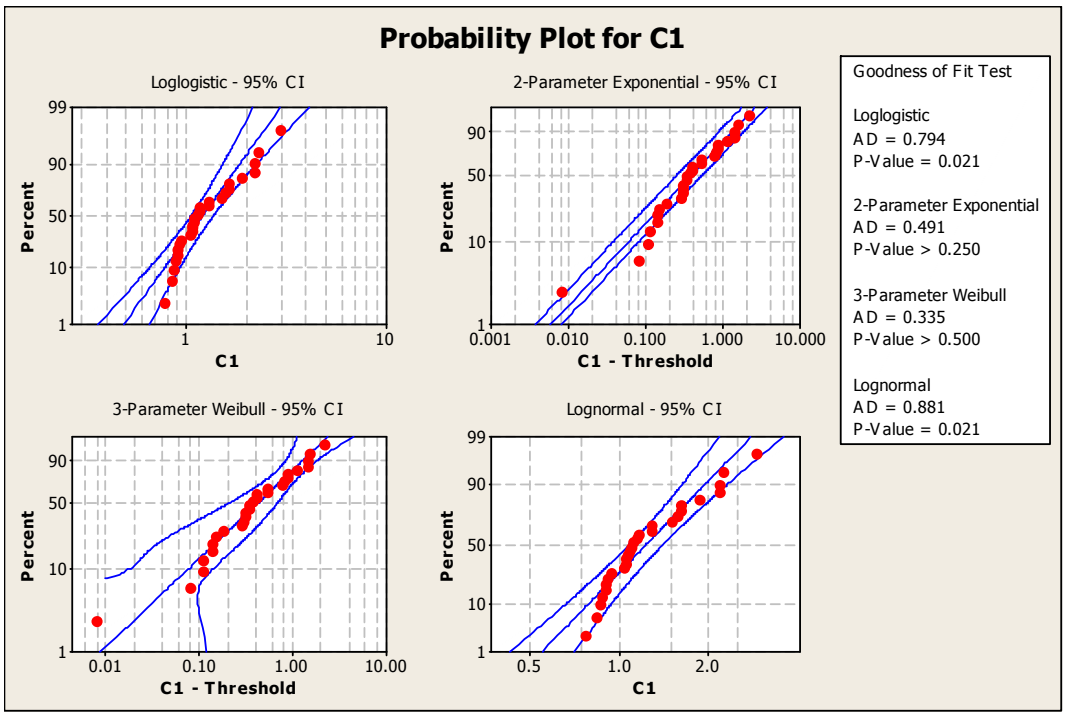


Figure 6. Histogram of Accumulated Homogenous and Nonhomogenous Data

Table 6. Bayesian Treatment of Hybrid Pool Using Case-Specific Likelihood Function

True Value	Expert Estimate	Expert Relative Error	Bayesian Update	Bayesian Relative Error	Error Reduction: + Error Increase: - No change: 0
3.6	3.9	1.083	5.0	1.389	-
3.6	5.8	1.611	4.6	1.278	+
7.5	13.9	1.853	4.9	0.653	+
7.5	21.7	2.893	8.7	1.160	+
7.5	8.6	1.147	7.3	0.973	+
7.2	5.5	0.764	5.4	0.750	-
7.2	11.2	1.556	8.0	1.111	+
Average		1.558		1.045	
Standard Deviation		0.695		0.270	
% of Estimates Improved		71% (5 out of 7)			



Descriptive Statistics (Minitab®)

N	Mean	StDev	Median	Minimum	Maximum
34	1.30	0.50	1.08	0.76	2.89

Figure 7. Distribution Identification for Accumulated Expert Relative Errors

The above demonstrated courses of actions for example case studies are repeated for all empirical expert judgment data (1922 data points) collected. As reflected in Table 7, the study reveals an average of 77% of estimates improved, applying the case-specific homogenous and nonhomogenous likelihood functions. The graphical presentation can be found in Figure 8.

The histogram of expert relative errors depicted in Figure 9 shows that over 45% of relative errors are equal or close to one (expert estimate ~ true value), about 45% of data points fall between (1 – 2) and about 5% falling in the range of (2 – 3). The average relative error is 1.2 and only 5% among all empirical relative errors data are greater than 3.

Table 8 shows the best-fitted probability distributions for relative errors, considering the producer risk of 5% ($\alpha = 0.05$). Lognormal is among the top fitting distributions, since it arises when independent random variables are combined in a multiplicative fashion, as relative error or 'E' is selected for the accuracy measure.

The distribution fitting tests also point to Wakeby and Cauchy distributions as the two first best fits. This fit seems logical since they are also ratio distributions. The random variable associated with ratio distribution comes about as the proportion of two Gaussian distributed variables with zero mean (the Cauchy distribution is also called the normal ratio distribution). The other best fits are Log-Logistic, Burr, and Dagum distributions, which are continuous probability distributions for a nonnegative random variable. The Pearson distribution is a fit since it can visibly contain skewed observations.

Among the above discussed distribution, lognormal seems a better choice for Bayesian models due to ease of use, flexibility to fit many types of data, and wide-spread application in many fields (i.e. environmental application of lognormal distribution, Ashok et al., 1997), and great utility in decision science (Johnson et. al, 2003). Johnson et al. note that some practitioners maintain "that the lognormal distribution is as fundamental as the normal distribution" and that the lognormal distribution has found applications in fields including the physical sciences, life sciences, social sciences, and engineering. He continues, "practitioners find few – if any – tables of its cumulative distribution function available to support their work". Additionally, distribution of the data seems to be positively skewed and for non-negative values, suggesting more reasons to select lognormal distribution as the choice. The lognormal (3P) distribution of expert relative error is depicted in Figure 10.

Table 7. Bayesian Treatment of Non-Homogenous (NH), Homogenous (H) and Hybrid Pools Using Case-Specific Likelihood Function

Case # - H/NH	%Estimates Improved
1 H	62%
2 NH	71%
3 NH	100%
4 NH	100%
5 NH	71%
6 NH	67%
7 NH	67%
8 NH	67%
9 NH	100%
10 NH	71%
11 NH	100%
12 NH	57%
13 NH	71%
14 NH	100%
15 NH	86%
16 NH	100%
17 NH	57%
18 NH	86%
19 H	80%
20 NH	57%
21 NH	86%
22 NH	57%
23 NH	86%
24 NH	86%
25 NH	57%
26 NH	100%
27 NH	57%
28 H (multiple cases)	63%
Average	*77%*
Minimum	*57%*
Maximum	*100%*

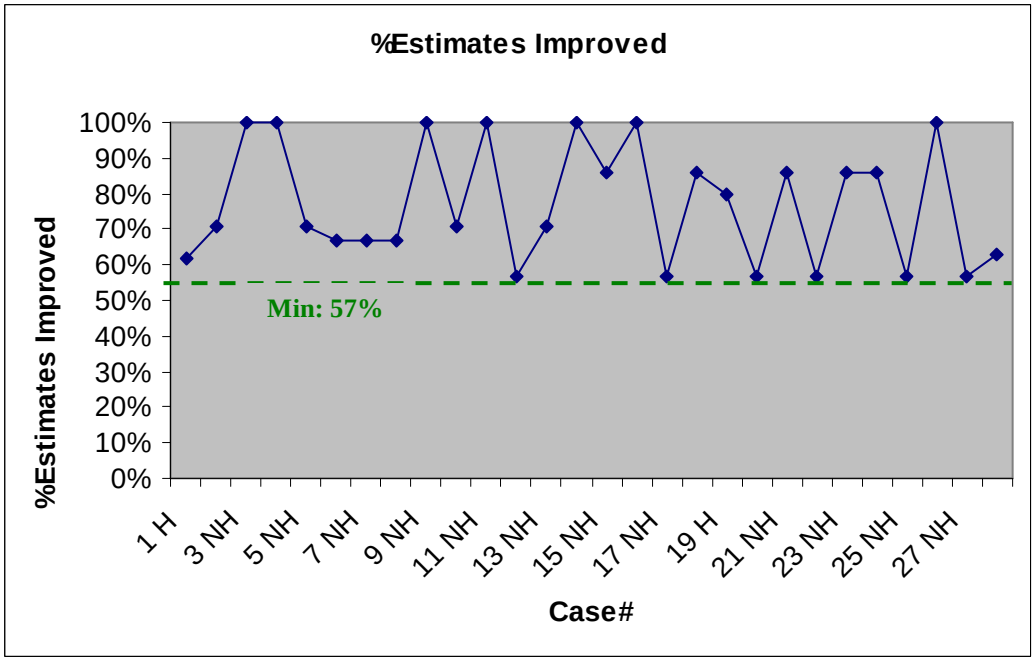
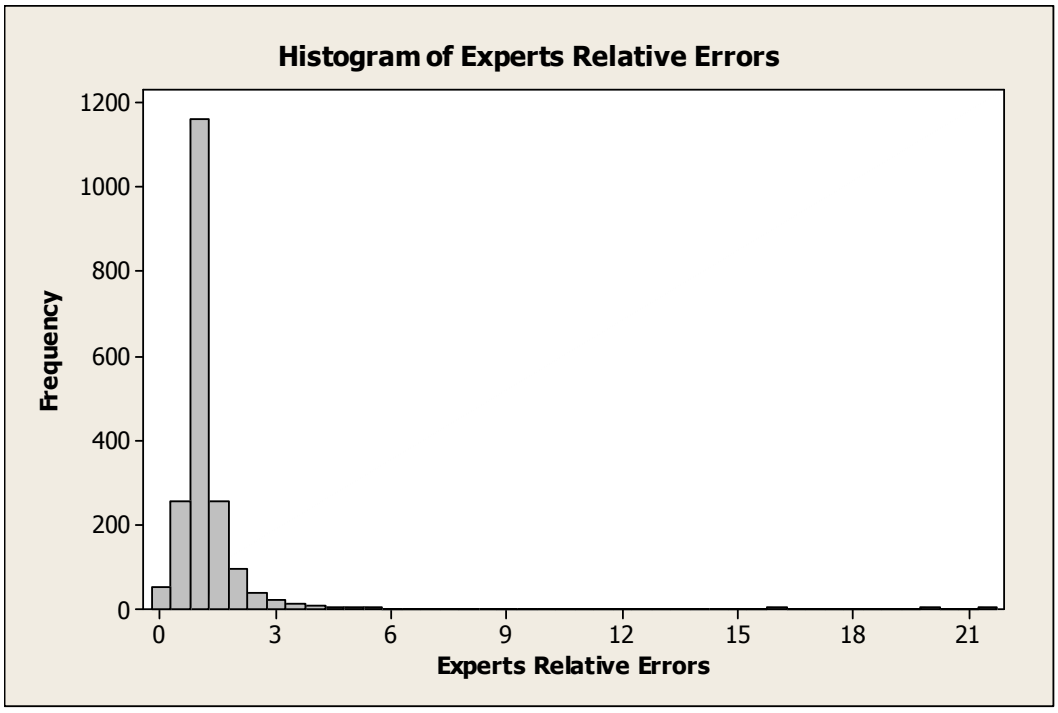


Figure 8. Improvement by Bayesian Treatment in All Empirical Cases



Descriptive Statistics: Relative Error (Minitab®)

Variable	N	Mean	StDev	Median	Min	Max
Relative Error	1922	1.2	1.5	1.0	0.0003	21.3

Figure 9. Histogram of All Relative Errors

Table 8. Best Fitted Distribution for Expert Relative Errors

Best Fitted Distribution (MathWave-EasyFit)	Kolmogorov Smirnov	Anderson Darling	Chi-Squared
	Rank	Rank	Rank
Wakeby	1	1	1
Cauchy	2	2	2
Dagum (4P)	3	5	5
Log-Logistic (3P)	4	4	4
Burr (4P)	5	3	3
Burr	6	7	7
Dagum	7	6	6
Pearson 6 (4P)	8	8	8
Lognormal (3P)	9	9	9

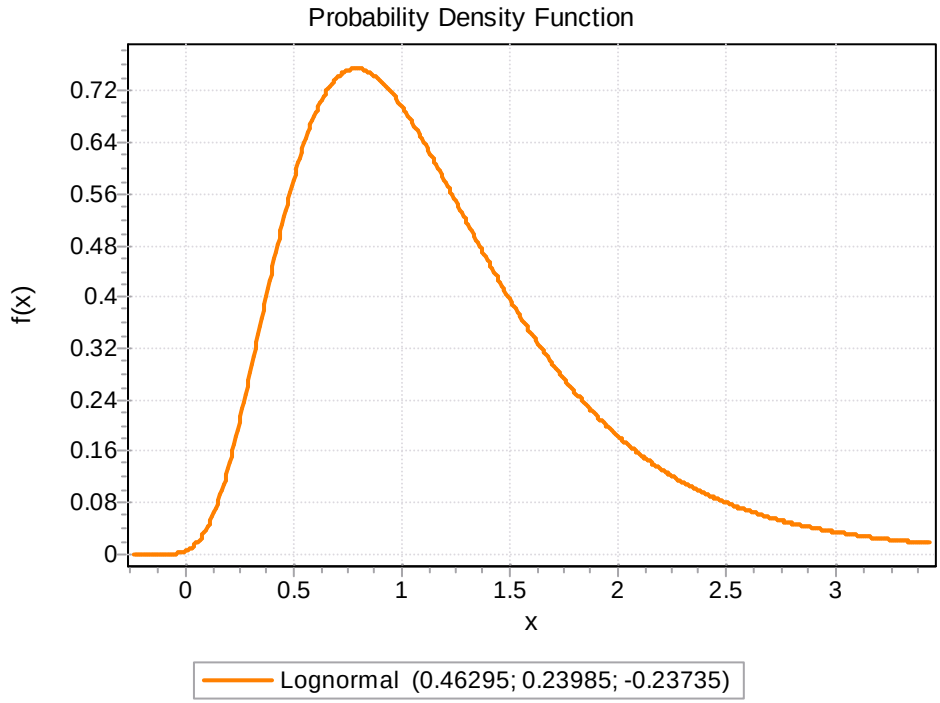


Figure 10. Lognormal (3P) Distribution of for All Relative Errors

5.4 Performance Assessment of Generic Likelihood Functions

The entire process of updating estimates was also repeated, using a generic likelihood function. If '*E*' is a lognormally distributed, its expected value is:

$$f(E;\mu,\sigma)=\frac{1}{E\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln E-\mu}{\sigma}\right)^2} \qquad \text{Equation 30}$$

From Figure 10:

$\mu = 0.24$

$\sigma = 0.46$

$$\mu=\ln E_{50} \qquad \text{Equation 31}$$

Median: $E_{50}=e^{\mu}=e^{0.24}=1.27$ Equation 32

The above parameters are prior and should be updated using hybrid formulations. From Equation 13:

$$L(u'|u,\underline{\theta})=\frac{1}{\sqrt{2\pi}\sigma_E u'}e^{-\frac{1}{2}\left(\frac{\ln u'-\ln u-\ln E_{50}}{\sigma_E}\right)^2}\Rightarrow$$

$$L(u'|u)=\frac{1}{\sqrt{2\pi}(0.46)u'}e^{-\frac{1}{2}\left(\frac{\ln u'-\ln u-0.24}{0.46}\right)^2} \qquad \text{Equation 33}$$

Representation of independent expert estimates by $(\overline{u'})=(u_1'...u_n')$:

$$L(\overline{u'}|u)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}(0.46)u'}e^{-\frac{1}{2}\left(\frac{\ln u'-\ln u-0.24}{0.46}\right)^2} \qquad \text{Equation 34}$$

$$\pi(u|\overline{u'})=\frac{e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'-\ln u-0.24}{0.46}\right)^2}\pi_0(u)}{\int e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'-\ln u-0.24}{0.46}\right)^2}\pi_0(u)du} \qquad \text{Equation 35}$$

It can be shown that the parameters of the above posterior distribution can be calculated as (Mosleh, 1981):

$$u_{50} = \prod_{i=1}^{n} \left(\frac{u_i'}{E_{50}} \right)^{\frac{1}{n}} \quad \text{Equation 36}$$

$$\sigma^2 = \frac{{\sigma_E}^2}{n} \quad \text{Equation 37}$$

$$\mu = u_{50} \times e^{\frac{\sigma^2}{2}} \quad \text{Equation 38}$$

u_{50}: is the posterior median

σ: is the posterior standard deviation

n: is the number of estimates or experts

μ: is the posterior mean

Mean and the median of the posterior are compared with the true value to explore whether the formulated likelihood distribution is able to reduce the expert error. An example is presented in Table 9 and Table 10 using TU Delft data (case 28). The complete study for all data tested reveals an overall improvement in the accuracy of expert, applying the formulated generic likelihood function considering available case-independent evidence.

Table 9. Numerical Example to Measure Performance of Generic Likelihood Function Using Mean (μ) of Posterior

u'	u	Other Available Experts Estimates on True Value	$\mu = u_{50} \text{Exp}(\sigma^2/2)$	u'/u	μ/u	Error Reduced: + Error Increased: -
0.019	0.027	(0.05, 0.02, 0.02, 0.035)	0.023	0.704	0.866	+
0.05	0.027	(0.019, 0.02, 0.02, 0.035)	0.018	1.852	0.680	+
0.02	0.027	(0.019, 0.05, 0.02, 0.035)	0.023	0.741	0.855	+
0.02	0.027	(0.019, 0.05, 0.02, 0.035)	0.023	0.741	0.855	+
0.035	0.027	(0.019,0.05, 0.02, 0.02)	0.020	1.296	0.743	+

Table 10. Numerical Example to Measure Performance of Generic Likelihood Function Using Median (u_{50}) of Posterior

u'	u	Other Available Experts Estimates on True Value	$u_{50}=\prod(u'/E_{50})^{1/n}$	u'/u	u_{50}/u	Error Reduced: + Error Increased: -
0.019	0.027	(0.05, 0.02, 0.02, 0.035)	0.023	0.704	0.844	+
0.05	0.027	(0.019, 0.02, 0.02, 0.035)	0.018	1.852	0.662	+
0.02	0.027	(0.019, 0.05, 0.02, 0.035)	0.022	0.741	0.833	+
0.02	0.027	(0.019, 0.05, 0.02, 0.035)	0.022	0.741	0.833	+
0.035	0.027	(0.019,0.05, 0.02, 0.02)	0.020	1.296	0.724	+

5.5 Conclusion

The questions answered include empirical assessment of expert errors and to explore whether the use of formulated likelihood functions would reduce future prediction errors. The empirical assessment of data revealed that approximately:

1. 45% of errors were close to one (expert estimate ~ true value)
2. 45% of data points were between (1 – 2)
3. 5% of relative errors were falling in the range of (2 – 3)
4. 5% among all empirical errors data was greater than 3
5. Lognormal was identified as one of the best fitted distributions
6. The average error was 1.2
7. The standard deviation was 1.5

Applying the case-specific likelihood function developed by relative errors showed:

- 77% of estimates improved

Application of generic likelihood function using the posterior mean and case-independent evidence revealed:

- 50% of estimates improved

Application of generic likelihood function using the posterior median and case-independent evidence showed:

- 52% of estimates improved

Results confirm that the developed generic likelihood function, in conjunction with available evidence, is able to update at least half of the estimates.

Chapter 6: Data-Informed Aggregation of Expert Opinions

6.1 Introduction

In uncertain situation, combining data can reduce error (Armstrong, 2001). Speculations made about the correlation between accuracy of expert estimates and the number of experts elicited, have led many to conclude that the more experts are elicited, the higher accuracy of estimates can be reached. This may seem similar to increasing the sample size in an experiment. Ashton and Ashton (1985) studied judgmental forecasts of the number of advertising pages in Time magazine. The conclusion was that by combining the forecasts of four experts, error of estimates is reduced by 3.5%. Batchelor and Dua (1995) showed increase in accuracy from 10 to 22 economists. Their study also revealed a small improvement from 22 to the remaining 12.

The two well-established mathematical approaches to aggregate opinions are axiomatic and Bayesian models (Boring, 2007; Clemen and Winkler, 1997). The first formal framework of the Bayesian methods for use of expert opinion was presented by Morris (1974, 1977). French (1985), Lindley (1985), and Genest and Zidek (1986) all conclude that a Bayesian updating scheme is the most appropriate method when a group of experts provide information for a decision maker. A comprehensive review of aggregation literature, including dependence, can be found in French (1985), Ouchi (2004), Genest and Zidek (1986), French and Ríos Insua (2000).

The objectives of this part of the research are to:

1. Investigate whether mathematical aggregation of expert opinions reduce the error of aggregated estimate
2. Assess the correlation between the number of experts and the accuracy of estimates through Bayesian aggregation.

The above questions are addressed, using empirical data in the Bayesian framework, applying likelihood distribution formulated in Chapter 4 by considering:

- Case-specific likelihood function
- Generic likelihood function

In this chapter, mathematical formulas for generic aggregation are presented. Using empirical data, aggregation performance and number of experts for optimum accuracy are determined in each method based on the results obtained.

6.2 Mathematical Model

Expert opinions are aggregated in the Bayesian framework using the likelihood function formulated by relative error of estimates as well as generic likelihood function developed. Postulating independent experts with lognormal likelihood distributions with parmaters (μ_i, σ_i) we have (see Chapter 4):

$$L(\overline{u'}|u)=\prod_{i=1}^{n}L(u'_i|u)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma u'_i}e^{-\frac{1}{2}\left(\frac{\ln u_i-\ln \mu_i}{\sigma_i}\right)^2} \qquad \text{Equation 39}$$

$\ln \mu_i = \ln u_i - \ln E_{50}$

u'_i : is the i^{th} expert estimate

u : is the unknown of interest

$\overline{u'}$: set of expert estimates

Expanding the above equation and rearranging the terms as a function of '*u*':

$$L(\overline{u'}|u)=\frac{1}{\left(\sqrt{2\pi}\right)^n\sigma^n\prod_{i=1}^{n}u'_i}e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'_i-\ln \mu_i}{\sigma_i}\right)^2} \qquad \text{Equation 40}$$

Using the above likelihood in Bayes' theorem the posterior distribution of the unknown of interest given set of errors:

$$\pi\left(u\left|\overline{u'}\right.\right)=\frac{e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'_i-\ln \mu_i}{\sigma_i}\right)^2}\pi_0(u)}{\int e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'_i-\ln \mu_i}{\sigma_i}\right)^2}\pi_0(u)du} \qquad \text{Equation 41}$$

For the generic likelihood formulated by relative errors in Chapter 5, (E_{50} = 1.2 and σ_E = 0.69). The assumption is that E_{50} and σ_E are the same for all experts. Therefore for expert '*i*':

$$L(u'_i|u)=\frac{1}{\sqrt{2\pi}\sigma_E u'_i}e^{-\frac{1}{2}\left(\frac{\ln u'_i-\ln u-\ln E_{50}}{\sigma_E}\right)^2} \qquad \text{Equation 42}$$

Postulating independence among experts, as in equation 31:

$$L\left(\overline{u'}\middle|u\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_E u'_i}e^{-\frac{1}{2}\left(\frac{\ln u'_i-\ln u-\ln E_{50}}{\sigma_E}\right)^2} \qquad \text{Equation 43}$$

$$L\left(\overline{u'}\middle|u\right)=\frac{1}{\left(\sqrt{2\pi}\right)^n {\sigma_E}^n\prod_{i=1}^{n}u'_i}e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'_i-\ln u-\ln E_{50}}{\sigma_E}\right)^2} \qquad \text{Equation 44}$$

This results in posterior distribution for this case:

$$\pi\left(u\middle|\overline{u'}\right)=\frac{e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'_i-\ln u-\ln E_{50}}{\sigma_E}\right)^2}\pi_0(u)}{\int_u e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\ln u'_i-\ln u-\ln E_{50}}{\sigma_E}\right)^2}\pi_0(u)} \qquad \text{Equation 45}$$

It can be shown that the median of this posterior distribution can be calculated as:

$$u_{50}=\prod_{i=1}^{n}\left(\frac{u'_i}{E_{50}}\right)^{\frac{1}{n}} \qquad \text{Equation 46}$$

$$\sigma^2=\frac{{\sigma_E}^2}{n} \qquad \text{Equation 47}$$

The mean of the posterior is

$$\mu=u_{50}\times e^{\frac{\sigma^2}{2}} \qquad \text{Equation 48}$$

The relative error of aggregated value (mean and median of posterior), $E_{Aggregate}=\frac{u'_{Aggregate}}{u}$, is compared with the expert's relative error, $E_i=\frac{u'_i}{u}$.

The number of estimates improved is monitored as the number of experts (n) increases in order to uncover whether this boost reduces the overall error of estimates and to unveil the minimum number of experts needed to obtain maximum accuracy.

6.3 Aggregation by Simulation

In this section, simulation-based performance assessment of Bayesian and representative models of axiomatic aggregation of point estimates is conducted. In addition, the impact of the number of experts on Bayesian aggregation performance is assessed through replication.

Simulation is carried out considering both cases of independence and dependence (for Bayesian method) among experts. The simulation process flow is depicted in Figure 11.

There are two loops constructed. Model inputs and random true values are produced in the first loop using sampling of lognormal distributions. In the second loop, expert estimates are generated within the same data range and aggregation is performed. The process is repeated in each of the loops for the calculated number of iteration.

The simulation loop iteration is calculated based on the formula proposed by Winston (2001):

$$m = \frac{4\left(z_{\alpha/2}\right)^2 \sigma^2}{D^2} \qquad \text{Equation 49}$$

In this formula,

m is the number of iterations needed,

σ is the estimated standard deviation of the output, and

D is the desired width of the confidence interval. Simulation is first run with just 100 iterations ($\alpha = 0.05$, and therefore $z_{\alpha/2}=1.96$) to obtain an estimate for the standard deviation. The number of iterations can then be calculated using the same formula and calculated standard deviation.

The selected axiomatic aggregation methods are arithmetic weighted sum and weighted geometric mean:

I. Arithmetic Unweighted Sum: this is just an unweighted linear combination of 'n' expert estimates (u': expert estimates).

$$u'_{\text{Aggeragte}} = \frac{1}{n}\sum_{i=1}^{n} u'_i \qquad \text{Equation 50}$$

II. Unweighted Geometric Mean: An unweighted geometric mean is obtained as the product of the estimates raised to the power equal to one over the number of estimates (n) (u': expert estimates).

$$u'_{\text{Aggeragte}} = \left(\prod_{i=1}^{n} u'_i\right)^{\frac{1}{n}} \qquad \text{Equation 51}$$

For the Bayesian aggregation simulation, posterior distribution is formed. The mean of the posterior, as the aggregated estimate and expert estimates are compared with the true value (refer to Chapter 4) imported from the first loop.

To address dependency among experts, choices of copulas are used for likelihood functions as listed in the following. The basis of applying a copula distribution is that a copula-based model is constructed by joining the copula function with the marginal distributions. According to Sklar's Theorem (1959), given a joint cumulative distribution function $F(x_1, \ldots, x_n)$ for random variables $(x_1 \ldots x_n)$ with marginal cumulative distribution $F_1(x_1) \ldots F_n(x_n)$, F can be written as a function of its marginal distributions:

$$F(x_1 \ldots x_n) = c[F_1(x)_1 \ldots F_n(x)_n] \qquad \text{Equation 52}$$

The function 'c' is called a copula. This means that the joint density $f(x_1 \ldots x_n)$ can be written as:

$$f(x_1 \ldots x_n) = f_1(x_1) \ldots f_n(x_n) c[F_1(x)_1 \ldots F_n(x)_n] \qquad \text{Equation 53}$$

It is clear that the above copula density '*c*' captures information about the dependence among the *X*s, and therefore, it is called a dependence function. There are many families of copulas which typically have several parameters related to the strength and form of the dependence. More discussion and properties of these selected copula functions can be found in Clayton (1978), Frank (1979), Gumbel (1960), Hougaard (2000), Silva and Lopez (2008). The selected families of copulas are:

1. Gaussian – Multivariate normal copula: this copula captures dependence like the multivariate normal distribution, by using only pair wise correlations among the variables. It accomplishes the task for variables with arbitrary marginal distributions. Moreover, the normal copula permits the use of any positive-definite correlation matrix, meaning that it is not limited to intra class correlation matrices.
2. Archimedean
    - 2.3 Frank: Frank can be used to capture positive dependence among random variables.
    - 2.4 Clayton: In the Clayton copula, the random variables are statistically independent.
    - 2.5 Gumbel: The Gumbel copula is asymmetric, with more weight in the right tail.

The simulation process for dependent experts is depicted in Figure 12. The simulation is executed using MATLAB® software produced by 'The Mathworks™'. MATLAB® is a technical computing language for algorithm development and numerical computation.

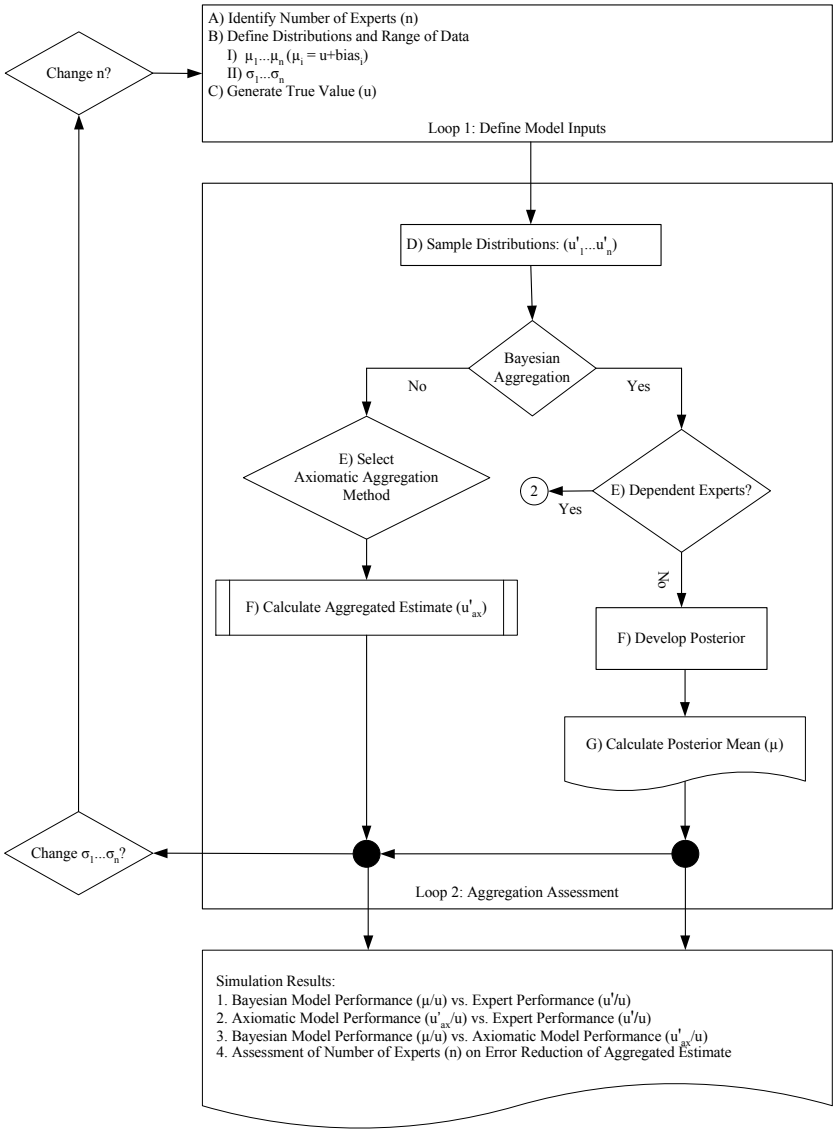


Figure 11. Aggregation Simulation Approach

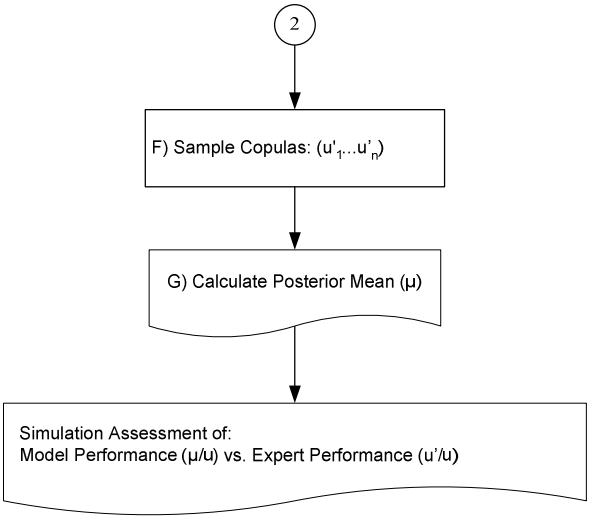


Figure 12. Aggregation Simulation for Dependent Experts

6.4 Simulation Results

6.4.1 Aggregation Performance

Simulation shows that Bayesian aggregation method results in less aggregation error than Axiomatic procedures, as depicted in Figure 13. In this graph, the x-axis is the number of experiments or cases simulated, unique to their generated inputs in both loops. The y-axis is the relative errors or $E = \frac{u'}{u}$, where *u*' is the estimate and *u* is the true value. The spikes which can be noted in the graphs show selection of high standard deviations (low expert expertise), which clearly reveals that the decrease of expertise increases the error.

6.4.2 Dependent Experts Performance

For dependent experts, Gaussian, Frank, Clayton and Gumbel copula families are used and minimum improvement among these choices are reported. Model error shows about 80% overall reduction in error of aggregated estimate compared to the mean of all expert errors with correlation of 0.25, about 75% with correlation of 0.50, and finally about 70% with correlation of 0.75. This means that the more independent experts are; the more accurate aggregated estimate becomes, however, the amount of improvement is not significant.

6.4.3 Size of Expert Panel

The simulation reveals that there is not a strong correlation between accuracy of aggregated estimate and the number of experts. As depicted in the Figure 14, about 50% of estimates are improved by increasing the number of experts to two. It seems that selecting more than two experts can lead to more improved estimates (over 60%). However, it can be noted that from 3 to 10 experts, the percentage of improved estimates is not noteworthy (less than 10%).

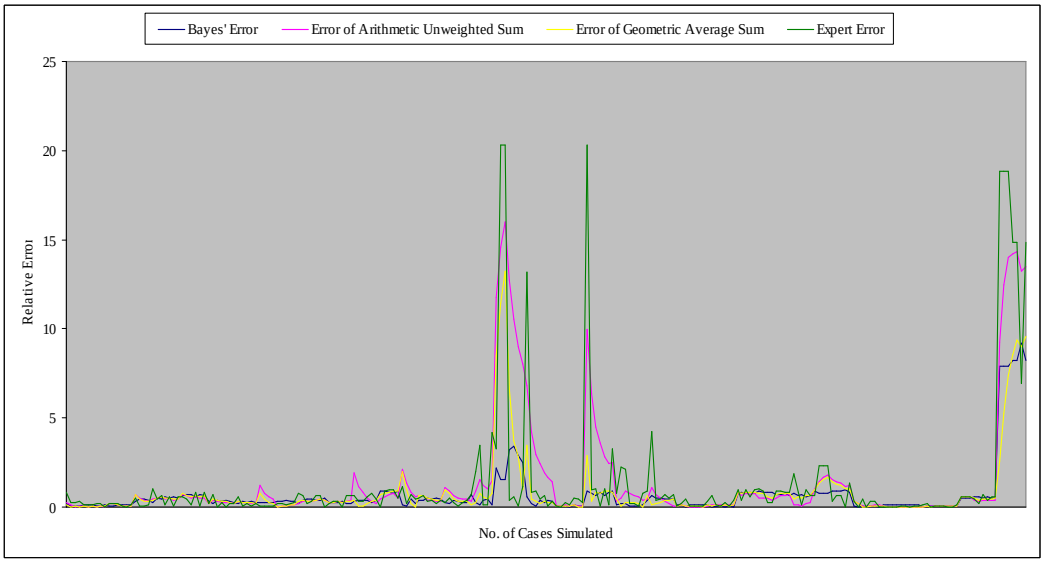


Figure 13. Performance of Aggregation Methods

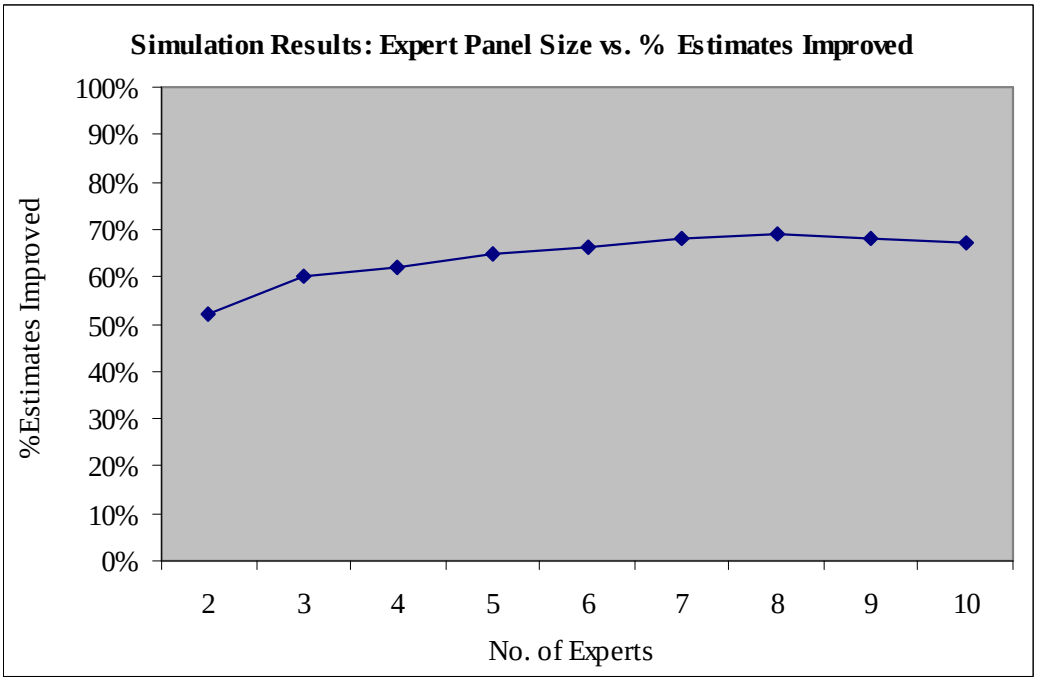


Figure 14. Simulation Results: Expert Panel Size vs. % Estimates Improved

6.5 Aggregation Using Empirical Data

In this section aggregation is performed using empirical data to assess:

(a) The Bayesian aggregation performance and

(b) Impact of number of experts on aggregation in a real world application.

Expert opinions are combined in a Bayesian framework using likelihood function formulated by relative error of estimates, considering independence among experts. The relative error of aggregate is compared with the expert error. This procedure is illustrated using sample data, reflected in Table 11.

Table 11. Numerical Example for Aggregation Procedure Illustration

Expert ID	Estimate	True Value
A	0.019	0.027
B	0.05	0.027
C	0.02	0.027
D	0.02	0.027
E	0.035	0.027

The available evidence to update the estimate of expert B is the relative error of expert A. The posterior is developed, and the average of this distribution is calculated as the Bayesian update. The result can be found in Table 12.

Table 12. Bayesian Update and Relative Error for Aggregation Example

Expert ID	Estimate	True Value	Bayesian Update	Expert Relative Error	Bayesian Relative Error
A	0.019	0.027		0.704	
B	0.05	0.027	0.06	1.852	2.222

In the next step, the first two expert relative errors are considered available evidence to update the estimate of expert C, as reflected in Table 13. The aggregated estimate is compared with all three expert estimates, revealing reduction in error. From the results obtained for this set of data, going from one to two experts increases the error for both estimates released by expert A and B. However, increasing the number of experts from two to three reduces the error for all experts A, B, and C.

Table 13. Continuation of Aggregation Example

Estimate	True Value	Bayesian Update	Expert Relative Error	Bayesian Relative Error	% Reduced Error (+) % Increased Error (-)	% Reduced Error (+) % Increased Error (-)	Expert ID
0.019	0.027		0.704		-	+	A
0.05	0.027	0.06	1.852	2.222	-	+	B
0.02	0.027	0.023	0.741	0.852		+	C

This process is continued to include all experts in the data set, as shown in Table 14.

Table 14. Aggregation Results for Example Data

Estimate	True Value	Bayesian Update	Expert Relative Error	Bayesian Relative Error
0.019	0.027		0.704	
0.05	0.027	0.06	1.852	2.222
0.02	0.027	0.023	0.741	0.852
0.02	0.027	0.025	0.741	0.926
0.035	0.027	0.039	1.296	1.444

Aggregate ID		A & B	A, B & C	A, B, C & D	A, B, C, D & E
ID	Expert Relative Error	Bayesian Aggregate Relative Error			
		2.222	0.852	0.926	1.444
A	0.704	-	+	+	-
B	1.852	-	+	+	+
C	0.741		+	+	-
D	0.741			+	-
E	1.296				-
Estimates Improved		0	3	4	1
Total (Expert Panel Size)		2	3	4	5

To treat the data completely random, another step is taken where a sample of 10% of the data sets are used for calculations out-of-reported order. Rearranging the raw data in previous example (Table 11) is shown in Table 15.

Table 15. Example for Aggregation Procedure: Out-of-order data

Expert ID	Estimate	True Value
A	0.019	0.027
D	0.02	0.027
C	0.02	0.027
B	0.05	0.027
E	0.035	0.027

The same process as described before in the example is repeated for this random set and results are shown in Table 16.

Table 16. Aggregation Results for Example: Out-of-order data

Estimate	True Value	Bayesian Mean	Expert Relative Error	Bayesian Relative Error
0.019	0.027		0.704	
0.02	0.027	0.029	0.741	1.074
0.02	0.027	0.024	0.741	0.889
0.05	0.027	0.040	1.296	1.481
0.035	0.027	0.069	1.852	2.556

Aggregate ID		A & D	A, D & C	A, D, C & B	A, D, C, B & E
ID	Expert Relative Error	Bayesian Aggregate Relative Error			
		1.074	0.889	1.481	2.556
A	0.704	+	+	-	-
D	0.741	+	+	-	-
C	0.741		+	-	-
B	1.296			-	-
E	1.852				-
Estimates Improved		2	3	0	0
Total (Expert Panel Size)		2	3	4	5

These calculation steps are executed for empirical data sets. The number of improved estimates is monitored as the number of experts increase for estimates involving 2 to 10 experts to:

3. Investigate whether mathematical aggregation of expert opinions reduce the error of aggregated estimate
4. Assess the correlation between the number of experts and the accuracy of estimates through Bayesian aggregation.

Bayesian calculations were performed by 'Uncertainty Modeling' software released and validated by The Center for Risk and Reliability (CRR), Droguette and Mosleh (2003).

The improvements (reduction in error) in all data sets per expert panel size are listed in Table 17 using case-specific likelihood. Additionally, the correlation between percentages of error reduction with the increase of the number of experts is investigated using best-fitted line, as depicted in Figure 15. The best fitted line reveals a positive correlation, but with a moderate adjusted coefficient of determination (R^2 = 63%). The computation was also performed using mean and median of the generic likelihood function, summarized in Table 18 and Table 19.

Table 17. Aggregation Performance: Case-Specific Likelihood

Expert Panel Size	Total Data	No. of Estimates Improved	% of Estimates Improved
2	98	52	53%
3	147	91	62%
4	184	109	59%
5	225	144	64%
6	240	160	67%
7	259	193	75%
8	152	100	66%
9	72	51	71%
10	60	42	70%

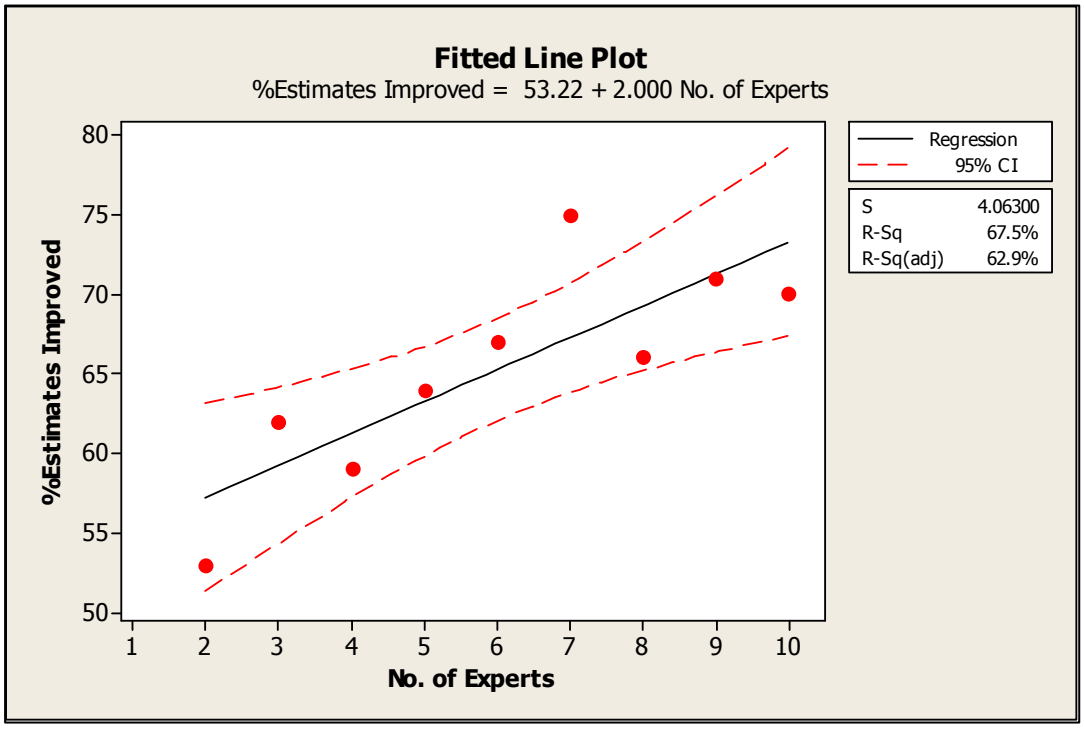


Figure 15. Fitted Line Plot: Improvement vs. Experts Panel Size

Table 18. Aggregation Performance Summary: Generic Likelihood – Mean

Expert Panel Size	Total Data	No. of Estimates Improved	% of Estimates Improved
2	70	25	53%
3	105	50	59%
4	140	76	60%
5	175	92	58%
6	192	108	52%
7	203	115	51%
8	128	70	55%
9	45	16	44%
10	40	12	43%

Table 19. Aggregation Performance Summary: Generic Likelihood – Median

Expert Panel Size	Total Data	No. of Estimates Improved	% of Estimates Improved
2	70	41	46%
3	105	62	52%
4	140	87	57%
5	175	111	54%
6	192	113	49%
7	203	118	46%
8	128	100	66%
9	45	26	42%
10	40	24	38%

6.5.1 Aggregation Performance

Bayesian aggregation resulted in less relative error on average. Application of the likelihood function developed by relative errors revealed on average 65% of estimates improved. Application of generic likelihood for homogenous data using posterior mean revealed on average 53% of estimates improved. Application of generic likelihood for homogenous data using posterior median showed on average 50% of estimates improved.

6.5.2 Expert Panel Size

Best-fitted line graphs for case-specific events, Figure 15, reveal that increasing the number of experts is positively correlated with the accuracy of aggregated estimate. The moderate coefficient of determination (R^2 = 63%) suggests that this association is not very strong. It seems that eliciting two experts (instead of one) can lead to reduction in error for more than 50% of estimates. It can be seen that increasing the number of experts from two to three, reduces the error for approximately 60% of estimates. However, from 3 to 10 experts, the percentage of improved estimates is not significant.

Chapter 7: Summary of Results

7.1 Research Contribution

This research contributes to the body of knowledge of expert judgment. In contrast to many studies revealing shortcoming in the expert judgment, this research reports how well experts are able to make a prediction in real world. This task was carried out by data-informed calibration and aggregation of experts in the Bayesian framework.

A generic likelihood was developed, which showed the ability to update the expert estimates. Additionally, specific likelihood distributions for homogenous, nonhomogenous and mixed data were formulated using expert relative errors of estimates, revealing that formulated likelihood functions can reduce future prediction errors.

To study the impact of number of experts on the accuracy of aggregated estimate collected expert judgments were combined in a Bayesian framework using likelihood distributions developed in the first part of the research study. Total number of estimates with reduced errors was depicted against corresponding expert panel size. The objective achieved was the determination of the correlation between the number of experts and the accuracy of the combined estimate to recommend an expert panel size. The result of the study showed weak to moderate correlation between the expert panel size and the accuracy of aggregate. It was noted that eliciting two experts (instead of one) could lead to reduction in relative error of estimates.

7.2 Data-Informed Calibration of Expert Judgment

The objective of this section was empirical assessment of expert judgment in different disciplines as well as feasibility and value of data-driven expert calibration within the Bayesian formalism.

The result of the conducted study revealed:

1. 45% of errors are close to one (expert estimate ~ true value)
2. 45% of data points are between (1 – 2)
3. 5% of relative errors are falling in the range of (2 – 3)
4. 5% among all empirical errors data are greater than 3
5. Lognormal is identified as one of the best fitted distributions
6. The average relative error is 1.2 with standard deviation of 1.5

Applying the case-specific likelihood function developed by relative error for homogenous and nonhomogenous cases showed:

- 77% of estimates improved

Application of generic likelihood function using posterior mean, considering the existing evidence revealed:

- 57% of estimates improved

Application of generic likelihood using posterior median, considering the existing evidence showed:

- 52% of estimates improved

7.3 Data-Informed Aggregation of Expert Judgment

The objective of this section was:

1. to determine if mathematical aggregation reduces the error of aggregate,
2. to explore the correlation between the number of experts and accuracy of aggregated estimate in order to recommend an expert panel size

Figure 16 gives a quick overview of the results obtained:

1. Mathematical aggregation reduces the error of estimate.
2. The accuracy of the aggregate increases by adding to number of experts.
3. The optimum expert panel size is 3, if the improvement of 50-60% of estimates is satisfactory.

Overall, the decision of eliciting more experts can be properly made, considering governing circumstances of the case on-hand. If possible, the panel should be large enough to capture complementary expertise and achieve diversity of opinion, to ensure a balanced and broad spectrum of viewpoints, expertise, and technical points of view. The decision-maker should assess if targeted improvement pays off the cost of hiring more experts.

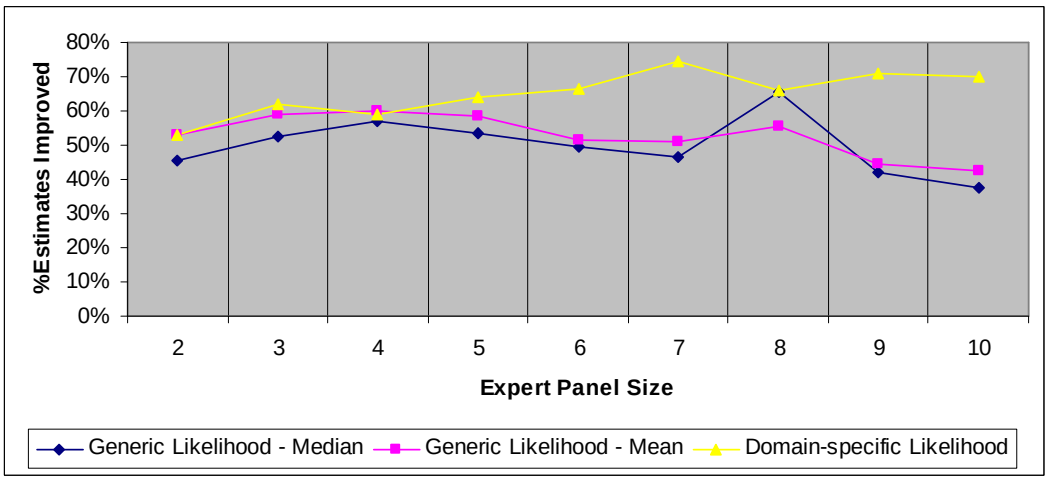


Figure 16. Bayesian Treatment vs. Expert Panel Size

7.4 Research Limitations

Reader should be aware of the limitations and restrictions encountered in conducting this research to have a complete picture of this study:

- Besides expressing their subjective judgments directly, experts in this study could use prototypes, models, destructive and nondestructive tests (among other tools) to gather data, gain practical knowledge to estimate the unknown.
- This study only focused on expert point estimates in discrete or continuous forms.
- The estimates provided by forecasting models were considered 'expert data'. This was because of expert input into construction of the model, or expert review and adjustment of the output.
- Experts were considered independent in model development and numerical calculations.
- Inconsistencies among experts were accepted as inherent variation in modeling and assessment processes. Inherent variation could help to capture real-world error causes (though the sources of these fallacies remain unknown) and examine the formulated likelihood functions in dealing with these variations.
- The focus of this research was on mathematical procedures for calibration and aggregation of expert point estimates. Specially, Bayesian method was the central point of the study.

7.5 Future Research

There are many factors which can impact the accuracy of expert judgment such as expert attributes, elicitation and aggregation methods and so on. This research focused on calibration and aggregation of expert judgment in a Bayesian framework, considering independent experts.

Dependency is a major factor affecting the quality of judgment. Future research using methods presented in this research should consider the case of dependence in both calibration and aggregation procedures and address the pertinent issues.

Additionally, the empirical data available for collection allowed this research to only consider up to 10 experts. If data is available, the study should continue to larger expert panel size, and perhaps, determine the variations seen in the % of improved estimates as number of expert increases.

References

[1] Abdullah Khan, H. T. A Comparative Analysis of the Accuracy of the United Nations' Population Projections for Six Southeast Asian Countries. Interim Report, IR-03-015.

[2] Adler, M. & Ziglio, E. (1996). Gazing into the oracle. JK Publishers: Bristol, PA.

[3] Alpert, M. & Raiffa, H. (1982). A progress report on the training of probability assessors, in Judgement Under Uncertainty: Heuristics and Biases, eds. D. Kahneman, P. Slovic & A. Tversky, Cambridge University Press, New York, 294–305.

[4] Anderson, U. & Wright, W. F. (1988). Expertise and the explanation effect. Organization Behavior and Human Decision Processes, 13, 431-446.

[5] Apostolakis, B. E. (1990). Interfuel and energy-capital complementarity in manufacturing industries. Applied Energy 35, 2, 83-107.

[6] Armstrong, J. S. (2001). Chapter 13 in Armstrong, J. S. (2nd eds). Principles of Forecasting: A Handbook for Researchers and Practitioners, Combining Forecasts. Norwell, MA: Kluwer Academic Publishers.

[7] Armstrong, J. S. & Fildes, R. (1995). On the selection of error measures for comparisons among forecasting methods. Journal of Forecasting, 14, 67-71.

[8] Armstrong, J. S. (1985). Long-term Forecasting: From Crystal Ball to Computer (2nd eds). New York: John Wiley.

[9] Arnell, N.; Tompkins, E. & Adger, N. (2005). Vulnerability to abrupt climate change in Europe. Technical Report 34, Tyndall Centre for Climate Change Research, Norwich Babbitt.

[10] Ashton, A. H. & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. Management Science, 31, 1499-1508.

[11] Ashton, A. H. (1986). Combining the judgments of experts: How many and which ones? Organizational Behavior and Human Decision Processes, 38, 405-414.

[12] Ashton, R. H. (1974). An experimental study of internal control judgments. Journal of Accounting Research, 12, 143-157.

[13] Asian Development Bank (2005). Accuracy of Asian Development Outlook Forecasts.

[14] Ayuub, B. M. (2001). A Practical Guide on Conducting Expert-Opinion Elicitation of Probabilities and Consequences for Corps Facilities. Prepared for U.S. Army Corps of Engineers, Institute for Water Resources.

[15] Ayyub, B. M. (2001). Elicitation of Expert Opinions for Uncertainty and Risks.

[16] BaFail, A. O (2004). Applying Data Mining Techniques to Forecast Number of Airline Passengers in Saudi Arabia (Domestic and International Travels). King Abdul Aziz University.

[17] Baldwin, J. M. (1975). Thought and things: a study of the development and meaning of thought or generic logic. New York: The Macmillan Company.

[18] Batchelor, R. A. & Dua, P. (1995). Forecaster Diversity and the Benefits of Combining Forecasts. Management Science, 41, 68-75.

[19] Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. Operational Research Quarterly, 20, 451-468.

[20] Batz, M. B.; Hoffmann, S. A.; Krupnick, A. J.; Morris, J. G.; Sherman, D. M.; Taylor, M. R. & Tick; J. S. (2004). Identifying the Most Significant Microbiological Food-borne Hazards to Public Health: A New Risk Ranking Model, Food Safety Research Consortium. Discussion Paper Series Number (1), September 2004 - FIRRM Food Attribution Percentages for Illnesses from Foodborne Campylobacter and Listeria monocytogenes.

[21] Bedford, T.; Quigley, J. & Walls, L. (2006). Expert Elicitation for Reliable System Design. Statistical Science, 21, 4, 428–450.

[22] Bonano, E. J. & Apostolakis, G. E. (1991). Thoretical foundation and practical issues for using expert judgments in uncertainty analysis of high-level radioactive waste disposal. Radioactive Waste Management and the Nuclear Fuel Cycle, 16, 2, 137–159.

[23] Bonano, E. J.; Hora, S. C.; Keeney, R. L. & von Winterfeldt, D. (1990). Elicitation and Use of Expert Judgment in PerformanceAssessment for High-Level Radioactive Waste Repositories. NUREG/CR-5411. Washington, DC: Nuclear Regulatory Commission.

[24] Bonano, E. J.; Hora, S. C.; Keeney, R. L. & von Winterfeldt, D. (1990). Elicitation and use of expert judgment in performance assessment for high-level radioactive water repository. Nuclear Regulatory Commission, NUREG/CR-5411, Washington, DC.

[25] Bond, C. E.; Gibbs, A. D.; Shipton, Z. K. & Jones, S. (2007). What do you think this is? ‘‘Conceptual uncertainty’’ in geoscience interpretation. GSA Today, 17, 11, 4–10.
[26] Booker, J. M. & Meyer, M. (1996). Elicitation and Analysis of Expert Judgment. Los Alamos National Laboratory, LA-UR-99-1659.
[27] Budescu, D. V. & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. Acta Psychologica, 104, 371-398.
[28] Campbell P. R. (2002). Evaluating Forecast Error in State Population Projections Using Census 2000 Counts. U.S. Bureau of Census, Population Division Working Paper No. 57.
[29] Cathers, C. A. & Thompson, G. D. (2005). Forecasting Short-Term Electricity Load Profiles, Cardon Research Papers.
[30] Chase, W. G. & Simon, H. A. (1973). The mind’s eye in chess. Visual Information Processing. W. G. Chase. New York, Academic Press**,** 215–281.
[31] Chen, Z. & Yang, Y (2004). Assessing Forecast Accuracy Measures. Iowa State University
[32] Chhibber, S. & Apostolakis, G. (1993). Some approximations useful to the use of dependent information sources. Reliability Engineering and System Safety, 42, 67-86.
[33] Christensen, C.; Christensen, R. & Huffman, M. D. (1985). Bayesian point estimation using the predictive distribution. The American Statistician, 39, 319-321.
[34] Claessens, M. (1990). An application of expert opinion in ground water transport. DSM Technical Report – R 90 8840, University of Delft, the Netherlands.
[35] Clemen, R. T. & Reilly, T. (February of 1999). Correlations and Copulas for Decision and Risk Analysis. Management Science, 45, 2.
[36] Clemen, R. T. & Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. Operations Research, 33, 427-442.
[37] Clemen, R. T. & Winkler, R. L. (1993). Aggregating Point Estimates: A Flexible Modeling Approach. Management Science, 39, 501-515.
[38] Clemen, R. T. & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. Risk Analysis, 19, 187-203.
[39] Clemen, R. T. & Winkler, R. L. (October of 1997). Combining Probability Distributions from Experts in Risk Analysis.
[40] Clemen, R. T.; Fischer, G. W. & Winkler, R. L. (2000). Assessing dependence: Some experimental results. Management Science, 46, 1100-1115.
[41] Clifford A. Cathers, C. A. & Thompson, G. D. (2006). Forecasting Short-Term Electricity Load Profiles. Sierra Southwest Cooperative Services, Inc. The University of Arizona, Cardon Research Papers, August 2006.
[42] Cooke, R. M. (1991). Experts in uncertainty: Opinion and subjective probability in science. New York: Oxford University Press.
[43] Cooke, R. M. (1994). Uncertainty in dispersion and deposition in accident consequence modeling assessed with performance-based expert judgment. Reliability Engineering and System Safety, 45, 35 – 46.
[44] Cooke, R. M. (2003). Book review of Elicitation of Expert Opinions for Uncertainty and Risks, by B. M. Ayyub. Fuzzy Sets and Systems, 133, 267–268.
[45] Cornish, E. (1977). The study of the future. World Future Society, Washington, D.C.
[46] Dalkey, N. C. (1970). The Delphi Method: An Experimental Study of Group Opinion. Technical Report RM-5888-PR, The Rand Corporation.
[47] Dawes, R. M. (1979). The robust beauty of improper linear models. American Psychologist.
[48] Dawid, A. P. (1982). The Well-Calibrated Bayesian. Journal of the American Statistical Association, 77, 605-613.
[49] de Finetti, B. (1937). Its logical laws, its subjective sources. In Kyburg, Jr. & Smokler, Studies in Subjective Probability, 2nd (eds), 53-118. Huntington, NY: Robert E. Krieger.
[50] De Finetti, B. (1964). La pervision: ses loislogique, ses source subjectives. Annales de Linstut Henri Poincare, 7:1-68 (1937). English translation in Kyburg & Smokler (eds.), Studies in Subjective Probability, Wiley, New York.
[51] Dechow, P. M., Sloan, R. G. (1997). Returns to contrarian investment: tests of the naïve expectations hypotheses. Journal of Financial Economics, 43, 3-28.
[52] DeGroot, M. H. (1988) A Bayesian view of assessing uncertainty and comparing expert opinion. Journal of Statistical Planning and Inference, 20, 295-306.
[53] DeWispelare, A. R.; Herren, L. T. & Clemen, R. T. (1995). The use of probability elicitation in the high-level nuclear waste regulation program. International Journal of Forecasting 11, 5–24.

[54] Draper, D.; Pereira, A.; Prado, P.; Saltelli, A.; Cheal, R.; Eguilior, S.; Mendes, B. & Tarantola, S. (1999). Scenario and parametric uncertainty in GESAMAC: a methoological study in nuclear waste disposal risk assessment. Computer Physics Communications 117, 142–155.
[55] Dumler, T. J. (2003). Rainfall and Farm Income. Risk and Profit Conference.
[56] Droguett, E. L., Mosleh, A. Framework for Integrated Treatment of Model and Parameter Uncertainties. Submitted to the Risk Anal.
[57] Droguett, E. L., Mosleh, A. Assessment of Model Uncertainty: Use of Model Performance Data. Submitted to the Risk Anal.
[58] Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (ed.), Formal Representation of Human Judgment, 17-52. New York: John Wiley.
[59] Ehrman, C. M. & Shugan, S. M. (1995). The Forcaster's Dilemma. Marketing Science, 14, 2, 123-127.
[60] Einhorn, H. J.; Hogarth, R. M. & Klempner, E. (1977). Quality of group judgment. Psychological Bulletin, 84, 158–172.
[61] Ericsson, K. A.; Krampe, R. T. & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363-406.
[62] Ericsson, K. A. & Staszewski, J. J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. In Klahr & Kotovsky (eds.), Complex information processing: The impact of Herbert A. Simon, 235-267. Hillsdale, NJ: Lawrence Erlbaum.
[63] Fischhoff, B. (1982). Debiasing. In Judgment under uncertainty: heuristics and biases (eds), Kahnemann, Slovic and Tversky, 422-444. Cambridge University Press.
[64] Fischhoff, B. (1982). For those condemned to study the past: Heuristics and biases in hindsight. In Judgment under uncertainty: heuristics and biases (eds), Kahnemann, Slovic and Tversky, 335-351. Cambridge University Press.
[65] Forrester, Y. (2005). The Quality of Expert Judgment: An Interdisciplinary Investigation. Doctor of Philosophy Dissertation, University of Maryland.
[66] French, S. (1985). Group Consensus Probability Distribution: A Critical Survey. In Bayesian Statistics (2nd eds), JM Bernado et al. Amsterdam, 183-201.
[67] French, S., Ríos Insua, D. Statistical decision theory. London: Arnold, 2000.
[68] Fullerton Jr., H. N. Evaluating the 1995 BLS Labor Force Projections.
[69] Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting prior distributions. Journal of the American Statistical Association, 100, 680-700.
[70] Genest, C. & McConway, K. J. (1990). Allocating the weights in the linear opinion pool. Journal of Forecasting, 9, 53–73.
[71] Genest, C. & Zidek, J. V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. Statistical Science, 1, 114-135.
[72] Genest, C., & Schervish, M. J. (1985). Modeling expert judgments for Bayesian updating. Annals of Statistics, 13, 1198-1212.
[73] Geomatrix Consultants (1998). Saturated zone flow and transport expert elicitation project. Deliverable Number SL5X4AM3. CRWMS M&O, Las Vegas, NV.
[74] Ghabayen, S. M. S.; McKee, M. & Kemblowski, M. (2006). Ionic and isotopic ratio for identification of salinity sources and missing data in the Gaza aquifer. Journal of Hydrology 318, 1–4, 360–373.
[75] Gigone, D. & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. Psychological Bulletin, 121, 149-167.
[76] Gilliland, M. (2002). Is Forecasting a Waste of Time? July/August issue of Supply Chain Management Review.
[77] Goossens, L. H. J. & Harper F. T. (1998). Joint EC/USNRC expert judgment driven radiological protection uncertainty analysis. J. of Radiological Protection, 18, 4, 249–264.
[78] Goossens, L. H. J.; Cooke, R. M.; Hale, A. R. & Rodic´-Wiersma, Lj. Fifteen years of expert judgement at TU Delft. Safety Science 46 (2008) 234–244
[79] Gustafson, D. H.; Shukla, R. K.; Delbecq, A. & Walster, G. W. (1973). A Comparative Study of Different Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups, Organizational Behavior Human Perform, 9, 200–291.
[80] Hammitt, J. K. & Shlyakhter, A. I. (1999). The Expected Value of Information and the Probability of Surprise. Risk Analysis 19, 1, 135-152.
[81] Hawkins N. C. & Evans J. S. (1989). Subjective estimation of toluene exposures: a calibration study of industrial hygienists. Appl. Ind. Hyg. J. 4, 3, 61–68.

[82] Helmer, O. (1977). Problems in futures research: Delphi and causal cross-impact analysis. Futures, 17-31.
[83] Hill, G. W. (1982). Group vs. individual performance: Are N+1 heads better than one? Psychological Bulletin, 91, 517-539.
[84] Hogarth, R. M. (1977). Methods for aggregating opinions. In Jungermann & DeZeeuw (eds.), Decision Making and Change in Human Affairs, 231–255. Dordrecht, Netherlands: Reidel.
[85] Hogarth, R. M. (1978). A note on aggregating opinions. Organizational Behavior and Human Performance, 21, 40-46.
[86] Hora, S. C. & Iman R. L. (1989). Expert opinion in risk analysis: The NUREG-1150 methodology. Nuclear Science and Engineering, 102, 323-331.
[87] Hora, S. & von Winterfeldt, D. (1997). Nuclear waste and future societies: A look into the deep future. Technological Forecasting and Social Change, 56, 155-170.
[88] Hora, S., Jensen, M., 2005. Expert panel elicitation of seismicity following glaciation in Sweden. SSI Report 2005:20, Swedish Radiation Protection Authority.
[89] Hynes M. E. & Van Marke E. H. (1977). Reliability of embankment performance predictions. In: Mechanics in Engineering, 1st ASCE-EMD Specialty Conference, University of Waterloo.
[90] Johnson, A. C. & Thomopoulos, N. T. Tables and Characteristics of Standardized Lognormal Distribution.
[91] Jouini, M. N. & Clemen, R. T. (1996). Copula models for aggregating expert opinions. Operations Research, 44, 444-457.
[92] Kadane, B. & Wolfson, J. (1998). Experiences in Elicitation. The Statistician, 47, 3–19.
[93] Kadane, J. B. & Winkler, R. L. (1988). Separating Probability Elicitation from Utilities. Journal of the American Statistical Association, 83, 402, 357–363.
[94] Kahn, H. & Wiener, A. J. (1967). The Year 2000: A Framework for Speculation on the Next Thirty-Three Years, London: Collier-Macmillan Limited.
[95] Kallen, M. J. and Cooke, R. M. (2002). Expert aggregation with dependence. In Probabilistic Safety Assessment and Management, In Bonano, Camp, Majors and Thompson (eds.), 1287–1294. North-Holland, Amsterdam.
[96] Keeney, R. L. & von Winterfeldt, D. (1989). On the uses of expert judgment on complex technical problems. IEEE Transactions on Engineering Management, 36, 83-86.
[97] Klugman, S. F. (1945). Group judgments for familiar and unfamiliar materials. Journal of General Psychology, 32, 103-110.
[98] Krishnamurti, T. N. et al. (1999). Improved weather and seasonal climate forecasts from multimodal ensemble. Science, 285, 1548-1550.
[99] Lacke, C. (1998). Decision Analytic Modeling of Colorectal Cancer Screening Policies. Operations Research Program, North Carolina State University.
[100] Lannoy, A. & Procaccia, H. (2001). L'utilisation du jugement d'expert en s^uret_e de fonctionnement, Tec & Doc (in French).
[101] Libby, R. & Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. Organizational Behavior and Human Performance, 21, 121-129.
[102] Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. Organizational Behavior and Human Decision Processes, 26, 149-7 1.
[103] Lindley, D. V. (1985). Reconciliation of discrete probability distributions. In Bernardo, DeGroot, Lindley, & Smith (eds.), Bayesian statistics 2, 375-390. Amsterdam: Holland.
[104] Linstone, H. & Turoff, M. (1975). The Delphi method: techniques and applications reading. Addison Wesley, Nass.
[105] Mc Laren, C. H. & Mc Laren, B. J. (2003). Electric Bill Data. Journal of Statistics, Online Edition, 11, 1.
[106] McIntosh, C. S. & Bessler, D. A. (1988). Forecasting Agricultural Prices Using a Bayesian Composite Approach. Southern Journal of Agricultural Economics.
[107] McKenna, S. A.; Walker, D. D. & Arnold, B. (2003). Modeling dispersion in three-dimensional heterogeneous fractured media at Yucca Mountain. Journal of Contaminant Hydrology 62, 3, 577–594.
[108] McLean, I.; Phil, M.; Anderson, M. & White, C. (2003). The accuracy of guestimates by 11 Forensic Clinicians. Journal of the Royal Society of Medicine, 96J, 96, 497–498.
[109] Mendel, M. & Sheridan, T. (1989). Filtering Information from Human Experts. IEEE Transactions on Systems, Man & Cybernetics, 36, 6-16.

[110] Meyer, M. A. & Booker, J. M. (1991). Eliciting and Analyzing Expert Judgment: A Practical Guide. London: Academic Press.
[111] Miklas, M. P. J.; Norwine, J.; DeWispelare, A. R.; Herren, L. T. & Clemen, R. T. (1995). Future climate at Yucca Mountain, Nevada proposed high-level radioactive waste repository. Global Environmental Change 5, 3, 221–234.
[112] Morgan, M. G. & Keith, D. W. (1995). Subjective judgments by climate experts. Environmental Policy Analysis 29, 10, 468–476.
[113] Morgan, M. G., & Henrion, M. (1990). Uncertainty: A Guide to Dealing with Uncertainty in Risk and Policy Analysis. Cambridge University Press, Cambridge.
[114] Morris, J. M. & D'Amore, R. J. (1980). Aggregating and Communicating Uncertainty. Pattern Analysis and Recognition Corp., 228 Liberty Plaza, NY.
[115] Morris, P. A. (1974). Decision analysis expert use. Management Science, 20, 1233-1241.
[116] Morris, P. A. (1977). Combining expert judgments: a Bayesian approach. Management Science, 23, 679-693.
[117] Morris, P. A. (1983). An axiomatic approach to expert resolution. Management Science, 29, 24-32.
[118] Morris, P. A. (1986). Observations on Expert Aggregation. Management Science, 32, 321-328.
[119] Mosleh, A. & Apostolakis, G. (1986). The Assessment of Probability Distributions from Expert Opinions with an Application to Seismic Fragility Curves. Risk Analysis Journal, 6, 4, 447-461.
[120] Mosleh, A (1981). On the Use of Quantitative Judgment in Risk Assessment: A Bayesian Approach. Dissertation, University of California, Los Angeles.
[121] Murphy, A. H. & Winkler, R. L. (1977). Reliability of Subjective Probability Forecasts of Precipitation and Temperature. Applied Statistics, 26, 41-47.
[122] Nelsen, R. B. (1999). Introduction to copulas. Springer Verlag, New York.
[123] O'Hagan, A. & Oakley, J. E. (2004). Probability is perfect, but we can't elicit it perfectly. Reliability Engineering & System Safety 85, 239–248.
[124] O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. The Statistician, 47, 1, 21–35.
[125] Ouchi, F. (2004). A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis. World Bank Policy Research Working Paper 3201.
[126] Parent, E. & Bernier, J. (2003). Encoding prior experts judgments to improve risk analysis of extreme hydrological events via POT modeling. J. of Hydrology, 283, 1–18.
[127] Pike, W. A. (2004). Modeling drinking water quality violations with Bayesian networks. Journal of the American Water Resources Association, 40, 6, 1563–1578.
[128] Quigley, M. A.; Chandramohan, D. & Rodrigues, L. (1999). Diagnostic Accuracy of Physician Review, Expert Algorithms & Data-derived Algorithms in Adult. Verbal Autopsies, International Journal of Epidemiology, 28, 1081-1087.
[129] Ramachandrani, G.; Banerjee, S. & Vincent, J. H. (April of 2003). Expert Judgment and Occupational Hygiene: Application to Aerosol Speciation in the Nickel Primary Production Industry. Ann. occup. Hyg., 47, 6, 461–475, British Occupational Hygiene Society, Published by Oxford University Press.
[130] Rantilla, A. K. & Budescu, D. V. (1999). Aggregation of Expert Opinions. Proceedings of the 32nd Hawaii International Conference on System Sciences.
[131] Rodier, C. J. (2005). Test of Model Error in Travel Forecast, Verify the accuracy of land use model used in transportation and air quality planning: a case study in Sacramento, California region. MTI Report 05-02.
[132] Rodríguez, J. L. (2005). Recent Applications of the Delphi Method in Social Sciences. Institute of Applied Business Economics, University of the Basque Country/Euskal Herriko Unibertsitatea (UPV/EHU).
[133] Sanders, N. R. (1992). Accuracy of judgmental forecasts. Omega: International Journal of Management Science, 20, 353–364.
[134] Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. Journal of the American Statistical Association, 66, 783-801.
[135] Schmittlein, D. C.; Kim, J. & Morrison, D. G. (1990). Combining forecasts: Operational adjustments to theoretically optimal rules. Management Science, 36, 1044-1056.
[136] Schwartz, Z. & Cohen, E. (2004). Hotel Revenue management Forecasting - Evidence of Expert-judgment Bias. Cornell University.
[137] Shanteau, J., (2002). Domain differences in expertise.

[138] Slovic, P. (1972). From Shakespeare to Simon: Speculation and some evidence about man's ability to process information. Oregon Research Inst. Research Monograph, 12, 2.
[139] Slovic, P. (1972). Information processing, situation specificity and the generality of risk-taking behavior. Journal of Personality and Social Psychology, 22, 128-134.
[140] Slovic, P. (1972). Psychological study of human judgment: Implications for investment decision making. The Journal of Finance, **27,** 4, 779-799.
[141] Sniezek, J. A. & Henry, R. A. (1989). Accuracy and confidence in group judgment. Organizational Behavior and Human Decision Processes, 43, 1-28.
[142] Spiteri, A.; Torpiano, J.; Bailey, M.; Mercieca, V. & Grech, V. (2004). A comparison of clinical pediatric murmur assessment with echocardiography. Malta Medical J. 16, 04.
[143] Stark, K. C.; Wingstrand, A.; Dahl, J.; Møgelmose, V. & Lo Fo Wong, D. M. A. (2002). Differences and similarities among experts' opinions on Salmonella entrica dynamics in swine pre-harvest. Preventive Veterinary Medicine, 53, 7-20.
[144] Stekler, H. O. & Thomas, R. (2000). Evaluating BLS Labor Force, Employment, and Occupation Projection for 2000.
[145] Stiber, N. A.; Pantazidou, M. & Small, M. J. (1999). Expert system methodology for evaluation reductive dechlorination at TCE sites. Environmental Science and Technology, 33, 17, 3012–3020.
[146] Stiber, N. A.; Small, M. J. & Pantazidou, M. (2004). Site-specific updating and aggregation of Bayesian belief network models for multiple experts. Risk Analysis 24, 6, 1529–1538.
[147] Stone, M. (1961). The opinion pool. Annals of Math. Statistics, 32, 1339-1342.
[148] Tennessee Valley Authority. Appendix B – Methodology and Results from Socioeconomic Modeling. Final Environmental Assessment.
[149] Tversky, A. & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. Science, 185, 1124-1131.
[150] Vegelin A. L.; Brukx, L. J. C. E.; Waelkens, J. J. & Van den Broeck, J. (2003). Influence of knowledge, training and experience of observers on the reliability of anthropometric measurements in children. Annals of Human Biology, 30, 1, 65-79.
[151] Walker, K. D.; Evans, J. S. & MacIntosh, D. (2001). Use of expert judgment in exposure assessment - Part 1. Characterization of personal exposure to benzene. Journal of Exposure Analysis and Environmental Epidemiology, 11, 308-322.
[152] Walker, K.; Catalano, P.; Hammitt, J. & Evans, J. (2003). Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene. Journal of Exposure Analysis and Environmental Epidemiology 13, 1-16.
[153] Williams, A. M., & Ericsson, K. A. (2005). Perceptual-cognitive expertise in sport: Some considerations when applying the expert performance approach, Human Movement Science, 24, 283-307.
[154] Wilson, J. M. (1994). Network representations of knowledge about chemical equilibrium: Variations with achievement. Journal of Research in Science Teaching, 31, 1133-1147.
[155] Winkler, R. L. & Makridakis, S. (1983). The Combination of Forecasts. Journal of the Royal Statistical Society, A, 146, 150-157 (1983).
[156] Winkler, R. L. & Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. Management Science, 39, 1526-1543.
[157] Winkler, R. L. (1968). The Consensus of Subjective Probability Distributions. Management Science, 15, 61-75.
[158] Winkler, R. L. (1981). Combining Probability Distributions from Dependent Information Sources. Management Science, 27 479-488.
[159] Wisse, B., Bedford, T. & Quigley, J. (2005). Combining expert judgments in the Bayes linear methodology. In Proc. CEA-JRC Workshop on the Use of Expert Judgment in Decision-Making, Devictor, Moulin and Bolado-Lavin, (eds.). CEC, Aix-en-Provence.
[160] Wissema, G. (1982). Trends in technology forecasting. R&D Management, 12, 1, 27-36.
[161] Zajonc, R. B. (1962). A note on group judgments and group size. Human Relations, 15, 177–180.
[162] Zarnowitz, V. 1984. Business Cycles Analysis and Expectational Survey Data. NBER Working Papers 1378, National Bureau of Economic Research, Inc.
[163] Zio, E. & Apostolakis, G. E. (1996). Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories. Reliability Engineering and System Safety 54, 225–241.