

ABSTRACT

Title of Dissertation: A META-DATA INFORMED EXPERT JUDGMENT AGGREGATION AND CALIBRATION TECHNIQUE

Ellis Steven Feldman, Doctor of Philosophy, 2016

Dissertation directed by: Professor Ali Mosleh
Department of Mechanical Engineering

Policy makers use expert judgment opinions elicited from experts as probability distributions, quantiles or point estimates, as inputs to decisions that may have economic or life and death impacts. While challenges in estimating probabilities in general have been studied, research that distinguished between non-probabilistic, i.e., physical, variables and probabilistic variables specifically in the context of meta-data based expert judgment aggregation techniques, and the errors associated with the predictions developed from such variables, was not identified.

This research demonstrated that for two combined expert judgment meta-data bases, the distinction between physical and probabilistic variables was significant in terms of the extent of multiplicative error between elicited medians and realized values both before and after aggregation. The distinction also impacts the widths of bounds around aggregated point estimates. The research compared nine methods of aggregating estimates and obtaining calibrated bounds, including ones based on alpha stable

distributions, quantile regression, and a Bayesian model. Simple parametric distributions were also fit to the meta-data. Methods were compared against criteria including accuracy, bounds coverage and width, sensitivity to outliers, and complexity. No single method dominated all criteria for either variable type.

The research investigated sensitivity of results to level of realized value for a variable, such as infrequent events for probabilistic variables, as well as sensitivity of results to number of experts.

A META-DATA INFORMED EXPERT JUDGMENT AGGREGATION AND
CALIBRATION TECHNIQUE

by

Ellis Steven Feldman

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
2016

Advisory Committee:
Professor Ali Mosleh, Chair
Professor Hugh Bruck
Professor Peter Sandborn
Professor Monifa Vaughn-Cooke
Professor Martin Dresner (Dean Representative)

© Copyright by
Ellis Steven Feldman
2016

Dedication

To my parents, Nathaniel and Clara Feldman.

Acknowledgements

I would like to extend my deepest appreciation to Professor Ali Mosleh for his encouragement, insights, and support throughout this dissertation. I would also like to thank all those at UMD and the Federal Aviation Administration who contributed to this research. I am most grateful to my son, daughter, and son-in-law, my sister and brothers and their spouses, and last but not least, to my wife.

Table of Contents

Table of Contents

Dedicationii
Acknowledgements.....	iii
Table of Contents.....	iv
List of Tables	vii
List of Figures	viii
Chapter 1: Introduction	1
1.1 Background	3
1.1.1 Definitions of Expert Judgment	3
1.1.2 Point Estimates.....	5
1.1.3 Aggregation of Expert Judgment.....	7
1.1.4 Data Sets	9
1.2 Problem Statement	10
1.3 Research Questions	10
1.4 Significance of the Study	11
1.5 Assumptions, Scope, Limitations, and Delimitations	12
1.6 Research Organization	13
1.7 Summary	14
Chapter 2: Data Sources.....	15
2.1 Introduction	15
2.2 Description of Data Sources.....	15
2.3 Data Sources Preparation	18
Chapter 3: Research Question 1 (Physical versus Probabilistic Expert Estimates).....	21
3.1 Introduction	21
3.2 Research Question 1 Significance	21
3.2.1 Justification of metric	22
3.3 Research Question 1 Literature Review.....	23
3.4 Research Question 1 Mathematical Formalism	28
3.5 Research Question 1 Methodology and Results.....	32

3.5.1. Comparison of Prediction Accuracy between Full Physical and Probabilistic Datasets.....	40
3.5.2 Maximum Multiplicative Error, MME	44
3.5.3 Bounding Intervals	46
3.6 Summary	47
Chapter 4: Research Question 2 (Aggregation Methods).....	51
4.1 Introduction	51
4.2 Research Question 2 Significance.....	53
4.3 Research Question 2 Literature Review	54
4.3.1 Choice of metric	61
4.3.2 Linear Pooling Techniques.....	63
4.3.4 Alpha-Stable distribution.....	71
4.4 Research Question 2 Mathematical Formalism	74
4.4.1 Calibration Score	75
4.4.2 Information score.....	78
4.4.3 Alpha-stable distribution	81
4.4.4 Gaussian mixture	83
4.4.5 Median	84
4.4.6 Bayesian aggregation.....	84
4.5 Research Question 2 Analysis and Results	91
4.5.1 Bayesian Likelihood Function.....	93
4.5.2 Aggregation via Alpha-Stable Distribution	105
4.5.3 Gaussian Mixture.....	110
4.5.4 Rule of Thumb (ROT)	110
4.5.5 Maximum Likelihood Estimate	110
4.5.6 Classical model.....	111
4.5.7 Median, with bounds via quantile regression	111
4.5.8 Accuracy of aggregated results.....	111
4.5.9 Absolute distances between coverage percentages and 90% Over Each Theme	113
4.5.10 Sensitivity to Outliers	114

4.6 Summary	132
Chapter 5: Research Question 3 (Magnitude of e).....	135
5.1 Introduction	135
5.2 Research Question 3 Significance.....	137
5.3 Research Question 3 Methodology and Results.....	137
5.4 Summary	152
Chapter 6: Research Question 4 (Number of Experts)	159
6.1 Introduction	159
6.2 Research Question 4 Significance.....	159
6.3 Research Question 4 Literature Review.....	160
6.4 Research Question 4 Mathematical Formalism	164
6.5 Summary	172
Chapter 7: Research Question 5 (Parametric Distributions for Bounds)	175
7.1 Introduction	175
7.2 Research Question 5 Significance.....	175
7.3 Research Question 5 Mathematical Formalism	175
7.4 Research Question 5 Methodology and Results.....	176
7.5 Summary	181
Chapter 8: Analysis of Test Data	182
8.1 Introduction	182
8.2 Test Data Description.....	182
8.3 Test Data Results.....	186
8.4 Comparison between EJE and TD Results.....	189
8.4.1 Probabilistic Data	190
8.4.2 Physical Data	190
8.4.3 Bounds Coverage Against Probabilistic TD.....	191
8.4.4 Bounds Coverage Against Physical TD	193
Chapter 9: Conclusions and Recommendations	200
Appendix A: EJE and TD Referencing.....	217
Appendix B: Shapiro-Wilk Analysis	219
Appendix C: Sample of MathWave EasyFit Outputs	225

Bibliography	230
--------------------	-----

List of Tables

Table 1: TUD Case Names and EJE Theme Names Mapping	16
Table 2: EJDS: Summary-level Listing of Record, Prediction and Theme Count by Data Source	17
Table 3: Chances of Under and Over-Estimation given a single prediction, e' before aggregation.....	53
Table 4: Chances of Under and Over-Estimation after e' predictions have been aggregated into \hat{e}	54
Table 5: Alpha-Stable parameters.....	106
Table 6: Gaussian Mixture Parameters for Physical and Probabilistic Data	110
Table 7: MME Scorecard for EJE Physical Data.....	119
Table 8: MME Scorecard for EJE Probabilistic Data.....	119
Table 9: MME Scorecard for EJE Physical Data- TUD Records	122
Table 10: MME Scorecard for EJE Probabilistic Data- TUD Records	122
Table 11: EJE Physical Data TUD Records - 90% Coverage by Aggregation Method .	125
Table 12: EJE Physical Data - 90% Coverage by Aggregation Method	125
Table 13: EJE Physical Data TUD Records – 80% Coverage by Aggregation Method	126
Table 14: EJE Physical Data - 80% Coverage by Aggregation Method	126
Table 15: EJE Probabilistic TUD Data - 90% Coverage by Aggregation Method	127
Table 16: EJE Probabilistic TUD Data - 80% Coverage by Aggregation Method	127
Table 17: Average MME by Aggregation Method for Record with Outliers – Physical Data	131
Table 18: Average MME by Aggregation Method for Records with Outliers – Probabilistic Data.....	131
Table 19: Complexity Rating.....	133
Table 20: Physical Data Scorecard	134
Table 21: Probabilistic Data Scorecard.....	134
Table 22: EJE Probabilistic Data – Spearman ρ by Aggregation Method for Bin Mid-Point	149
Table 23: EJE Physical Data- Spearman ρ by Aggregation Method for Bin Mid-Point	150
Table 24: Stratification EJE Probabilistic Data	152
Table 25: Stratification EJE Physical Data Bins [17.2, -11) – [1,2).....	154
Table 26: Stratification EJE Physical Data Bins [4,4.3333) – [15,21.14)	156
Table 27: Ranks of EJE Physical Data Bins Stratification	158
Table 28: Costs of Expert Judgment Studies (Yucca Mountain Site Characterization Project).....	162
Table 29: Distribution of Number of Predictions by Number of Records in EJE	165
Table 30: Cauchy Parameters for Bounds	178
Table 31: MME Scorecard for TD Probabilistic Data	188
Table 32: MME Scorecard TD Physical Data	189
Table 33: Weighted Probabilistic TD Bounds Coverage.....	191
Table 34: Probabilistic TD - Parametric Fit Bounds Developed Using Cauchy Distribution	191

Table 35: Weighted Physical TD Bounds Coverage	193
Table 36: Weighted Physical TD Bounds Coverage Subset "All themes except Software Hours Bid Phase"; mass 0.8.....	193
Table 37: Physical TD - Parametric Fit Bounds Developed Using Cauchy Distribution	194
Table 38: EJE Physical Data Aggregation Methods for Comparison with TD Physical Data	196
Table 39:TD Physical Data Aggregation Methods for Comparison with EJE Physical Data	196
Table 40: EJE Probabilistic Data Aggregation Methods for Comparison with TD Probabilistic Data.....	197
Table 41: TD Probabilistic Data Aggregation Methods for Comparison with EJE Probabilistic Data.....	198

List of Figures

Figure 1: Scatterplot of $e' - e$ vs. e	29
Figure 2: Scatterplot of e'/e vs. e	29
Figure 3: CDFs of $r \equiv e/e'$ for Physical Subset and Probabilistic Data	33
Figure 4: Overestimation Probabilities by Factor for Physical Subset and Probabilistic Data	36
Figure 5: Underestimation Probabilities by Factor for Physical Subset and Probabilistic Data	36
Figure 6: CDFs of e/e' for Physical and Probabilistic Data	41
Figure 7: Physical Data and Probabilistic Data – Overestimation Probability by Factor.	42
Figure 8: Physical Data and Probabilistic Data – Underestimation Probability by Factor	42
Figure 9: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data	45
Figure 10: Physical Data and Probabilistic Data – MME Exceedance Probability by Factor	46
Figure 11: Schematic depiction of seed item realizations for a well-calibrated expert	76
Figure 12: Density of $\ln(e/e')$ for Physical and Probabilistic Data—Detail over [-3,3] ..	92
Figure 13: Percentiles of $e' - e e$ for Probabilistic Data.....	94
Figure 14: Quantiles of e' versus e for Probabilistic Data	96
Figure 15: Quantiles of e' versus e for Physical Data	97
Figure 16: Distance between Ln-transformed 95 th and 5 th Percentiles versus $\ln(e)$..	100
Figure 17: Q-Q Plot of Probabilistic Data Alpha-Stable fit.....	107
Figure 18: Q-Q Plot of Physical Data Alpha-Stable fit	108
Figure 19: CDF of e/\hat{e} for Physical and Probabilistic Data	112
Figure 20: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data Using Aggregated Estimate, \hat{e}	113
Figure 21: Quantiles of \hat{e} versus e for Probabilistic Data	138
Figure 22: Quantiles of \hat{e} versus e for Physical Data.....	139
Figure 23: Probability $\{e' \leq s \cdot e, s < 1; e' \geq s \cdot e, s > 1\}$ for specified values of s —Probabilistic Data	141

Figure 24: Probability $\{e' \leq s \cdot e, s < 1; e' \geq s \cdot e, s > 1\}$ for specified values of s—Physical Data	142
Figure 25: Probability $\{\hat{e} \leq s \cdot e, s < 1; \hat{e} \geq s \cdot e, s > 1\}$ for specified values of s—Probabilistic Data	144
Figure 26: Probability $\{\hat{e} \leq s \cdot e, s < 1; \hat{e} \geq s \cdot e, s > 1\}$ for specified values of s—Physical Data	146
Figure 27: Weighted Average MME vs Number of Heads, s for Probabilistic Data	167
Figure 28: Weighted Average MME vs Number of Heads, s for Physical Data	169
Figure 29: Sensitivity of Total Cost for a Probabilistic Variable to Number of Heads, s	171
Figure 30: Sensitivity of Total Cost for a Physical Variable to Number of Heads, s	172
Figure 31: Quantiles of $e \hat{e}$ for Probabilistic Data.....	180
Figure 32: Quantiles of $e \hat{e}$ for Physical Data	180

Chapter 1: Introduction

In general, expert judgment is used chiefly “where there is uncertainty due to insufficient data, when such data is unattainable because of physical constraints or lack of resources” (Slottje, Sluijs, & Knol, 2008, p. 7), or to develop probabilistic assessments given cost and feasibility constraints in obtaining hard data. Policy makers use expert judgment opinions elicited from experts, in the form of probability distributions, quantiles or point estimates, as inputs to decisions. These decisions can have significant economic or even life and death consequences. For example, expert judgment was used to assist civil authorities on the island of Montserrat in setting alert levels during the volcanic eruptions of the mid-1990s (Aspinall and Cooke, 1998). Expert judgment elicitation has been used in connection with nuclear reactor safety. The Board of Governors of the Federal Reserve System (2013) recognized that material changes in bank holding companies’ businesses, or limitations in relevant data may lead some of these organizations to rely solely on expert judgment for certain loss, revenue, or expense projections. The U.S. Federal Aviation Administration (GAO, 2008a) has used expert judgment elicitations to predict the performance of surface radar based alerting systems in preventing serious runway incursions.

Clemen and Winkler (1999) noted that “a set of experts can provide more information than a single expert” (p.188). Winkler (1971) asserted that such input needs to be aggregated or combined into single distribution to be used “as the basis of decision making” (p. B-62). How best to combine these predictions into a single probability distribution - which can be used as an input for policy decisions - is an area of active

research. Notwithstanding the potential significant ramifications of an erroneous expert prediction, actual and predicted values can diverge by orders of magnitude.

Similar discrepancies have been observed between predictions made by different experts. These discrepancies need not arise from “incompetence, venality and ideology”, but rather “may be attributable to the character and fallibilities of human judgement itself” (Mumpower & Stewart, 1996, p.193). Although point estimates are less useful than interval estimates, the former continue to be observed in U.S. Government (USG) expert judgment elicitation studies, e.g. USDA (2007) and USDA (2012).

This research developed and applied a meta-data informed technique for generic calibration and aggregating expert judgment estimates from sets of elicited point estimates, and providing intervals bounding the estimate. The background for this research as well as the problem statement and its significance are discussed below. In addition, the specific research questions, assumptions, scope, limitations and delimitations and the structure in which they are addressed, and the overall organization of the research are also presented.

1.1 Background

The 1978 report to the U.S. Nuclear Regulatory Commission by Lewis et al. (1978) stated that “faced with the problem of estimating the probability of occurrence of an extremely rare event - core melt - in a system of great complexity, a nuclear power reactor”, where the event had “never occurred in a commercial reactor”, and system complexity rendered “a complete and precise theoretical calculation impossibly difficult”, it was “necessary to invoke simplified models, estimates, engineering opinion, and in the last resort, subjective judgments.” Such subjective judgments include point estimates and probability distributions. Formal elicitation of such estimates from Subject-Matter Experts (SMEs) are referred to as expert judgments. When these point estimates or probability distributions from multiple SMEs are combined, an aggregated expert judgment is formed.

1.1.1 Definitions of Expert Judgment

Meyer and Booker (2001) defined expert judgment as “data given by an expert in response to a technical problem” (p. 3). Hammitt and Zhang (2013) observed that “Expert judgement (or expert elicitation) is a formal process for eliciting experts’ beliefs or opinions about the value of a quantity that may be used as input to a model to inform policy decisions or for other purposes” (p. 109). Cooke and Probst (2006) asserted that the “role of structured expert judgment is to quantify uncertainty, not remove it” (p. 5) and presented nine theses regarding expert judgment. These theses stated that expert judgment is not knowledge; experts can disagree, are capable of quantifying uncertainty as subjective probability, are not consistently overconfident, and favor performance assessment. Additionally, these theses concluded that equal weighting is not optimal, and

that citation-based weights do not perform better (Theses 8 and 9 state respectively: “Uncertainty from random sampling … omits important sources of uncertainty”, and “The choice is not whether to use expert judgment, but whether to do it well or badly” (Cooke & Probst, 2006, p. 4).

The practice of expert judgment necessitates a definition of who is an expert. Johnston (2003) defined expert as “individuals with specialized knowledge suited to perform the specific tasks for which they are trained, but that expertise does not necessarily transfer to other domains”. In their study regarding the differences between expert performance and process, Camerer and Johnson (1991) defined an expert as a person “who is experienced at making predictions in a domain and has some professional or social credentials” (p. 196). Mumpower and Stewart (1996) defined experts as “those who are regarded as such by others within their field” (p. 193) for the purpose of examining disagreement among experts. Weinstein (1993) claimed that there are two kinds of experts; epistemic, i.e., “a capacity to provide strong justifications for a range of propositions in a domain” (p. 58) and performative, i.e., a capacity to perform a skill.

Murphy (1993) observed that from a forecaster’s perspective, the “goodness of a forecast is generally related… to the degree of similarity between the forecast conditions and the observed conditions” (p. 281). Such a determination is analogous to an expert judgment claim. However, if multiple predictions for a given event are available, Ranjan and Gneitin (2009) asserted that “there is strong empirical evidence that combined probability forecasts that draw on all the experts’ or models’ strengths result in improved predictive performance” (p. 71-72). The advantages of combining forecasts are

discussed, for example, by Clemen (1989) in a review of over 200 related articles and in Hendry and Clements (2002).

Cooke and Goossens (2008) claimed that expert judgment “is sought when substantial scientific uncertainty impacts on a decision process” (p. 657). Hora and Jensen (2002) identified circumstances in which expert judgment is used. Such circumstances include, but are not limited to: a lack of complete evidence; data existing only from analog situations; conflicting models; scaling up from experiments to target processes being indirect; and, significant uncertainties in achieving compliance.

According to Lin and Bier (2008), expert judgment can be used to “provide estimates regarding new, rare, complex, or poorly understood problems or phenomena” (p. 711).

1.1.2 Point Estimates

Snedecor and Cochran (1978) defined a point estimate as a single number that states “an estimate of some quantitative property of the population” (p. 29). The use of point estimates in USG cost estimates has been criticized by the General Accountability Office in their Cost Estimating and Assessment Guide as opposed to using confidence intervals. Specifically, “a point estimate, by itself, provides no information about the underlying uncertainty other than that it is the value chosen as most likely” (GAO, 2009, p.155). Sample GAO (1997) findings include the Federal Aviation Administration’s practice of using point estimates “instead of a range of estimates or a realistically qualified estimate”, resulting in “uninformed, …potentially unwise, investment decisions” (p. 36). The Census Bureau (GAO 2008b) did not “provide a level of confidence associated with the point estimate” (p. 19) of its short form cost estimate. A GAO review of Missile Defense Cost Estimates (GAO, 2012) found deficiencies in

computing intervals for point estimates. In 2014 the GAO (GAO, 2014) found that the US Nuclear Regulatory Commission cost estimation practices lacked “a risk and uncertainty analysis to assess variability in point estimates due to factors such as a lack of knowledge about the future or errors resulting from historical data inconsistencies” (p. 18).

Use of point-estimates in USG is not limited to cost estimates. The USDA (1999) used point estimates to assess Stocks to Use Ratio (the level of carryover stock for any given commodity as a percentage of the total demand or use) for price forecasting of crops. Further, up until the FY 2007 Presidential Budget (GPO, projected farm program outlays were based only on point estimates. (USDA, n.d.). Due to under-estimation, a transition to stochastic outlay estimates was made since these estimates “account for the price and corresponding outlay variability around the deterministic estimate” (USDA, n.d.).

According to the National Research Council (1994) “Use of a single point estimate suppresses information about sources of error that result from choices of model, data sets, and techniques for estimating values of parameters from data” (p. 185). Chernick (2011) asserted that point estimates are “useful but do not describe the uncertainty associated with them” (p. 62). An interval estimate is more informative than the point estimate “because the width of the interval conveys how well, in a probabilistic sense, the point estimate is known” (Lee & McCormick, 2012, Section 1.2).

Notwithstanding the limitations of point-estimation, it continues to be used in expert judgment elicitation. For example, FAA has used expert judgment elicitation involving FAA controllers and pilots in estimating the effectiveness of surface radar-

based alerting systems in preventing runway incursions and accidents. The promotion of eliciting interval-estimates is a challenge in and of itself. For example, in their recommendations regarding expert elicitation, the European Food Safety Authority (EFSA) noted that “All people involved in the process need training on making probability judgements” (EFSA, 2014, p. 119).

1.1.3 Aggregation of Expert Judgment

Keith (1996) suggested that aggregation of multiple expert inputs depends on the question “Who is the audience for the analysis and what do they need?” Booker and Meyer (2001) argued that multiple experts are required so that a problem will be addressed from different perspectives. Additionally, use of “a single expert will slant results towards the content and functioning of his or her memory” (Booker & Meyer, 2001, p. 87). Seaver (1976) observed that “In general the group judgment will be more accurate than the individual judgments primarily due to a decrease in error variance” (p. 25). Martz, Bryson, and Waller (1985) concluded that “Aggregation of expert opinion using group medians does give some improvement in accuracy” (Conclusions section).

The number of experts required for such aggregation is an open question. Aspinall (2010) observed that personal “experience with more than 20 panels suggests that 8–15 experts is a viable number — getting more together will not change findings significantly, but will incur extra expense and time. However, this has not been rigorously tested” (p. 295).

Clemen and Winkler (1999) noted that aggregation (or combination) procedures of expert judgment are frequently “dichotomized into mathematical and behavioral approaches” (p. 188) although a combination of the two approaches is common.

Mathematical approaches, which are the focus of this research, use modeling techniques to produce a single probability distribution whereas behavioral approaches aim to establish agreement among experts by having them interact. How best to aggregate expert judgments is an ongoing research issue.

Mathematical models for aggregating expert judgment can be categorized as either Non-Bayesian Axiomatic models or Bayesian models (Ouchi, 2004). Non-Bayesian Axiomatic models consist of Linear Opinion Pooling and Logarithmic Opinion Pooling methods (Clemen and Winkler, 1999; Genest and Zidek, 1986). As stated in Clemen and Winkler (1999) these methods are combined in Cooke's (1991) *classic model*. Bayesian models include single point estimate combination methods as well as methods for combining distributions (Clemen and Winkler, 1999).

The real-world situation modeled here is that expert judgment elicited information consists of sets of elicited point estimates, assumed to be medians. No information is assumed to be available as to the relative calibration of the eliciters. Policy considerations may preclude unequal weighting of elicited responses, e.g. from expert judgment panels of unionized air traffic controllers. This research compares and contrasts nine models, i.e., Alpha-Stable, Arithmetic, Bayesian, Classical, Geometric mean, Harmonic Mean, Median, Maximum Likelihood Method, and Rule of Thumb (related to the Alpha-Stable) for aggregating expert judgment into an estimate, \hat{e} and for developing bounds around \hat{e} . One of the models, Cooke's (1991) classical model, referenced hereinafter as the classical model, is allowed to make full use of elicited triplets of 5th, 50th and 95th percentiles for each meta-data variable (where available), along with calibration information developed from related variables for which the

elicitees also provided such triplets. The advantage this additional information gives to Cooke's (1991) aggregated estimates and bounds, over those developed without such information, discussed in Chapter 4.

1.1.4 Data Sets

The data sets used in this study for addressing the research questions posed in this dissertation are extracts from Delft University of Technology (TUD) expert judgment meta-databases, described in Cooke and Goossens (2008) and University of Maryland Center for Reliability and Risk Analysis (UMD) expert judgment meta-data sources. The TUD source is cited extensively in the peer-reviewed literature. The UMD source was derived from two dissertations addressing expert judgment (Forrester (2005); Shirazi (2009)) and related graduate coursework, conducted in the Department of Mechanical Engineering and thus provides an opportunity to introduce other similar data. All data consisted of named variables with known realized values. 5th and 95th percentiles were also provided for TUD data, but used only for the classical model. Each variable along with elicited values is called a record. Sets of variables related in terms of subject-matter were assigned to themes where each theme is unique to either TUD or UMD. In addition, a set of test data records was compiled during this research. The number of records from TUD and UMD are 606 and 1,182 respectively, for a total of 1,788 records. The test data consists of 37 records. The data sources are summarized in Chapter 2.

In summary, the objective is to provide tentative guidelines for aggregating and bounding elicited point estimates to increase decision support reliability using diverse data compilations.

1.2 Problem Statement

Given that reliance on expert judgments may have life and death, economic, or political impacts, their assessment in terms of reliability needs to be scrutinized. Challenges in estimating probabilities have been explored, e.g., Hogarth (1975a), Hogarth (1975b), Tversky and Kahneman (1974), Hilbert (2011), and Huer (1981). However, extensive literature contrasting the assessment of non-probabilistic, i.e., physical, variables and probabilistic variables in the context of meta-data based expert judgment aggregation techniques, and the errors associated with the predictions developed from such variables, was not identified. Because of these difficulties, a distinction between probabilistic and physical variables appears to be warranted in order to separate the differences between observed and predicted values, i.e., multiplicative excursions, generated for each variable type in the meta-database. Further, the distinction needs to be applied to the aggregation models for both point estimates and bounds as well as sensitivity to the number of experts, and to level of realized value, e.

1.3 Research Questions

The problem statement can be addressed in terms of the following research question sets:

Research Question One Set:

Does the type of quantity estimated, "physical"—variables having units of mass, time, etc.—or "probabilistic"—variables representing likelihood of an event, or a frequency of occurrence—matter? Percentiles, standard deviations or ratios of physical quantities were considered to be physical. If the meta-data are disaggregated into physical and probabilistic-related subsets, does the accuracy of elicited predictions change? Further,

given a point estimate elicited from an expert, what upper and lower bound multiplicative factors should be applied to that estimate, in order to bound it in an interval—with a corresponding level of probability?

Research Question Two Set:

Given a set of point estimates elicited from experts, how should they be combined to yield an aggregate estimate and associated bounds?

Research Question Three Set:

Does the magnitude of the quantity estimated matter? How does the range of multiplicative factor change between estimates of infrequent events, and estimates of frequent events?

Research Question Four Set:

How does the quality of the aggregated estimate vary with the number of experts used? Does a larger number of experts result in a tighter credible interval for the same level of probability? How significant is domain expertise?

Research Question Five Set:

Are there simple parametric distributions which perform reasonably well in yielding bounds around the estimate?

1.4 Significance of the Study

Decision makers may not fully appreciate the magnitude of range factors which must be applied to aggregated expert judgment elicited point estimates (e') in order to bound them with a given level of confidence, e.g., 90%. The significance of the type of variable (physical and/or probabilistic) for possible prediction error also may not be understood. This study used a large metadata set with several methods of aggregating

point estimates into an aggregate point estimate (\hat{e}) with associated bounds, to contrast the accuracy and precision of these quantities along with other scorecard factors such as complexity of calculation and stability of estimates to outliers in the data and calibration of bounds in the data. Results are compared to those resulting from a linear pooled calibration-based weighting scheme (classical model) which incorporated triplets of elicited percentiles for multiple seed variables. According to Cooke (1991), a model may need to be “...‘seeded’ with other events whose outcomes are known, or become known within a short time” if a decision maker requires information regarding specific events that will not occur within “...a required time frame” (p. 194). Sensitivity of accuracy to number of experts and to level of the true value of the variable, is also reported.

1.5 Assumptions, Scope, Limitations, and Delimitations

The assumptions, scope, limitation, and delimitations in this section are applicable to the entire body of research; research question-specific framing elements are presented in Chapters 3 through 7.

The following assumptions apply to this research:

1. Data used in this study per the sources listed in Chapter 2 were considered to be correct.
2. All processing packages used in this research; specifically, MathWave EasyFit, Python (Enthought Canopy), and Dell Statistica 13, were considered to be operating as intended
3. The distinction between physical and probabilistic data within the TU Delft database is assumed to be correct notwithstanding difficulties in translation and absence of descriptions in some cases (see Chapter 2).

The scope of this research was an experiment that addressed the five specific research question sets listed in the Research Questions section using quantitative techniques specified in Chapters 3, 4, 5, and 6.

Limitations are potential weaknesses or problems identified by the researcher and here include potential threats to reliability, internal validity, and external validity (generalizability). Carmines and Zeller (1979) defined reliability as “the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials” (p. 11). This research used static data extracts, which were subjected to identical computational procedures that were exercised repeatedly on the same platform. No resulting discrepancies were identified and consequently threats to reliability were not detected.

Delimitations identify exclusions from the research, re-enforcing its generalizability and scope. Furthermore, delimitations function to make the research feasible. The delimitations applicable to this research were the methods of elicitation used to obtain the data sets and expert judgment qualifications.

1.6 Research Organization

Chapter 1 introduces the research framework. Chapter 2 presents a summary of the data sources. Analysis of the research questions listed in Section 1.3 is presented in Chapters 3 through 7. Chapter 8 provides the results of applying the techniques developed in Chapters 3 through 7 to a set of expert judgment test data developed exclusively for this research and compares the results with the data used in that specific Chapter. Chapter 9 presents conclusions and recommendations developed from Chapter 3 through Chapter 8.

1.7 Summary

Since the circumstances in which expert judgment can be used may have life and death, economic, or political ramifications, it is necessary to understand how such judgments are constructed in terms of the types of estimates, data and models that are used. The use of point estimates precludes information about uncertainty, but is widely used. Lack of training and understanding is often cited as the reason for not providing interval estimates. Further, the aggregation of expert judgment is an ongoing research area. There is no hard and fast guideline as to the number of experts required for such aggregation.

Hora and Jensen (2002) identified circumstances in which expert judgment is used, e.g., lack of complete evidence; conflicting models exist; scaling up models from experiments to target processes is not direct, and uncertainties may be significant in achieving compliance. According to Lin and Bier (2008), expert judgment can be used to “provide estimates regarding new, rare, complex, or poorly understood problems or phenomena” (p. 711).

This research described an experiment for comparing and contrasting the accuracy and precision of aggregated expert judgment elicited values and associated bounds, compared to realized values for physical and probabilistic data types, using eight different models, and the classical model.

Chapter 2: Data Sources

2.1 Introduction

The data records used to develop the aggregation methods for this research include data from the Delft University of Technology (TUD) and University of Maryland Center for Reliability and Risk Analysis (UMD). These data sources are collectively referred to the Expert Judgment Extracts (EJE). Expert judgment records that were collected solely during the course of this research for testing aggregation methods applied to EJE are referenced as Test Data (TD), which is described in Chapter 8. The EJE source consists of 606 TUD records and 1,182 UMD records for a total of 1,788 records. The TD table consists of 37 records. The EJE and TD records constitute the Expert Judgment Data Sources (EJDS) database constructed for this research.

2.2 Description of Data Sources

The TUD data source is described in Cooke and Goossens (2008) and has been used in peer-reviewed studies including Lin and Cheng (2008), Lin and Bier (2008), Boone et al. (2009), Lin (2011), and Eggstaff, Mazzuchi, & Sarkani (2014). These data sources cover sectors such as nuclear applications, chemical and gas industry, groundwater transport, water pollution, dike ring, barriers, aerospace, occupational safety health, financial activities, volcanoes, and dams. The TUD data was provided under a proprietary basis by Professor R.M. Cooke. UMD data sources included Department of Mechanical Engineering graduate course work (Mosleh, personal communication, October, 2013) and the references and specific tables (reference Appendix A: EJE and TD Referencing) cited in two dissertations; Forrester (2005) and Shirazi (2009).

Within the EJE table, the term *theme* is used to describe a set of *records* that relate to a common topic area, e.g., industrial accidents or information security. The themes are unique to the TUD and the UMD sources. This term was introduced to prevent confusion with the term *case* used in the TUD source per Cooke and Goossens (2008), and the TUD data repository, EXCALIBUR. The mapping of cases and themes is provided in Table 1: TUD Case Names and EJE Theme Names Mapping. This distinction was necessary since not all TUD records were incorporated in EJE due to the reasons discussed Section 2.3.

Table 1: TUD Case Names and EJE Theme Names Mapping

TUD Case Name	EJE Theme Name
A_SEED	A_SEED
ACNEXPTS	ACNEXPTS
AOTDAILY	AOTDAILY
AOTRISK1	AOTRISK1
ATCEP Error 5 experts 31 jan 08	Aviation
FCEP Error 5 experts 31 jan 08	
pilots	
BSWAAL	BSWAAL
dams	dams
DEPOS1	DEPOS1
DISPER1	DISPER1 (amalgamated with TNODISP1)
TNODISP1	
EXP_DISP	EXP_DISP
EXP_WD	EXP_WD
GL-invasive-species	GL-invasive-species
CARMAExpStudy	Greece_NL_CARMA
CARMA-Greece-Assessments	
GROND5	GROND5
INFOSEC	INFOSEC
Ladders	Ladders
NH3EXPTS	NH3EXPTS
ONINX	ONINX
OpRiskBank	OpRiskBank

TUD Case Name	EJE Theme Name
PBEARLYH	PBEARLYH
PBINTDOS	PBINTDOS
pm25	pm25
RETURNafter	RETURNafter
S_SEED	S_SEED
SO3EXPTS	SO3EXPTS
ESTEC1	Space Flight Risk
ESTEC-2	
ESTEC-3	
MONT1	Volcanoes
Volcrisk	

Table 2: EJDS: Summary-level Listing of Record, Prediction and Theme Count

provides a summary-level listing of the record, prediction and theme count in EJDS.

Although the number of themes for TUD and UMD totals 52, there are 43 unique themes.

Specifically, there are nine themes that have both probabilistic and physical data

(Aviation, dams, Greece_NL_CARMA, INFOSEC, Ladders, pm25, Space Flight Risk, UMD Campus, and Volcanoes).

Table 2: EJDS: Summary-level Listing of Record, Prediction and Theme Count by Data Source

Record, Predictions and Theme Counts by Data Category	TUD	UMD	EJE Total	TD	EDJS Total
Number of records in Physical Category	540	1,181	1,721	11	1,732
Number of records in Probabilistic Category	66	1	67	26	93
<i>Subtotal Number of Records in Both Categories</i>	<i>606</i>	<i>1,182</i>	<i>1,788</i>	<i>37</i>	<i>1,825</i>
Number of predictions in Physical Category	4,661	1,445	6,106	130	6,236
Number of predictions in Probabilistic Category	516	13	529	60	589
<i>Subtotal Number of Predictions in Both Categories</i>	<i>5,177</i>	<i>1,458</i>	<i>6,635</i>	<i>190</i>	<i>6,825</i>
Number of themes in Physical Category	27	16	43	5	48

Record, Predictions and Theme Counts by Data Category	TUD	UMD	EJE Total	TD	EDJS Total
Number of themes in Probabilistic Category	8	1	9	5	14
Subtotal Number of Themes in Both Categories	35	17	52	10	62

References made to specific EJE records within this research are denoted as SeqID (Sequence ID) which is used as a prepend to the type of data, i.e., Physical or Probabilistic and the specific record number. Accordingly, SeqID EJEPHYS(1 to 1721) and SeqID EJEPROB(1 to 67) are used to describe EJE Physical Data and Probabilistic Data respectively. Analogously, specific TD records are referenced as TDPHYS(1 to 11), and TDPROB(1 to 26). Appendix A: EJE and TD Referencing provides a listing of the SeqIDs assigned to each TUD case, and EJE and TD theme.

2.3 Data Sources Preparation

Appendix A: EJE and TD Referencing also identifies the source for each EDJS theme. With respect to the UMD themes, other than PHD Surveys (Forrester, 2005), UMD Campus, and CALCE Experts, all data were retrieved from the source specified in Appendix A: EJE and TD Referencing.

The TUD data were provided under a proprietary agreement via receipt of the EXCALIBUR application and data files from Professor R. Cooke (personal communication, April, 2011). Their source is accordingly denoted as EXCALIBUR in Appendix A.

The following data processing conditions were applied to the TUD data:

1. Only TUD records that contained realized values were considered as potentially suitable records for EJE

2. Only the 5th, 50th (medians), and 95th percentiles were selected from the TUD records
3. Duplicated records were rejected
4. Analytical research and literature reviews enabled the determination as to whether data were probabilistic or physical, and traceability between TUD data and Cooke and Goossens (2008).

With respect to condition 3, elimination of duplicate records was not a straightforward matter of data de-duplication. For example, the Expert Names for the CARMA-Greece-Assessment and CARMAExpStudy cases are not duplicated in Greece_NL_CARMA case. However, the Variable Name, Scale, and the values for the 5th, 50th and 95th percentiles when the CARMA-Greece-Assessment and CARMAExpStudy cases are combined are identical to the corresponding variables in the Greece_NL_CARMA case. The data for the CARMA-Greece-Assessment and CARMAExpStudy cases were uploaded to the EJE.

As an illustration of condition 4, one of the TUD data cases per EXCALIBUR is ONINX which corresponds to Case #30 Dike Ring Failure in Cooke and Goossens (2008). The mapping was ascertained via a review of Cooke and Slijkhuis(2003) and Lin and Bier (2008). This review also confirmed that the data were physical.

In addition to a SeqID, each EDJS record is uniquely identified by a variable name. In the Dike Ring Failure case, as analyzed by Lin and Bier (2008), the elicitation questions are represented as variables, the responses to which are predictions. Three of these questions ask for actual flow rate on occasions when the calculated flow rate is 1,

10, and 100 liters per second per meter, and are represented by variables Mo1, Mo2, and Mo3 respectively.

All EJDS records were formatted into a defined structure. This structure can be illustrated using Forrester's (2005) use of data cited in Walker, Catalano, Hammitt, and Evans (2003):

1. SeqID: EJEPHYS195
2. Theme: Benzene Concentration
3. Variable: Ambient Benzene Concentrations
4. Number of Obs: 7 (this value represents the number of experts)
5. Value: 3.6 (the realized value of the variable)
6. $e'_1-e'_7$: 3.2; 3.2; 3.7; 3.9; 4.6; 5.8; 5.7 (the predictions provided by the experts).

The EDJS records are stored in an Oracle 11g Express database.

Chapter 3: Research Question 1 (Physical versus Probabilistic Expert Estimates)

3.1 Introduction

Does the type of quantity estimated, "physical"—variables having units of mass, time, etc.—or "probabilistic"—variables representing likelihood of an event, or a frequency of occurrence—impact the accuracy of elicited predictions? A quantitative Figure of Merit (FOM) is the ratio of realized value to predicted median, e/e' . A derived metric, Maximum Multiplicative Error, or MME is the maximum of this ratio and its inverse. Percentiles, standard deviations or ratios of physical quantities were considered to be physical. For example, a physical variable in the meta-database is maximum pumice clast dimension in mm; a probabilistic variable is the likelihood that an attack on a computer information system will be successful. Given a point estimate elicited from an expert, what multiplicative factors should be applied to that estimate, in order to bound it in an interval—with a corresponding level of probability? It will be shown that the type of quantity estimated—physical or probabilistic—does indeed make a difference in prediction accuracy; and that bounding intervals differ for the two types.

3.2 Research Question 1 Significance

As stated in Chapter 1, expert judgment opinions (or estimates) are used by decision makers as inputs to decisions. Since there is a cost associated with inaccurate predictions, the fact that the type of quantity estimated—physical or probabilistic—impacts accuracy assists decision makers in assessing the risk associated with the decisions. An example of a physical variable in the EJE database is maximum pumice

last dimension in mm; a probabilistic variable is the likelihood that an attack on a computer information system will be successful.

Decision-makers rely on expert opinions to shape their decisions – and it is clear that there may be important cost implications when poor decisions are made as a result of inaccurate predictions. By exploring whether the quality of predictions depends in part on the type of quantity being estimated, it might be determined that some types of expert opinions ought to be treated differently than others. Indeed, knowing that different types of quantities – physical or probabilistic – may impact the accuracy of predictions could help decision-makers better evaluate expert predictions and thereby better assess the potential risks of a particular decision.

3.2.1 Justification of metric

Financial planning activities in private and public sector organizations strive to avoid predictions having large multiplicative excursions on either side of the realized value. For example, the California Department of Transportation (2007) states that its “goal is to avoid project cost overrun and also avoid excessive cost underrun. Cost overrun leads to shortage of funding to deliver the project, while cost underrun leaves unused funds that could have been used to deliver other important projects.”, p. 20-4). The use of MME as a metric penalizes multiplicative excursions of estimates on either side of the realized value equally. The following example illustrates this premise. Consider a proposed project where the benefit, B is fixed at 3 units, but costs, C are uncertain, with a point estimate of 2 units. A cost overrun of factor of two high, i.e., $MME=2$, causes the Benefit/Cost (B/C) ratio to be half of what was claimed (e.g., 0.75 instead of 1.5 initially claimed). Similarly, for a project where cost is fixed at 2 units, but

benefits are uncertain, with a point estimate of 3 units, scaling linearly according to the effectiveness of a proposed safety system, a factor of two underrun in the latter, i.e., MME=2, causes the B/C ratio to be half of what was claimed. Neither project should have been approved based on benefit cost ratio; funds would have been better spent on alternative projects.

Additionally, for the EJE database, no negative values are present. If absolute percentage error were used instead of MME, the largest excursion associated with an underestimate would be 100%, whereas the largest excursion associated with overestimates could exceed 10^7 . However, the consequences of underestimates can be as serious as those associated with overestimates. Use of MME is an approximation which treats them equally.

3.3 Research Question 1 Literature Review

Aitken, Roberts, and Jackson (2010) defined probability as “a branch of mathematics which aims to conceptualise uncertainty and render it tractable to decision-making”, (p. 14). The importance of estimating probabilities, according to Huer (1981), is that “Social, political, military, and economic developments are not rigidly determined but occur or fail to occur with some degree of probability” (p.301). Nicholls (1999) asserted that one of the impediments in using meteorological forecasts correctly is attributable to “our difficulties in quantifying and dealing with probabilities, uncertainty, and risk” (p. 1386).

Martz et al. (1985) examined 20 subjective expert elicitations for each of 48 Los Alamos employees belonging to a “relatively homogeneous group” in terms of education and work activities. One of the conclusions was:

There was a consistent tendency on the part of subjects to underestimate the amount of uncertainty in their estimates. This was true for all response modes and all forms of questions (stimuli), but was worst for stimuli relating to probability or chance estimates. Relative errors were also worst for those stimuli. (p. 12)

Slovic (1987) observed that such difficulties in quantifying risks can sometimes cause risks to be overestimated or underestimated and asserted that “Experts’ judgments appear to be prone to many of the same biases as those of the general public, particularly when experts are forced to go beyond the limits of available data and rely on intuition”, (p. 281). Additionally, by way of an example, Aitken et al. (2010) observed that “There is, in short, no group of professionals working today in the criminal courts that can afford to be complacent about its members’ competence in statistical method and probabilistic reasoning” (p. 4).

Arguably, such competence in probabilistic reasoning may be associated with challenges in teaching probability. Garfield and Alhgren (1988) observed that “Inadequacies in prerequisite mathematics skills and abstract reasoning” (p. 43) may contribute to such challenges. Zientek, Carter, Taylor, and Capraro (2011) noted that until the 1980s, teaching of probability was “not viewed as important in primary and secondary education” (p. 25). Batanero and Díaz (2012) argued that lack of adequate teacher preparation in the field of probability due to time pressures undermine its significance. Ben-Tvi and Garfield (2008) asserted that “most students and adults do not think statistically about important issues that impact their lives” (p. 356), citing research compiled by Kahneman, Slovic, and Tversky (1982) that examines why specific heuristics are inconsistent with correct statistical thinking.

Tversky and Kahneman (1974) identified three heuristics, specifically, representativeness, availability, and adjustment from an anchor, which are used in making judgments under uncertainty, and the cognitive biases attributable to such heuristics. According to the authors, although these heuristics are useful, “sometimes they lead to severe and systematic errors” (p. 1124). The authors summarized these heuristics as follows:

- (i) Representativeness, which is usually employed when people are asked to judge the probability that an object or event A belongs to class or process B;
- (ii) Availability of instances or scenarios, which is often employed when people are asked to assess the frequency of a class or the plausibility of a particular development; and
- (iii) Adjustment from an anchor, which is usually employed in numerical prediction when a relevant value is available (p. 1131).

Tversky and Kahneman (1974) described cognitive biases derived from the reliance on judgmental heuristics as “not attributable to motivational effects such as wishful thinking or the distortion of judgments by payoffs and penalties.” (p. 1130). According to the National Institute of Standards and Technology (2012) “Cognitive bias results from computational trade-offs carried out in the brain and is not a conscious act or an act that can be avoided at will” (p.11). Huer (1981) examined such biases in estimating probability in deception/counter-deception activities and concluded that such estimates are biased by availability and anchoring heuristics.

The availability heuristic is reflected in a tendency for watch officers to “overestimate the probability of whatever they are watching for. Cases of deception are more memorable, hence more available, than instances when deception was not employed” (Huer, 1981, p. 315). With respect to anchoring, Huer (1981) identified that adjustment from the initial starting point was generally insufficient, because it is “easier to reinforce a target's existing preconceptions than to change them” (p. 315).

Hogarth (1975a) observed that “man has a well-established tendency to avoid using the ends of the probability scale. Small probabilities are generally overestimated and large probabilities underestimated”, (p. 274). Furthermore, “man frequently just ignores uncertainty” (Hogarth, 1975a, p. 273). Winkler (1975) issued a comment to Hogarth (1975a) and questioned whether “uncertainty is ignored or whether it is simply considered implicitly rather than explicitly” (p. 291). Hogarth (1975a) also discussed response mode: “Responses in probabilistic tasks can be seen to vary, e.g., as a function of whether probability is directly estimated on a scale of from zero to one, as opposed to the use of ‘odds’ or even ‘log-odds’” (p. 276).

Edwards (1975) commented that Hogarth’s (1975a) assertion regarding response modes was questionable in that if “the processes being hypothesized were really stable, regular features of human intellectual life, should they not be insensitive to the language in which the questions are phrased” (p. 293). In a rejoinder Hogarth (1975b) asserted that “evidence suggests that the tendency to avoid uncertainty is not some form of mental aberration but present in all of us to a greater or lesser extent” (p. 291).

According to Gayer (2015) “one of the most-documented biases people exhibit is that they tend to overestimate low probability risks of death (such as the risks of

botulism, lightning strikes, and natural disasters) and they tend to underestimate high probability risks of death (such as the risks of stroke, cancer, and heart disease” (Keynote Address, Brookings Institute). Nunes and Kirlik (2005) noted that a consistent feature of research regarding probability judgments is “firstly, the overestimation of low probability events, and secondly, consistent underestimation of high frequency events” (p. 422). Additionally, Israelski (2010) observed that there is a tendency for over-estimation of the “true probability of events viewed as favorable and under-estimation of events viewed as unfavorable” (p. 53). Barberis (2013) asserted “why people over- or under-estimate the likelihood of tail events; and of why they over- or underweight these events in their decision making [...] remains an open question” (p. 615).

According to Hilbert (2011), conservatism “refers to the experimental finding that people tend to underestimate high values and high likelihoods/probabilities/frequencies and overestimate low ones” (p. 14). The author further distinguished between “non-normalized” numbers, e.g., an interval, in which it is difficult to detect a conservatism bias, and probabilities, which are normalized between 0 and 1. This normalization enables the definition of “high (close to 1) and low (close to 0) when estimating likelihoods/probabilities/frequencies” (Hilbert, 2011, p. 14).

Hilbert (2011) argued that empirical studies differentiate between two methods for obtaining subjective probability estimates. These two methods are (1) likelihood and probability estimations and (2) and frequency probability of each estimate. The former addresses questions such as what is the likelihood of an event expressed as a percentage; the latter is used to compute for example, how often an event occurs. Tversky and Kahneman (1982) denoted these two methods as “singular” and “distributional” modes of

judgment respectively, and noted that “There are many instances in which the same question can be approached in either singular or distributional mode” (p. 518). Further according to the authors, “The distributional mode of judgment is more likely than the singular to yield accurate estimates of probability” (Tversky & Khaneman, 1982, p. 517).

Given the particular difficulties associated with eliciting expert judgment probabilistic type data, and the fact that previous meta-data informed aggregation techniques do not appear to distinguish between probabilistic and physical data types, there appears to be a research gap in this area.

3.4 Research Question 1 Mathematical Formalism

The measure of prediction accuracy used was the ratio $r \equiv e/e'$ of realized value to predicted value. For the TUD subset of EJE data, the predicted median value was used. For the UMD subset of EJE data, the point estimate prediction value was used. In considering whether to define error using “the difference between the measurement and the truth ... or ... the ratio between the two” (Tian, 2013), scatterplots of $(e' - e)$ versus e , and $\ln(e'/e)$ versus $\ln(e)$ for probabilistic data, were examined. The respective scatterplots are shown in Figure 1: Scatterplot of $e' - e$ vs. e and Figure 2: Scatterplot of e'/e vs. e .

Figure 1: Scatterplot of $e' - e$ vs. e

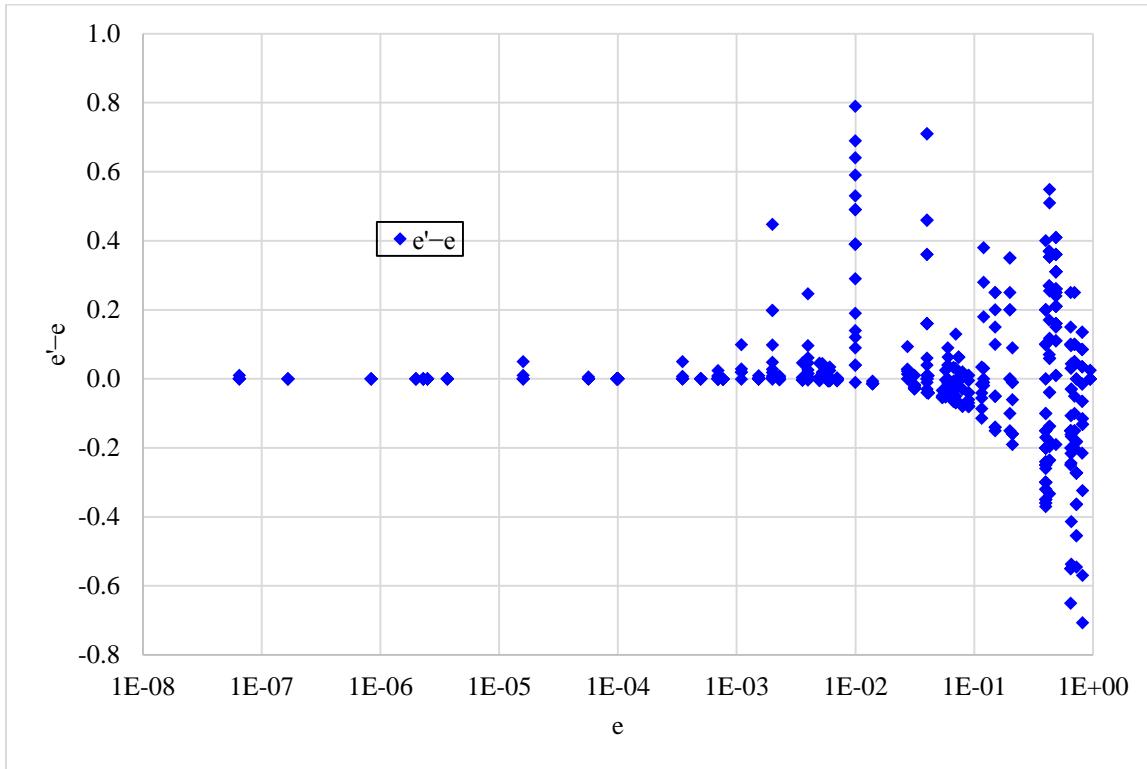
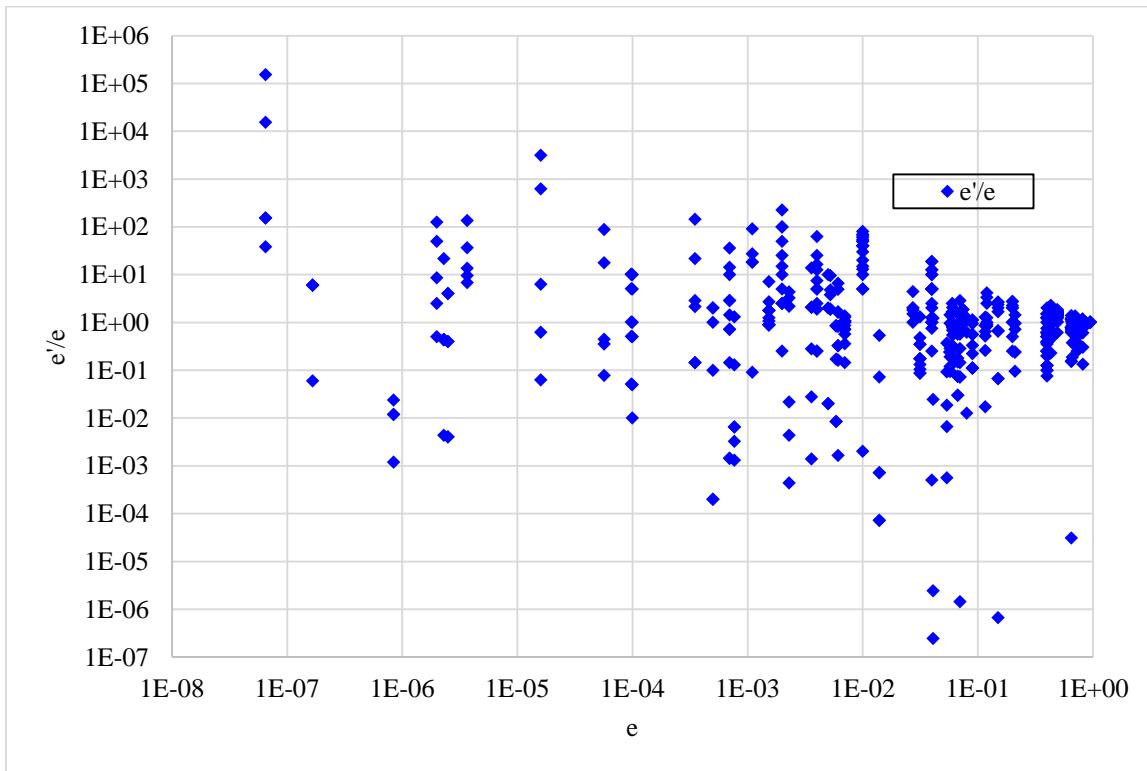


Figure 2: Scatterplot of e'/e vs. e



The multiplicative (ratio) scatterplot (Figure 2) highlights large discrepancies between realized and predicted values. Consider for a moment how economically significant the impact of such differences would be if one data point or the other were relied upon by a decision-maker. For example, the four points closest to the left edge of Figure 2 clearly represent large multiplicative discrepancies—as large as a factor of $1.5 \cdot 10^5$ —between predicted (10^{-2} , 10^{-3} , 10^{-5} , and $2.5 \cdot 10^{-6}$) and actual ($6.5 \cdot 10^{-8}$) probability of runway overrun on takeoff.

In the scatterplot based on the differences between e' and e (Figure 1) however, these four data points appear as a single undistinguished pair of points close to the line $e'-e=0$, near the left edge of the graph. This means that a decision-maker relying on a difference-based metric would miss the economically significant discrepancy between predicted and actual values.

Similarly, the ratio plot, Figure 2, highlights six anomalous points in the lower right corner of the plot, having ratios less than 10^{-4} , and running counter to the funnel pattern formed by the other ratios as e increases from 0.01 to 1. All of these six points except one arose from a single theme, “Space Flight Risk”, which may have been subject to different interpretations by the experts used for elicitation in this area. Since the space flight risk anomalies have $e'-e$ values ranging between -0.01 and -0.15 , and are undistinguished from other points in the difference-based scatterplot having e values between 0.01 and 1, and $e'-e$ values close to zero, a decision-maker relying on a difference-based metric would miss the significant discrepancy between predicted and actual values.

These observations argue for using a ratio rather than a difference when assessing the reliability of predictions in the meta-data. The EJE physical and probabilistic meta-data points span approximately twenty orders of magnitude; the ratio metric accommodates this wide range.

The ratio $r=e/e'$ of realized value to predicted value is the foundational measure of accuracy for this study. In order to assess the relative accuracy of estimates for physical and probabilistic variables, the cumulative distribution function of the ratios was computed for the two data types. In particular, the CDF of the ratios was computed for the subset of physical predictions for which the realized value, e was less than one. This subset will be referred to as the *physical subset*. This CDF was compared to the CDF of the ratios for probabilistic predictions. To compute a CDF, each prediction in each data category was assigned a weight. Consider first the physical subset category. To compute the CDF of r for the physical subset, each e' prediction associated with a given variable in a given theme was assigned a weight, $w_{e'}$ in a manner that allowed for equal weighting of all e' associated with the variable, all variables, within the theme, and all themes within the subset:

$$w_{e'} = [n\text{themes}_{e<1} \cdot n\text{var}_{e<1} \cdot n\text{obs}]^{-1} \quad (\text{Equation 1})$$

where $n\text{themes}_{e<1}$ is the number of themes containing variables with $e<1$; $n\text{var}_{e<1}$ is the number of such variables in the given theme, and $n\text{obs}$ is the number of observations e' for a given variable.

The physical subset contained 19 themes having at least one variable for which $e<1$; a total of 194 records met this criterion. (Compared to the full physical data set

containing 43 themes and 1721 records, the subset comprised approximately 20% of the total mass.) An analogous formula was used to compute the weights $w_{e'}$ associated with each e' for probabilistic data. For this category, all values of e were already less than one, therefore all of its nine themes and 67 records were applicable.

For each prediction e' in a given data category, the ratio r was computed. The r values and their associated weights (the $w_{e'}$ associated with the predictions e') were sorted by increasing value of r . The cumulated weights at each such value of r corresponded to the CDF.

3.5 Research Question 1 Methodology and Results

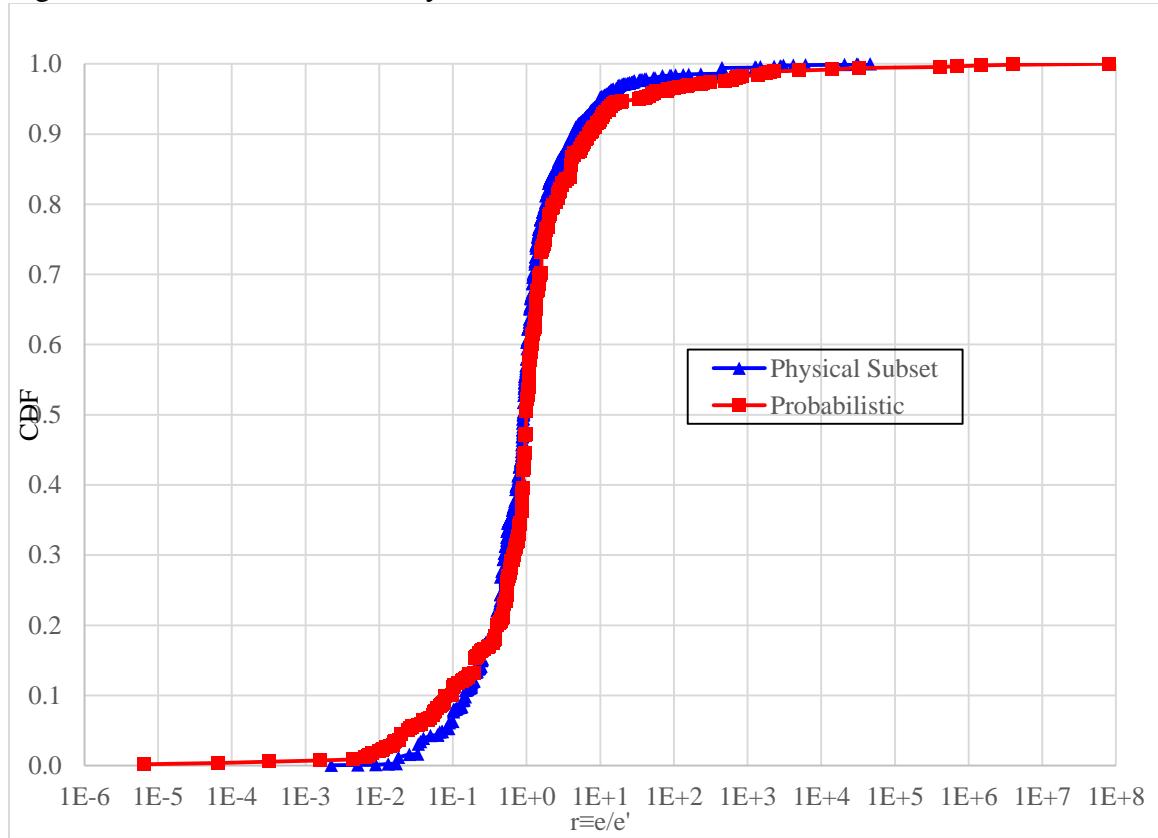
For each data category, weights were consolidated for successive r values which differed by less than 10^{-5} . These differences could arise from rounding errors in data processing and from repeated e' values. Over the full physical data set, they also arose from small differences in ratios of relatively large numbers. For example, the theme AOTDAILY, which includes 38 variables, involves predictions of next day Amsterdam stock market opening price. The AOTDAILY variable D060798 had e and e'_1 values 1242 and 1243, respectively; variable D070798 had e and e'_4 values 1246 and 1247, respectively. The corresponding e/e' ratios differed by approximately $3 \cdot 10^{-6}$, and were consolidated. The consolidation did not affect several small ratios, r less than 10^{-4} , since they differed from each other by more than 10^{-5} .

A single record from the Crop Yield theme (SeqID EJEPHYS 220) was excluded, as it was the only representative of this theme, which contained over 700 variables. Its “realized value” for e , 0.01, was an artifact employed to avert an infinite ratio for $\ln(e/e')$. The use of 0.01 reflected the fact that as a practical matter, there was no significant

economic difference between crop yields of zero and 0.01 bushels per acre. The single e/e' ratio of $0.000141=0.01/7080$ would have received an unreasonably large weight of approximately five percent.

After consolidation of weights, $n_1=822$ and $n_2=261$ unique (ratio, weight) pairs remained for the two data categories, respectively. Cumulating the weights yielded the CDFs for each category which are presented in Figure 3: CDFs of $r \equiv e/e'$ for Physical Subset and Probabilistic Data.

Figure 3: CDFs of $r \equiv e/e'$ for Physical Subset and Probabilistic Data



Given the CDFs, the probability that an elicited e' value over- or underestimates the realized value, e by a given factor is known. For example, the physical subset CDF at $r=0.1$ is 0.0627. This reflects an approximately six percent probability that the prediction overestimates the realized value by a factor of ten or more. The corresponding value for

the probabilistic CDF at this same value of r is 11.1 percent. This means that for probabilistic data, there is roughly twice the likelihood of overestimating the realized value by a factor of ten or more. Similarly, the physical subset CDF at $r=10$ is 0.9517. This means there is a probability of approximately 4.8 percent that the realized value is underestimated by a factor of at least ten.

The probabilistic CDF at $r=10$ is 0.9193. This means there is a probability of approximately 8.1 percent that the realized value is underestimated by a factor of at least ten; again, roughly twice the likelihood compared to physical data.

Figure 4: **Overestimation Probabilities by Factor for Physical Subset and Probabilistic Data** gives the probability of overestimating e by a given factor for the two data categories, for selected factors. Figure 5: Underestimation Probabilities by Factor for Physical Subset and Probabilistic Data depicts the analogous underestimation probabilities.

Figure 4: Overestimation Probabilities by Factor for Physical Subset and Probabilistic Data

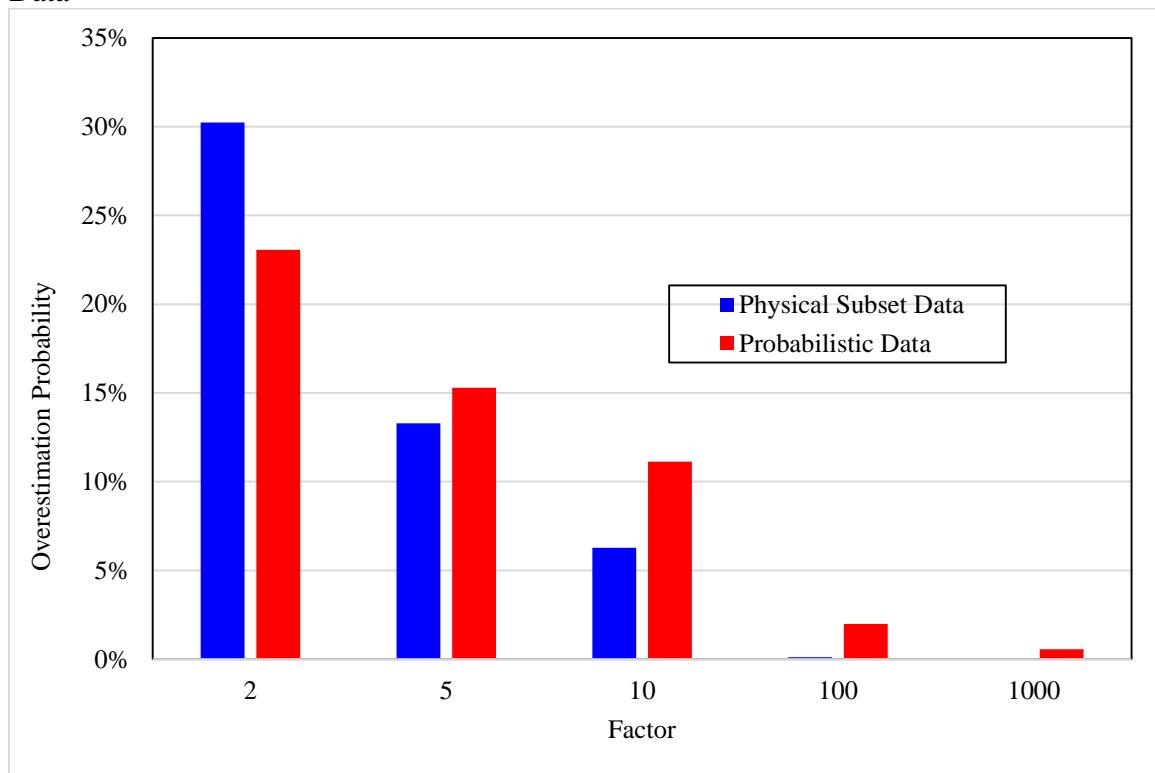


Figure 5: Underestimation Probabilities by Factor for Physical Subset and Probabilistic Data

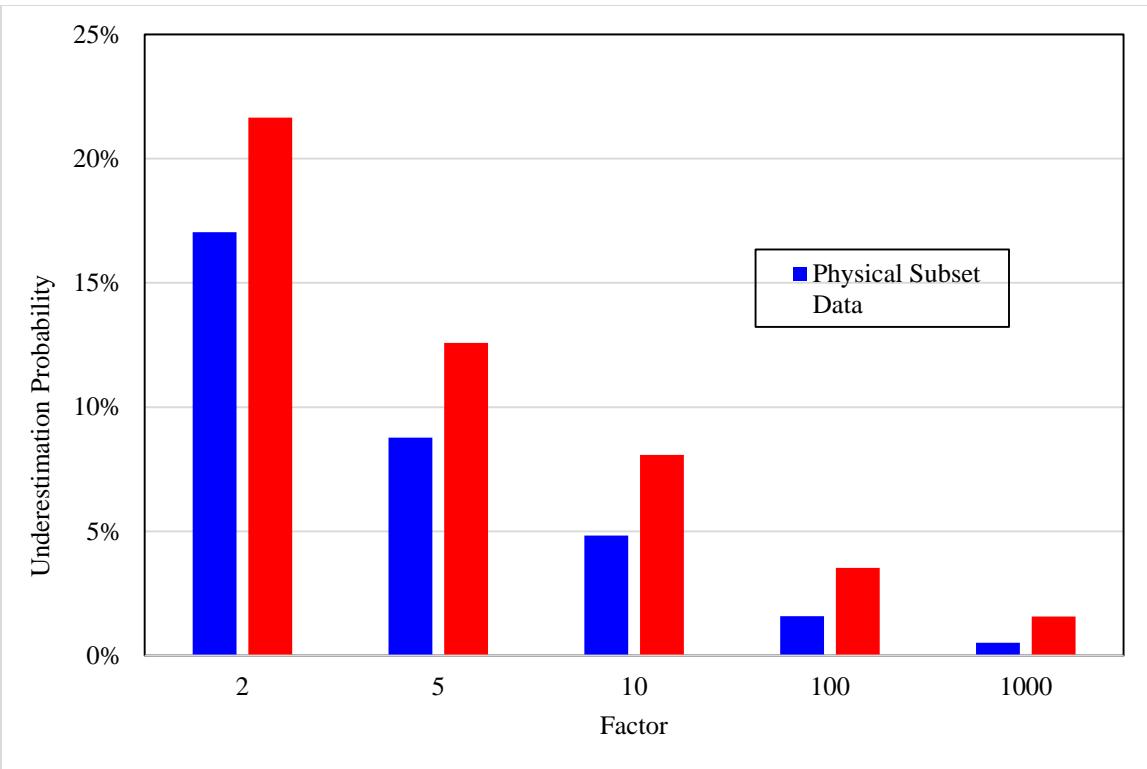


Figure 4 and Figure 5 demonstrate that while rare, overestimation errors of a factor of 100 or more are far more likely to occur for probabilistic data (2% chance) than for physical data (0.1% chance). For underestimation errors, the disparity is much less—about a factor of two for the two data types (3.5% versus 1.6% for probabilistic and physical subset data, respectively). For smaller factors ranging from 10 down to 2, the disparity is much less: approximately 1.5 for underestimation errors. For overestimation errors, at the factor of two level, physical subset data is *more* likely to be overestimated than probabilistic data (30% versus 23%, respectively). Probabilistic data is 1.2 and 1.8 times more likely to be overestimated than physical data at the factor of five and factor of ten levels, respectively.

Although the two CDFs were computed analytically, it may be noted that if they are treated as empirical CDFs, and the Kolmogorov-Smirnov (K-S) two-sided non-parametric test for equality of distributions is applied, the null hypothesis H_0 that there is no difference between the distributions of the ratios e/e' for the two data categories would be rejected.

The reason for the rejection is that the largest absolute difference, D between the two CDFs is 0.113, occurring at $r \approx 0.95$, which exceeds the critical value for the K-S test. The numbers of pairs are sufficiently large to justify using the large-sample approximation for the K-S two-sided test. The critical value $D_{crit} = C_\alpha [(n_1+n_2)/(n_1 \cdot n_2)]^{0.5}$, (reference Critical Values for the Two-sample Kolmogorov-Smirnov test (2-sided) (n.d.)). For a level of significance $\alpha=0.05$, $C_\alpha=1.36$, $n_1=822$ and $n_2=261$, $D_{crit} \approx 0.097$. Since $D > D_{crit}$, H_0 is rejected.

The null hypothesis of no difference between the distributions would also have been rejected if the K-S test had been applied to unweighted data, i.e. to physical subset and probabilistic CDFs constructed giving all ratios e/e' equal weights, regardless of the theme or variable to which they belong. In this case, the difference, D increases to 0.13 (at $r=0.1$), which is significant at the $\alpha=0.01$ level of significance. ($D_{crit} \approx 0.116$).

The test for equality of binomial proportion is an additional statistical test which could be applied to the two sets of unweighted ratios. If the likelihood of overestimating by a factor of ten is compared for physical subset and probabilistic data, using unweighted but consolidated ratios, the following results are obtained:

1. There are 22 out of the 822 physical subset ratios that are less than 0.1, while there are 41 out of the 261 probabilistic ratios which are less than 0.1.

2. Applying the large sample test for equality of binomial proportion per

<http://itl.nist.gov/div898/software/dataplot/refman1/auxiliar/binotest.htm> yields:

$$Z = (\hat{p}_1 - \hat{p}_2) / [(1/n_1 + 1/n_2) \cdot \hat{p}(1-\hat{p})]^{0.5} = -7.8, \quad (\text{Equation 2})$$

where $n_1=823$, $n_2=261$, $\hat{p}_1=22/n_1=0.027$, $\hat{p}_2=41/n_2=0.157$, $\hat{p}=(n_1\cdot\hat{p}_1 + n_2\cdot\hat{p}_2)/(n_1+n_2)=0.058$, and the test statistic Z is a standard normal.

3. This value of Z rejects the null hypothesis H_0 of no difference between proportions, at a level of significance $\alpha << 0.01$.

The above calculation is conservative, since it ignores the fact that ratios can be consolidated due to multiple instances of experts predicting the same e' value against a given realized value e . If weights are to be ignored, the ratios should not have been consolidated.

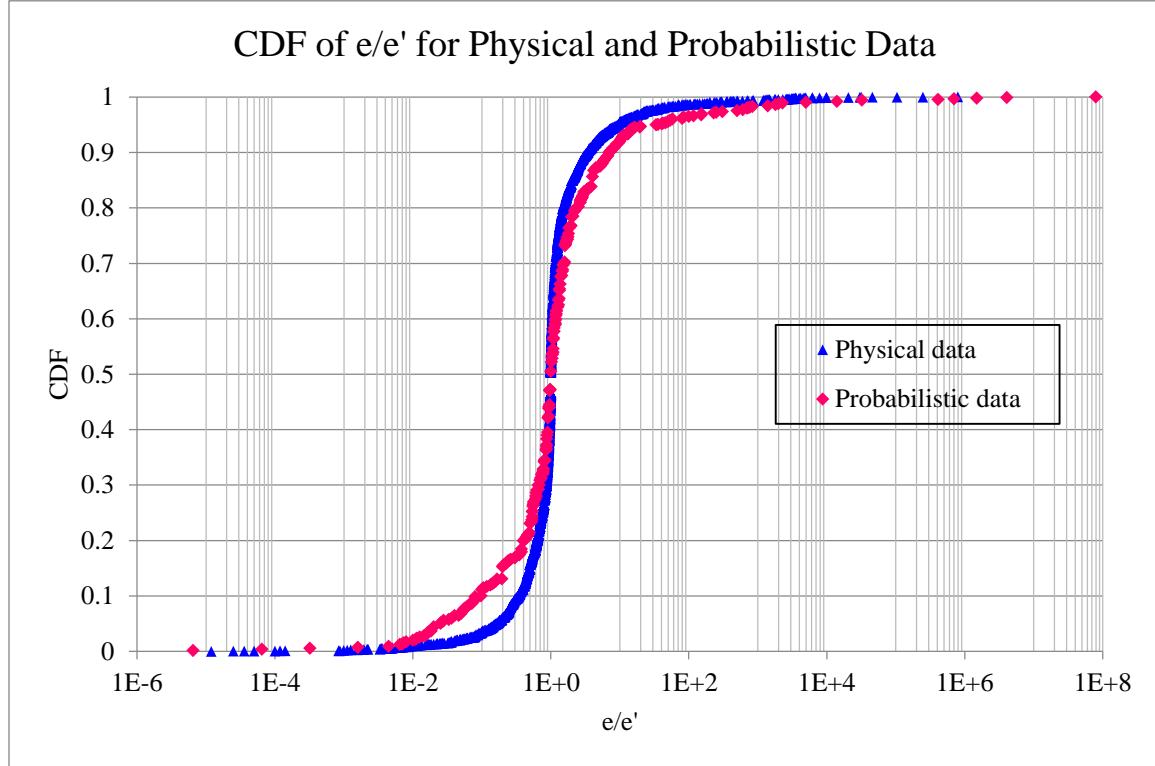
Accordingly, without consolidation, there are $n_1=1291$ physical subset ratios, for which $\hat{p}_1=23/n_1=0.0178$ have values less than 0.1. There are $n_2=528$ probabilistic ratios, for which $\hat{p}_2=59/n_2=0.1117$ have values less than 0.1. The pooled $\hat{p}=n_1\cdot\hat{p}_1+n_2\cdot\hat{p}_2=0.0451$, and $Z=-8.8$, which represents an even greater excursion than the previous result $Z=-7.8$.

Based on the above discussion, it can be concluded that there is a significant difference between the distributions of the physical data subset and probabilistic data.

3.5.1. Comparison of Prediction Accuracy between Full Physical and Probabilistic Datasets

Having established that a difference exists between the physical subset and probabilistic data, the CDF of the ratio e/e' for the entire physical data set was computed and compared to the CDF of the probabilistic data set. The calculation was analogous to that used for the physical data subset. All 43 physical themes and all variables within each theme were included. After consolidation of ratios differing by less than 10^{-5} , there were $n_1=3,403$ unique (r, weight) pairs for physical data. For probabilistic data, there was no change, as all $e<1$: there were $n_2=261$ unique (r, weight) pairs. The resulting CDFs are shown in Figure 6: CDFs of e/e' for Physical and Probabilistic Data.

Figure 6: CDFs of e/e' for Physical and Probabilistic Data



As before, the probability that an elicited e' value over- or underestimates the realized value, e by a given factor is specified by the CDF.

Figure 7: Physical Data and Probabilistic Data – Overestimation Probability **by Factor** gives the probability of overestimating e by a given factor for the two data categories, for selected factors. The analogous probabilities of underestimating e by those same factors are illustrated in Figure 8: Physical Data and Probabilistic Data – Underestimation Probability by Factor.

Figure 7: Physical Data and Probabilistic Data – Overestimation Probability by Factor

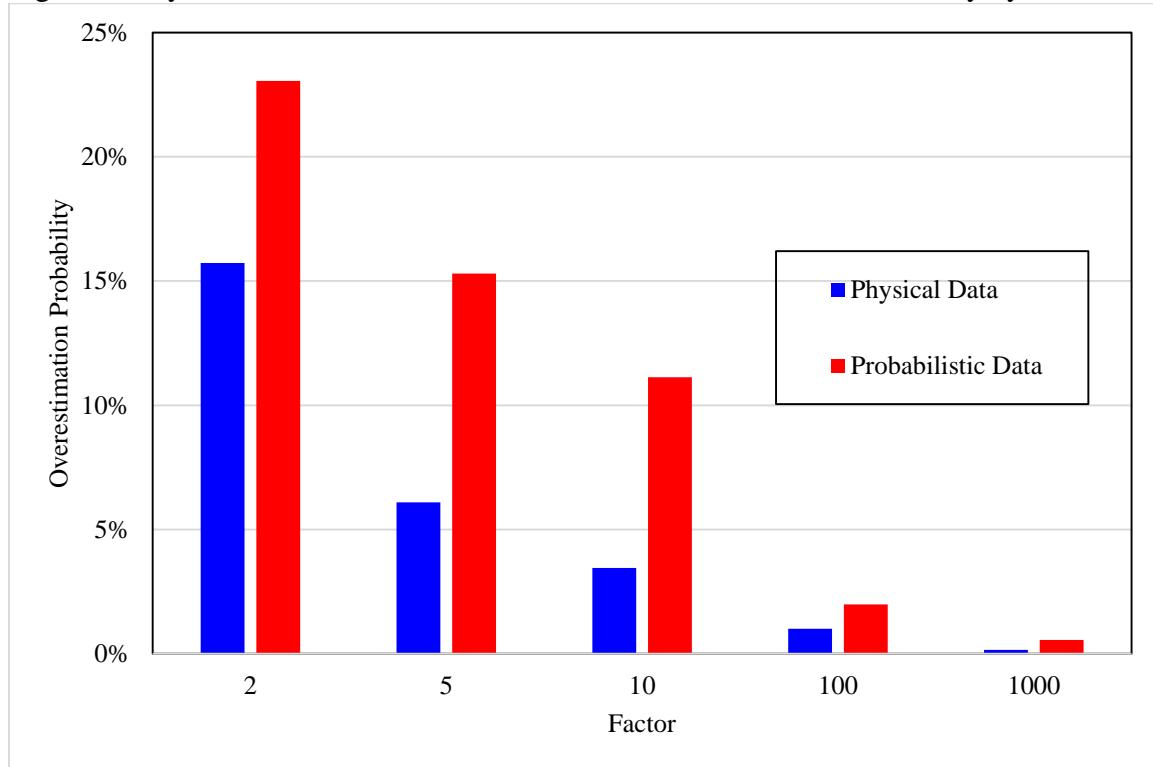
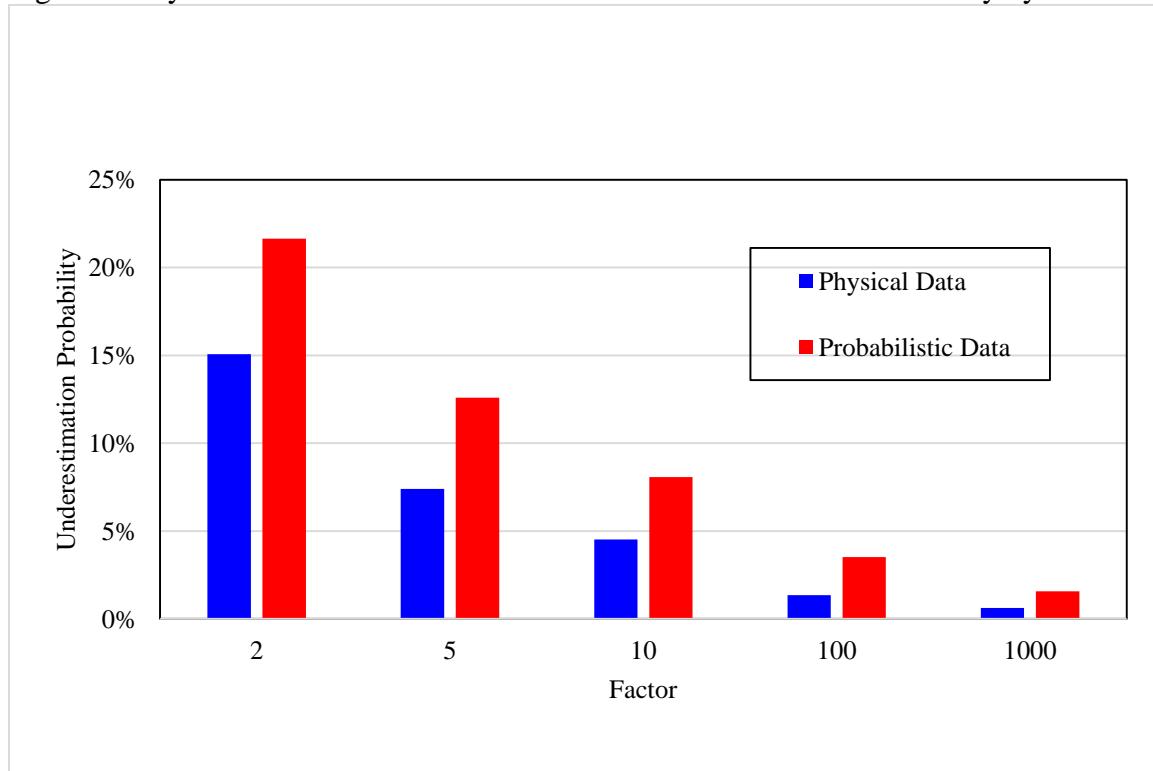


Figure 8: Physical Data and Probabilistic Data – Underestimation Probability by Factor



The individual values appear in Table 3: Chances of Under and Over-Estimation given a single prediction, e' before aggregation.

Overestimation errors by factors of 2, 5, and 10 are approximately half as likely (to within ten percent) to occur for the full physical data set as for the physical data subset. However, very large overestimations are more likely to occur over the full set. For example, the probability of a factor of 100 overestimation is 1% for the full set, while it is only 0.1% for the subset. The probabilistic data overestimation errors range from 1.5 to 3.5 times more likely to occur, at each factor. The largest discrepancy is at a factor of 1,000.

The likelihood of overestimating by a factor of ten can be compared for the two data sets using the large sample test for equality of binomial proportion, with the following results:

1. There are $n_1=6106$ non-consolidated ratios for the full physical data set, of which $\hat{p}_1=173/n_1=0.028$ are less than 0.1.
2. As before, there are $n_2=528$ probabilistic ratios, of which $\hat{p}_2=59/n_2=0.1117$ are less than 0.1.
3. The pooled $\hat{p}=0.0350$, and $Z=-10$.
4. Thus, the statistically significant difference between the distributions persists for overestimation by a factor of ten.

Underestimation probabilities at each factor agree to within approximately 15% with their counterparts for the physical data subset. For example, the probability of a factor of two underestimation error is 17% for the subset, and 15% over the full physical

data set. The probability of underestimation ranges from 1.5 to 2.5 times more likely for probabilistic data as for physical data over these factors.

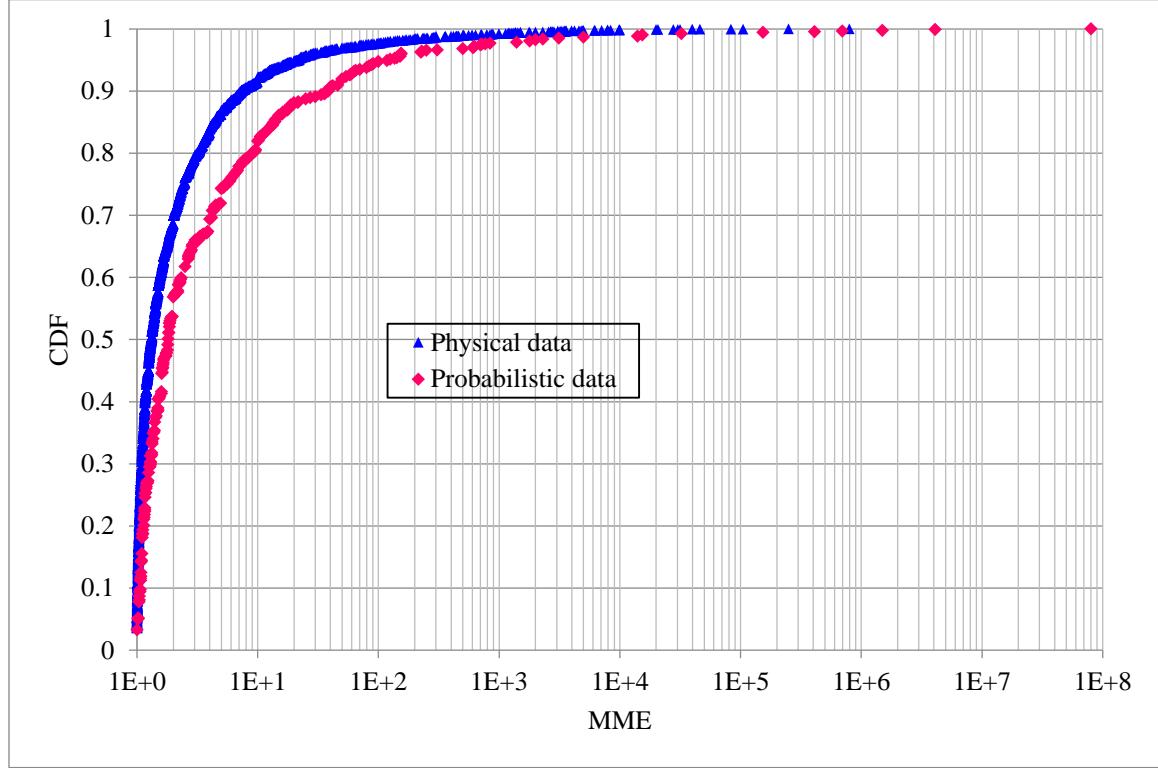
3.5.2 Maximum Multiplicative Error, MME

Over- and underestimation probabilities can be consolidated using the derived metric, MME. As previously stated, MME is the maximum of r and its inverse. Also as previously stated, use of MME as a metric penalizes multiplicative excursions of estimates on either side of the realized value equally, and represents an approximation that under- and over-estimates have equal consequences.

$$\text{MME} = \max(e'/e, e/e') \quad (\text{Equation 3})$$

The MME was computed for each (e, e') pair along with a corresponding weight, for each data type. The MMEs were sorted, and those differing by less than 10^{-5} had their weights consolidated. This left $n_1=3146$ and $n_2=244$ unique (MME, weight) pairs for physical and probabilistic data, respectively. The resulting CDFs are shown below in Figure 9: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data.

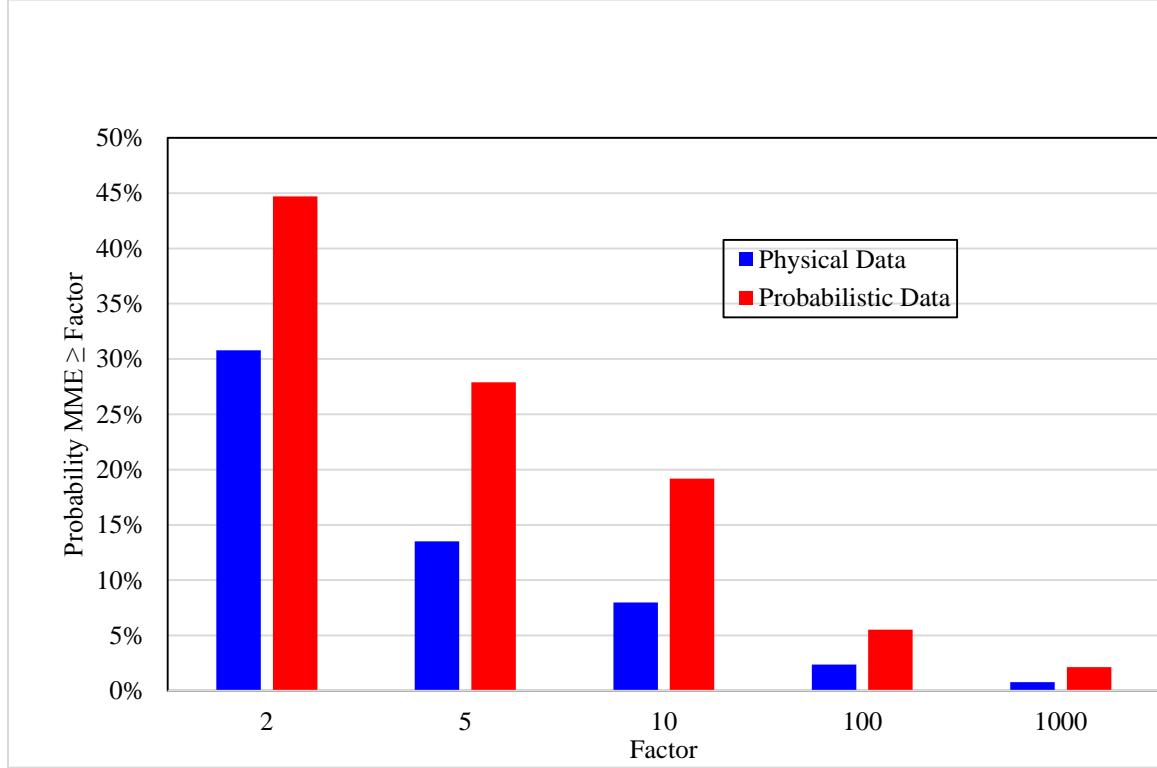
Figure 9: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data



The 80th and 90th percentiles of the CDF of MME for physical data are 3.2 and 7.3, respectively; the analogous percentiles for probabilistic data are 9 and 40, respectively. An interpretation of the latter figure, 40, is that there is a ten percent chance that a single prediction, e' will either overestimate or underestimate the realized value of a probabilistic variable, e , by a factor of forty or more. Similarly, there is a ten percent chance that a single prediction will either over- or underestimate the realized value of a physical variable, e by a factor of 7.3 or more.

The complement of the CDF at any value of MME gives the probability that multiplicative error will exceed that value. Exceedance probabilities at various factors were computed for the two data categories, and are provided in Figure 10: Physical Data and Probabilistic Data – MME Exceedance Probability by Factor.

Figure 10: Physical Data and Probabilistic Data – MME Exceedance Probability by Factor



It can be seen from the figure that at each factor, the probability of exceedance is greater for probabilistic than for physical data; the disparity averages about 2.2 (ranging from 1.5, at a factor of 2, to 2.7, at a factor of 1000). Note that for probabilistic data, the probability that MME will equal or exceed two is nearly fifty percent. This declines to approximately 20% at a factor of ten, and two percent at a factor of 1000. The corresponding values for physical data are 31 percent, eight percent and one percent, respectively

3.5.3 Bounding Intervals

An interval with an associated level of probability is more informative than a single point estimate in that the latter does not provide an indication about risk. Given a single point estimate prediction, e' , percentiles of the previously-developed CDFs for the

appropriate data type could be used to obtain intervals around e' , with associated levels of probability. For example, the 10th and 90th percentiles of the CDF of e/e' for physical data, and the 5th and 95th percentiles of the same CDF, can be used to obtain 80% and 90% bounding intervals for physical data e given e' : [0.340e', 3.45e'] and [0.165e', 9.84e'], respectively. An analogous procedure gives the corresponding 80% and 90% bounding intervals for probabilistic data e given e' [0.098e', 7e'] and [0.025e', 36e'], respectively. The 90% intervals for probabilistic data are approximately five times wider on each side of e' multiplicatively, compared to physical data; the 80% intervals are approximately 2.5 times wider.

Since probabilistic data is bounded by one, adjustments are necessary in cases where the upper bound, 7e', exceeds one. In this case, the lower bound associated with the two-sided 80% interval, 0.098e', may be used as a one-sided 90% lower bound. The one-sided 80% lower bound is obtained from the 20th percentile of the CDF of $r \equiv e/e'$ for probabilistic data, which occurs at $r=0.4$, hence there is an 80% chance that e will be at least 0.4e'. For cases where the upper 90% bound, 36e' exceeds one, the same lower bound associated with the two-sided 80% interval, 0.098e', may be used as a one-sided 90% lower bound, as before.

3.6 Summary

This chapter investigated whether there is a difference between physical and probabilistic data, in terms of the accuracy of predictions versus realized values. The FOM used was the ratio, r of realized value to predicted median value: $r \equiv e/e'$. The distribution of r for the two data types was computed, using an analytical weighting

scheme which assigned equal weights to all predictions, e' associated with a given variable; to all variables within a given theme; and to all themes within a given data type. First, only the subset of physical data for which the realized values e were less than one, were compared to probabilistic data.

A comparison of the CDFs of r for the two types revealed that the type makes a difference: probabilistic data predictions are less reliable than predictions for physical data. For example, there is a clear and persistent difference of approximately twice the likelihood of a factor of ten or greater overestimate of e by e' . This difference is applicable regardless of whether the probabilistic data type is compared to the physical subset only, or to the full physical data set.

It was noted that if the CDFs were treated as empirical, the maximum difference between them would be significant at the significance $\alpha=0.05$ level under the Kolmogorov-Smirnov test. Over- and under-estimation probabilities were presented for factors of 2, 5, 10, 100 and 1000.

The differences between the two data types persisted when the entire physical data set was used instead of only the subset. Over- and under-estimation excursions by the above-listed factors, were approximately twice as likely to occur for probabilistic data as for physical data. A metric consolidated the two types of excursions into a single derived FOM: the maximum multiplicative error, $MME = \max(e/e', e'/e)$. The CDF of MME for the two data types was obtained, and the probability that MME exceeded the above-listed factors was computed. This probability was nearly 50% at a factor of 2, for probabilistic data; 30% for physical data. The respective probabilities decline to approximately 5% and 2%, at a factor of 100.

Bounding intervals based on the above CDFs were presented for the two data types. For physical data, 80% and 90% bounds for e given a single prediction e' are $[0.340e', 3.45e']$ and $[0.165e', 9.84e']$, respectively. For probabilistic data, the respective bounds are $[0.098e', 7e']$ and $[0.025e', 36e']$ for the two data types, respectively. One-sided lower bounds were presented for probabilistic data, for situations where the upper bound would have exceeded unity. The 90% intervals for probabilistic data are approximately five times wider on each side of e' multiplicatively, compared to physical data; the 80% intervals are approximately 2.5 times wider.

These bounds cannot be used when predictions are available from more than one expert, which is generally the case. In this situation, an aggregation method must be used. One advantage of aggregation is that it yields smaller multiplicative errors between predictions and realized values. As an exercise, the first e' value was taken from each physical record. Over 1000 of the physical records have a single e' value. The above-referenced bounds factors were applied. It was found that the coverage of the 90% bounds was 89%. However, the weighted average MME was 1699. This exceeds by a factor of 100 the MME of 8 obtained using one method of aggregation, the median of the e' values. The exercise was repeated for probabilistic data, with similar results: the coverage was 91%, but the weighted average MME was 1913, exceeding by a factor of 100 the MME of 14 obtained via the median of the e' values.

With the exception of adjustments for probabilistic data where the upper bound would otherwise exceed unity, the multiplicative factors used in the above bounding intervals do not vary with the magnitude of e . However, it will be seen in Chapter 5 that a relationship exists between quantiles of e' , and the magnitude of e ; the spread between

upper and lower bounds in Ln-space is greater at very small values of e, for probabilistic data. For physical data, the spread is larger at both very large and very small values of e. This relationship was developed during the exploration of a number of different methods for aggregating individual e' predicted medians into aggregated estimates \hat{e} , with associated intervals.

Chapter 4: Research Question 2 (Aggregation Methods)

4.1 Introduction

Given a set of point estimates (e') elicited from experts, how should they be combined to yield an aggregate estimate and associated calibrated bounds?

This chapter compares a number of methods for aggregating individual expert point estimates, e' into a single aggregated estimate of the median, \hat{e} , in conjunction with calibrated bounds around \hat{e} . It will be generally demonstrated that aggregation produces higher quality estimates, in the sense that the multiplicative excursion between the estimate and e is less than that between individual e' and e . The associated ranking criterion is the weighted average MME.

A criterion reflecting calibration of the aggregation methods is the average absolute value of the difference between the coverage percentage over each theme, and 0.9 (the desired coverage percentage of 90%), hereinafter ABSDist. Suppose a particular method achieves a global coverage percentage of 90%, but actually has coverage of either 80% or 100% over each theme. The absolute distance criterion will penalize such a method, compared to one which has a more even coverage closer to 90% for each theme.

A third criterion reflects informativeness, in the sense that narrow bounds convey more information than wide bounds. Theoretically, an aggregation technique could achieve an excellent score on the ABSDist criterion by employing the following strategy: for a theme with $n=10$ records, set the bounds

for nine of the ten at $\pm\infty$; set the bounds for the tenth record at $(0.999999\hat{e}, 1.000001\hat{e})$. The first nine bounds would include e , the last would not, except in those rare cases where \hat{e} is already equal to e . Therefore, the coverage of the theme would be 90%. To defeat this strategy which exploits non-informative bounds, a third criterion is employed: the median (weighted) one-sided multiplicative bounds width for 90% bounds. The one-sided width is the square root of the ratio of the estimated 0.95 bound to the estimated 0.05 bound. These widths are sorted in ascending order along with the weights corresponding to their associated records; the width at the first record having cumulated mass 0.5 or greater is reported as the median bounds width.

The reason for using a median instead of an average width is that the classical model has a few very wide widths, such as $3.2 \cdot 10^5$ for one of the physical variables associated with the volcanoes theme, which reflect great uncertainty in the true value of a variable. Including these widths in an average would swamp bounds for variables in other themes, for which there is much less uncertainty. Examples of the latter are those associated with Amsterdam stock market prices, for which the average bounds width is 1.04. It would be unfair to penalize a method such as the classical which might achieve a low ABSDist value, but which has a few very large widths compared to those of other methods, by applying an average rather than a median multiplicative bounds width. These widths will be reported in the same tables that report ABSDist for 90% bounds.

A fourth criterion is sensitivity to outliers. Do individual e' values that represent outliers within a record cause the aggregated estimate \hat{e} to diverge from

the true e farther for some methods than for others? Average MME by aggregation method will be reported for records with outliers, for both physical and probabilistic data.

The fifth and final criterion for ranking methods is ease or complexity involved in development and application. For example, the various types of mean (arithmetic, geometric, harmonic) and median are simple to compute; the Alpha-Stable technique has intermediate complexity, and the Bayesian is the most complex and computer resource intensive. A scorecard summarizing the rankings is provided at the end of the chapter.

4.2 Research Question 2 Significance

As an example of the advantage conferred by aggregation, contrast the chances of under- or overestimating the true e by a given factor, e.g., 2, 5, 10, 100, or 1000 for probabilistic data, and for physical data—before and after aggregating e' into \hat{e} , using the median of the e' .

Table 3: Chances of Under and Over-Estimation given a single prediction, e' before aggregation

Factor	Probabilistic Data Chance of Underestimation	Probabilistic Data Chance of Overestimation	Physical Data Chance of Underestimation	Physical Data Chance of Overestimation
2	22%	23%	15%	16%
5	13%	15%	7%	6%
10	8%	11%	4.5%	3.4%
100	3.5%	2.0%	1.3%	1.0%
1000	1.6%	0.6%	0.6%	0.2%

Table 4: Chances of Under and Over-Estimation after e' predictions have been aggregated into \hat{e}

Factor	Probabilistic Data Chance of Underestimation	Probabilistic Data Chance of Overestimation	Physical Data Chance of Underestimation	Physical Data Chance of Overestimation
2	11%	20%	11%	13%
5	7%	17%	5%	3%
10	3%	8%	3%	1.3%
100	0.93%	0.9%	0.34%	0.38%
1000	0.46%	0.00%	0.14%	0.003%

Table 3: Chances of Under and Over-Estimation given a single prediction, e' before aggregation and

Table 4: Chances of Under and Over-Estimation after e' predictions have been aggregated into \hat{e} show that while aggregation does not reduce under- or overestimation probability in every case, the *combined* chances at each factor—that is to say, regarding the factors as maximum multiplicative errors, MME, with associated probabilities of occurrence—decrease by factors of between 1.2 and 5.6, under aggregation. Average weighted MMEs will be compared for different aggregation methods, along with percentage of realized values contained within nominal 90% bounds, ease of computation, sensitivity to outliers, and so forth. The chapter will conclude with a scorecard comparing the different methods.

4.3 Research Question 2 Literature Review

Clemen and Winkler (1999) observed that “Combination, or aggregation procedures are often dichotomized into mathematical and behavioral approaches, although in practice aggregation might involve some aspects of each” (p. 188). The former approaches range from summary metrics such as medians of estimated probabilities to complex models, whereas the latter are largely focused

on generating agreement among experts. Although this research focuses on mathematical approaches, a discussion regarding the Delphi method, which is a behavioral approach is provided by way of contrast. The justification for this inclusion is historical; in Helmer and Rescher (1959) argued that in fields that do not yet have scientific laws, the testimony of experts is warranted. The issue is how to use this testimony and, specifically, how to combine the testimony of a number of experts into a single useful statement which is characterizes aggregation in general. The Delphi method was developed to support this argument.

The Delphi method was “developed at the RAND corporation in the early 1950s as a spin-off of an Air Force-sponsored research project” on Soviet strategic selection of U.S. industrial targets for atomic bombing, and “is undoubtedly the best-known method of eliciting and synthesizing expert opinion” (Cooke, 1991, p. 12). An account of a Delphi experiment conducted at that time, was released for open publication a decade later by Dalkey and Helmer (1963). The technique involved exposing the views of experts to each other anonymously (to avoid direct confrontation), and asking those whose predictions fall outside the lowest 25th percentile and highest 25th percentile of responses, to justify their responses. Another round of elicitation was then conducted, and the process was repeated, typically for several iterations. The median obtained on the final iteration was used as the synthesized opinion.

Brown and Helmer (1964) noted a reduction by a factor of two in the interquartile range of responses to a number of questions following several

iterations of the Delphi process. The authors (1964) also found that self-rated “elite” responders produced medians closer to the true value, compared with randomly-chosen responders. However, Cooke (1991) stated (p. 15) that Brockhoff (1975) found in “an extensive study …that self-ratings of participants did not coincide with “objective expertise” as measured by relative deviation” of the estimate from the true value. Additionally, the Delphi technique itself was criticized by Sackman (1975), who found that the iterations tended to force consensus by requiring respondents whose views fall outside the interquartile range to justify their responses. According to Cooke (1991), “group interaction … tended to make the participants more confident [without leading] … to an increased relative frequency of correctness.” (p. 17).

Winkler (1968) included Delphi and self-rating in a discussion of behavioral reassessment approaches. Winkler also discussed weighted-average methods; in one example, the decision-maker subjectively ranks N respondents from worst (rank 1) to best (rank N), then obtains a composite distribution by weighting each elicited distribution by a factor equal to its rank divided by the sum of the ranks. Self-rating (with or without reassessment) can also be used to weight distributions. Winkler (2015, p. 17) noted that “while still being open to more complex combining models/methods [he] had gravitated over the years to a feeling that there are considerable advantages to simple models when combining subjective probability forecasts”.

O’Hagan et al. (2006) characterized mathematical aggregation as Bayesian and opinion pooling, where the latter includes linear pooling (LinOp) techniques

and Logarithmic Opinion Pooling (LogOp) pooling techniques. The LinOp is a “weighted average of the individual densities, where the weights w_i sum to 1” (p. 181). The LogOp replaces the sum by a normalized product of densities each raised to the power w_i . In the case where distributions are replaced by discrete point estimates (also called forecasts), and equal weights w_i are applied, LogOp reduces to the geometric mean, and LinOp to the arithmetic mean.

Winkler (1968) discussed an alternative method of aggregating distributions, by applying weights in the Bayesian framework to the parameters of a natural conjugate prior pair distribution. Winkler (1968) used the example of the beta-Bernoulli conjugate pair. In this example, the i -th expert’s prior distribution for a quantity, p follows the beta distribution with parameters r_i and n_i . If the sample distribution is Bernoulli with r successes in n trials, then the posterior distribution obtained via Bayes’ Theorem can be shown to be beta distributed with parameters $r_i + r$, $n_i + n$. Winkler (1968) stated that if there are k experts, each with a beta prior with parameters r_i and n_i , $i=1,2,\dots,k$, and if weights w_i are applied to each and combined in the Bayesian framework, the resulting prior is beta distributed with parameters $\sum_{i=1}^k w_i r_i, w_i n_i$. (Note: the “ r ” and “ n ” obtained from the sample would be added to $\sum w_i r_i$ and $\sum w_i n_i$, respectively, to obtain the parameters of the posterior beta distribution.) The weights could range from $1/k$ (all experts have the same distribution) to 1 (completely independent beta distributions). Winkler (1968) noted that the beta Bernoulli conjugate pair is unimodal.

An axiomatic approach was taken by Morris (1977), who obtained an “interesting and surprisingly simple result”—the decision-maker’s posterior distribution equals the product (after normalization) of a calibrated expert’s prior and the decision-maker’s prior. However, these results were obtained by making “prohibitively strong” (Cooke, 1991, p. 180) assumptions. One assumption is scale invariance: “the variance of the expert’s prior *alone* provides no information to the decision maker about the uncertain quantity” (Morris, p. 681). Cooke (1991) stated that this assumption is “gratuitous” and that in assessing log failure probabilities “it is plausible that higher probabilities will be better known and hence have smaller variance” (p. 181).

Another assumption is shift invariance: “the decision maker’s assessment of how surprised the expert is likely to be when the true value of the uncertain variable is revealed is not conditional on the true value” (Morris, 1977, p. 681). Cooke (1991) stated that this is “a very strong assumption”, and gave an example of a company director who “thinks that [a competitor’s price next year] will be about \$20. The director “is confident in his advisor, and thinks there is a probability of $\frac{1}{2}$ that the true price will fall between the 25% and 75% quantiles of whatever distribution the advisor gives. At this moment an industrial spy delivers a memo recently purloined from the competitor in which the competitor’s price is revealed to be \$5. The competitor is making a surprise move and starting a price war.” Under shift invariance, the decision maker would have to still believe that \$5 falls within the advisor’s interquartile range. “However, if hearing the price \$5

leads him to think that the adviser's 25% quantile will probably be greater than \$5, then shift invariance is violated" (p. 182).

Several assumptions were made by Mosleh and Apostolakis (1986) in developing maximum likelihood estimates of the parameters of an aggregate fragility curve distribution in a Bayesian framework. A fragility curve specifies a conditional failure frequency of a component, given the level of a stress input. Each expert's likelihood function was assumed to be lognormal, and his elicited percentiles were also assumed to be independent of each other "an unrealistic assumption" (Mosleh & Apostolakis, 1986, p. 452). The model allowed for dependencies among experts as well as biases. An example of a bias in a lognormal framework is a multiplicative factor b , which is applied to the true but unknown value x , to produce the median of an expert's elicited distribution given x . The formulas used for the posterior distribution yielded an aggregate median equal to the geometric mean of the elicited medians after adjusting for each expert's bias, when other elicited percentiles were not used. The geometric mean result will be considered in the current research.

Incorporating multiple sets of interdisciplinary data into an aggregation scheme was explored by Forrester (2005) in an interdisciplinary meta-analysis of expert judgment studies using absolute percentage error (APE) as an error metric. Over 1500 studies were initially considered; summary results from 58 studies were used for this metric. APE was "either implicitly stated or extracted from data in each case study" (Forrester, 2005, p. 29). Forrester (2005) explored fitting APEs using exponential likelihood with a gamma conjugate prior, but concluded

that the fit was poor. A different likelihood and prior was also explored which produced a lognormal posterior distribution for APE. APE will not be used as a metric in this research, since it treats large multiplicative underestimates, e.g., by factors of 10^3 or 10^6 as essentially identical, whereas in reality, they could be associated with very different economic consequences. Forrester (2005) recommended “performing additional expert judgment case studies” (p. 118) to refine the empirical relationships between true values and estimates. This was necessitated by the limited range of error in the APE data used by Forrester (2005).

It is noted that Forrester (2005) also explored predicting accuracy using expert attributes; accuracy is calculated as $(TP + TN)/(P + N)$, where $P\equiv\#$ of positive instances (e.g., of a disease); $N\equiv\#$ of negative instances (e.g., disease not present); $TP\equiv\#$ of true positive predictions by an expert (e.g., states disease present when it is present); $TN\equiv\#$ of true negative predictions by an expert (e.g., states disease absent when it is absent). Eighty data points for accuracy values were obtained from as many studies. Nine categorical attribute variables X_i were defined; these included years of education, specialized training, membership in professional organizations in the area of expertise, and so on. The attributes took both binary values and ordinal values where an “intrinsic hierarchy” (Forrester, 2005, p. 53) applied. An example of the latter is years of practical experience in the area of expertise: if <5 years, $X_i=1$, if between 5 and 10 years, $X_i= 2$, else $X_i= 3$.

3. Correlations between the dependent variable, accuracy, and each independent variable, X_i were computed. This approach cannot be applied to TUD as it has

been anonymized. Similarly, FAA SME input in the Test Data (Seq IDs TDPHYS2; TDPHYS3; TDPHYS4) has been anonymized.

4.3.1 Choice of metric

Although “it is generally accepted that there is no single best accuracy measure, and selecting an assessment method is essentially a subjective decision” (Shirazi, 2009, p. 39), an error metric involving the prediction, \hat{e} , divided by the realized value, e is a simple measure which satisfies “the majority of established requirements, while being easy enough for numerical calculations: [It is] scale-independent, interpretable, minimally impacted by outlier observations or errors and can eliminate the bias introduced by possible trends, and seasonal components.” (Shirazi, 2009, p. 39). However, an argument can be constructed for considering the inverse of this metric: e/\hat{e} . Pursuant to the Nunn-McCurdy Act of 1983, the Department of Defense is required “to report to Congress whenever a major defense acquisition program experiences cost overruns that exceed certain thresholds.” (Library of Congress, 2010). The thresholds are defined in terms of percentage increases such as 15%, 25%, 30%, or 50%. Similarly, the FAA is required to report to Congress baseline cost breaches exceeding certain thresholds. For performance or benefits estimates, a “breach” is defined in terms of a threshold percentage below the estimate. As previously stated, a derived FOM which will be used is the maximum multiplicative error, $MME \equiv \max(\hat{e}/e, e/\hat{e})$, where e is the realized value of a given metadata variable. This reflects the assumption that over the large set of metadata variables, the consequence of an overestimate is equal to the consequence of an underestimate.

Shirazi (2009) examined several hundred studies to obtain over 1900 pairs of point estimates u and realizations u' , representing “over 60 different disciplines” (p. 21). Shirazi (2009) used the relative error $E \equiv u'/u$ as the FOM, and developed equations in the Bayesian framework for the posterior distribution $\pi(u)$ given a prior $\pi_0(u)$ and an expert’s estimate, u' , along with additional evidence in the form of observed relative errors of previous estimates, E_1, E_2, \dots, E_n . The previous estimates were against a single, true value u in the case of a homogeneous pool. The likelihood function for u' was assumed to depend on both u and a parameter set θ ; the latter is assumed to have a conditional distribution given the observed E_i . This too is computed in the Bayesian framework, following which $\pi(u)$ is computed. Formal equations were developed for the non-homogeneous case, where multiple realizations for u are possible, and for the hybrid case, a combination of the homogeneous and non-homogeneous cases.

Shirazi (2009) applied a generic lognormal likelihood function to a small subset of TUD data (260 points), and found that approximately half of the estimates were improved by using the generic function. However, no value of u'/u for this data was less than .001 or greater than 1000, while 90% of the ratios fell within the range (1/3,3). In contrast, the EJE data contains ratios ranging from $6.5 \cdot 10^{-6}$ to $8 \cdot 10^7$, with over twenty percent of the ratios falling outside the range (1/3,3). Finally, results of analysis of the EJE data indicate that whether taken as a whole, or broken into individual cases, the data is not distributed lognormally. The sensitivity of the error metric to the number of experts, n , was

suggested by Shirazi (2009), as a future research area, particularly for the case $n > 10$. This was acknowledged as the EJE data includes records with 45 experts.

4.3.2 Linear Pooling Techniques

One aggregation technique which does not make overly restrictive assumptions is averaging. Ashton and Ashton (1985), in a study of forecast sales of *Time* magazine advertising pages over a 14-year period, found that:

The simple approach of weighting all the individual forecasts equally appears to be a promising solution to the problem of choosing a forecast weighting method. This conclusion is supported by the finding that the incremental accuracy of differential over equal weighting was small (p.1506).

Differential weighting incorporated additional information about the accuracy of individual forecasters' predictions. Averaging is one of the techniques considered in the current research.

Another approach for combining forecasts via linear pooling applies weights based on calibration and information scores. This is the "classical" model of Cooke (1991), who used the term "classical" because the computation of the calibration score has the same form as the computation of a significance level in a classical statistical test involving the right tail mass of a χ^2 distribution. The model combines experts' elicited distributions in a linear pool, weighting them according to the product of experts' calibration and information scores computed against seed variables for which the actual values were known. The formulae used to compute these scores are presented in Section 4.4. It suffices here to note

that for a well-calibrated expert, the fraction of the time the elicited intervals (e.g., between the 5th and 50th percentiles) contain the actual seed variable values will approximately match the theoretical probabilities (e.g., 45%). Similarly, the information score for an expert reflects the relative narrowness of his or her particular distribution compared to that of other experts. The overall score for an expert is the product of the calibration and information scores. The distribution corresponding to the weighted average of individual experts' distributions is called a decision maker, or DM.

Cooke and Goossens (2008) reviewed the application of the classical model in 45 case studies of expert panels over a 17-year period. The model was applied using a software tool called "EXCALIBUR" developed at TUD. The case studies represent "over 67,000 experts' subjective probability distributions" (Cooke & Goossens, 2008, p.657) in a variety of fields. Fields include, but are not limited to nuclear applications, chemical, aerospace, industrial accidents, health, finance, and volcanoes.

For each of the cases, Cooke and Goossens (2008) compared three weighting schemes for obtaining a DM: 1) equal weighting (EQ) where each expert's distribution is weighted by 1/N, where N is the number of experts; 2) best expert (BE) where the DM is identical to the distribution of that expert having the highest product score of calibration and information; and 3) performance weighting (PW) where experts' distributions are assigned weights proportional to an adjusted product score. The adjustment zeroes the product score for each expert having a calibration score less than a preset cutoff value, α ; for all other

experts, the adjustment leaves the product score unchanged. The value of α is selected so as to maximize the DM's unadjusted product score. For each case and for each DM weighting scheme, Cooke and Goossens (2008) compared calibration score, information score, and combined score (product score). For 90% of the cases, they found that PW outperformed EQ (for one third of the cases, BE was identical to PW, in that a single expert was assigned weight one).

Aspinall (2008) stated that the classical model's weights "are constructed to be 'strictly proper scoring rules' in an appropriate asymptotic sense: experts receive their maximal expected long-run weight by, and only by, stating their true degrees of belief. With these weights, statistical accuracy strongly dominates informativeness – one cannot compensate poor statistical performance by very high information." (Aspinall, 2008, p. 2)

However, Clemen (2008), in "Comment on Cooke's Classical Method", pointed out that an expert could, in theory, game the system by using a certain strategy. Suppose that 10th, 50th and 90th percentiles are elicited from ten seed variables. To attain a very high calibration score, the expert needs to capture a) one realization in his lowest interval (i.e., the realization needs to fall below the 10th percentile); b) one realization in the highest interval (i.e., above the 90th percentile); and c) four realizations each above and below the median. To accomplish a), the expert reports a very large value for the 10th percentile of one of the seed variables. To accomplish b), the expert reports a very small value for the 90th percentile of a second seed variable. To accomplish c), the expert reports

very wide intervals for the 10th to 50th, or for the 50th to 90th percentiles, as appropriate, for four seed variables each.

Clemen (2008) stated this process comes at the expense of a reduced information score due to the use of wide intervals; however, an optimum tradeoff point could exist. Clemen (2008) added that, in his experience, “experts who become engaged in the assessment process want nothing more than to express their beliefs as clearly and accurately as they can” (pp. 761–2). Clemen (2008) also noted that the comparison of PW and EQ performance in Cooke and Goossens (2008) relied on within-sample data, when “weights are calculated on the basis of the available data, and then performance scores are calculated using the same data.” (Clemen, 2008, p. 762) The conclusion was that “it should come as no surprise that PW performs so well relative to the other two combination methods when evaluated within-sample”. (Clemen, 2008, p. 762).

Clemen (2008) recomputed scores for PW, EQ and BE for 14 of the 45 studies, using a leave-one-out out-of-sample procedure. The results showed there was no statistically significant difference in combination score between PW and EQ. BE failed “utterly on calibration and hence on the combination score” (Clemen, 2008, p. 762). However, the results also showed that PW had considerably more variability in the combination score than EQ (PW and EQ both ranged from approximately zero to 1.4 and 0.6, respectively). Clemen (2008) also compared PW and EQ using the absolute difference between the predicted median and the realized value for each seed in a given case. Because different scales were involved, the differences could not be compared directly. However, the p-

value from a non-parametric sign test could be applied to test the null hypothesis that PW was at least as accurate as EQ. Clemen (2008) found that for four out of the 14 cases, PW was more accurate than EQ. In the other ten cases, EQ was significantly more accurate than PW in four of them, in that the p-values were less than 0.1. Clemen (2008) stated that the “results suggest that, overall” (p. 764) PW is less accurate than EQ, but suggested that all 45 cases be studied.

Aspinall and Cooke (1998) described one application of the expert judgment model in which it was necessary to adjust the weighting scheme in real time. The adjustments took place during the eruption of the Soufrière Hills volcano during the 1990s on the Caribbean island of Montserrat. According to the authors, expert elicitation procedure “had to be adapted to the exigencies of real time crisis management.... it was felt better to focus the procedure on quantifying the informativeness/conservatism of each individual’s views, rather than rely too heavily on a hurried and questionable calibration score” (p. 3). The authors stated that calibrating a group of volcanologists was “non-trivial at the best of times, let alone with an eruption going on outside the window!” (p. 3).

The range of views was compared to thresholds in order to recommend adjustments to public alert level by civilian administrators. Aspinall and Cooke (1998) concluded that the “most important attribute of the structured expert judgement approach for the Montserrat emergency has been that it provided an appropriate means to accommodate in the decision-making process the participation of local technical people involved long-term in the volcano

monitoring: those most directly affected by and concerned about the eruption in their home land” (p. 6).

Lin and Cheng (2008) used Clemen’s (2008) suggestion to examine more of the TUD cases using the out-of-sample procedure. The leave-one-out cross validation procedure removes one seed variable from the pool, and then computes performance weights for experts using the remaining seeds. Weights were then used to compute quantiles for the excluded variable. The variable was then restored to the pool, and a different variable excluded. Once quantiles were computed for all seed variables, a score was calculated for the set of distributions using the classical model’s calibration and information scoring rule. Lin and Cheng (2008) compared PW, EQ and BE weightings under this scheme, finding that the combined score for PW dropped considerably when leave-one-out cross-validation was applied (from 0.871 to 0.350, averaged across all cases), and was not significantly different from EQ (0.349, averaged across all cases) when out-of-sample analysis was applied. The authors stated that “using seed questions to sift out better calibrated experts may still be a feasible approach.

However, whether the cost of extra efforts used in generating seed questions and evaluating experts based on their performance on these calibration questions is justifiable remains a question.” (Lin & Cheng, 2008, p. 160). They also found that the calibration score is not “robust enough to help identify a clear trend and threshold” in the number of seed variables “beyond which Cooke’s performance weighting scheme will be consistently better than the equal weight or the best expert approaches” (Lin & Cheng, 2008, p. 158).

Similarly, an extensive leave k-out (where k ranges from 1 to N-1, where N is the number of seeds in a given case) across all 45 cases by Eggstaff et al. (2014) found that, using a sign test p-value approach similar to Clemen (2008), PW was superior to EQ “in roughly two thirds of the datasets” (p. 81); however, it was not possible state a priori which method would be superior in any given situation. Eggstaff et al. (2014) stated that the authors reserved “judgment without clear parameters to specifically compare the performance scores across studies and can only conclude that the use of either method might hinge on the situation.” (p. 81)

In a leave-one-out study of the out-of-sample behavior of Cooke’s weighting scheme, Lin (2011) stated that the “threshold for minimum number of seed variables required for achieving stable aggregation results remains a question to be clearly answered.” (p. 473). As in other leave-one-out studies, Lin eliminated all judgments relating to a given seed variable, then computed calibration, information and performance weight scores for each expert based on the remaining variables. The experts’ performance weight scores sum to unity, and constituted a vector in R^N , where N is the number of experts; the i-th component of the vector corresponded to the i-th expert’s score. Given n seed variables, the jackknife process generated n such points in succession, as a different seed variable was eliminated each time. Lin (2011) computed the Euclidean distance between each point and the average of the points, and reported the mean distance. Lin (2011) found considerable variation in this metric, even for cases having the same number of seeds. However, the “mean distance ...

[dropped] considerably with an increase in the number of questions if there are only several seed questions; however, the addition of seed questions ... [introduced] only marginal changes in the mean distance when the number of seed questions is larger than 15" (p. 474).

Lin (2011) noted that fluctuations in the weight distribution did not necessarily imply large fluctuations in the combined distribution. Lin (2011) stated that the problem domain may influence the fluctuation in the distance metric, and that Bayesian and other aggregation methods may also perform well. Lin (2011) concluded that "more comparison studies are still needed to establish guidelines for combining probabilistic judgments" (p. 478). This research indirectly addressed the problem domain gap by exploring the impact on aggregation of probabilistic versus physical quantities, as well as level of probabilistic quantity estimated.

Bolger and Rowe (2015) asserted that the acid test is how forecasts perform against unknown real-world target variables, not seed variables. The authors suggested that certain problem domains were amenable to controlled experiments using combination of simulation and aggregation, and that this was likely to be more productive than "trawling through old data using increasingly arcane cross validation" (p. 25).

How best to combine expert's judgments remains an open research area, as discussed by Hammitt and Zhang (2013) in comparing several methods of combining expert's judgments, such as best expert, performance weighting, and equal weighting, but used synthetic data with only two experts who are assumed

to be perfectly calibrated. They acknowledged that in many cases experts are not well calibrated, e.g., they found that physicians estimating the probability that patients admitted to an intensive care unit are discharged alive, as an exception and that most field data involve more than two experts. Hammitt and Zhang (2013) stated that they are consistent with studies based on field data in finding that equal weighting yields lower combined scores (the score is the product of the calibration and information scores for the decision maker). Hammitt and Zhang (2013) suggest that “future studies should explore how relative performance of the combination rules depends on the number of experts”. Shirazi (2009) also recommended exploring changes in the performance of the aggregation with numbers of experts exceeding ten. Over 150 variables in the meta-database contain more than ten e' values. The sensitivity of performance of two aggregation methods with numbers of e' values, including numbers exceeding ten, is examined in Chapter 6.

4.3.4 Alpha-Stable distribution

Rimmer and Nolan (2005) observed that “stable distributions are a rich class of probability distributions that allow skewness and heavy tails and have many intriguing mathematical properties” (p. 766). Borak, Hardle, and Weron (2005) asserted that “Since stable distributions can accommodate the fat tails and asymmetry, they often give a very good fit to empirical data” (p. 22). Although “Student’s t, hyperbolic, normal inverse Gaussian, or truncated stable” (Borak et al., 2005, p. 21) are heavy-tailed alternatives to the Gaussian law, the underlying motivation for using Alpha-Stable distributions for modeling financial data is that

they are supported by the Generalized Central Limit Theorem. This theorem states “that stable laws are the only possible limit distributions for properly normalized and centered sums of independent, identically distributed random variables” (Borak et al., 2005, p. 22). These laws subsume infinite variance models.

Nolan (2009) also noted that “some observed quantities are the sum of many small terms - the price of a stock, the noise in a communication system, etc. and hence a stable model should be used to describe such systems” (p. 4). Further, “many large data sets exhibit heavy tails and skewness. The strong empirical evidence for these features combined with the Generalized Central Limit Theorem is used by many to justify the use of stable models.” (Nolan, 2009, p. 4). Although, in general, closed form PDF and CDF does not exist for the Alpha-Stable distribution, it can always be specified by its characteristic function. An examination of published literature designed to identify use of Alpha-Stable distributions highlighted applications such as finances, network traffic, signal processing, and copyright protection of digital images. Several examples of Alpha-Stable applications are provided in this section.

Fama (1965) discussed Mandelbrot’s assertion that academic research had not paid attention to the “implications of the leptokurtosis usually observed in empirical distributions of price changes” (p. 42). Further, Fama (1965) noted Mandlebrot’s claim that if outliers “are numerous” (p. 42), their exclusion detracts from the significance of tests performed on the remaining data. According to Fama (1965) Mandelbrot argued that empirical distributions of price changes

“conform better to stable Paretian distributions with characteristic exponents less than 2 than to the normal distribution (which is also stable Paretian but with characteristic exponent exactly equal to 2” (, p. 89).

Following a study of price changes in thirty stocks of the Dow-Jones Industrial Average (DJIA) from approximately the end of 1957 to September 1962, Fama (1965) concluded that Mandelbrot’s hypothesis was substantiated by the data. Fama (1965) argued that the price level of a given security in a Gaussian market will be fairly continuous, in a stable Paretian market with $\alpha < 2$ it will usually be discontinuous”, (p. 89). From a practical standpoint, Fama (1965) noted that this finding underlined the variability of an expected yield, i.e., the expected yield is higher and the probability of larger losses is greater in a stable Paretian market than a Gaussian market. Similarly, Borak et al. (2005) found that the Alpha-Stable distribution (where $\alpha=1.64$) offered a superior fit for changes to the DJIA between 1987 and 1994 than the Gaussian distribution.

Xiahou, Guangxi, and Yaoting (2004) used a data set containing one million data packets developed by the Bellcore Morristown Research and Engineering Facility to model the number of packet arrivals and number of byte arrivals in network traffic, with the objective of predicting congestion. Based on the assumption that network traffic is characterized by three main properties, i.e., “the burstiness in all time scales, the long-range dependence (LRD), and the heavy tailed distribution” (Xiahou et al., 2004, p. 447), the authors determined that an Alpha-Stable distribution should be used since “it can analyze performance in Gaussian or non-Gaussian case” (Xiahou et al., 2004, p.448). The

authors concluded that based on the comparisons of the distributions of actual network data with hypothetical distributions, “we think that distributions of packet arrival and the distribution of byte arrival in network traffic are Alpha-Stable distributions” (Xiahou et al., 2004, p. 456).

Additionally, Mahmood, Chitre, and Armand (2012) asserted that large groups of snapping shrimp produce a noise effect, which is “detrimental for sonar and under-water communication systems” (p. 1). This noise is impulsive, and the authors noted that the symmetric Alpha-Stable distribution provided good practical estimates of snapping shrimp noise. In this chapter, Alpha-Stable distributions are fit to the physical and probabilistic meta-data following transformations.

4.4 Research Question 2 Mathematical Formalism

As noted in Chapter 1, an objective of this research is to recommend a method for combining expert judgment data by comparing and contrasting Bayesian, linear pooled calibration-based weighting and other schemes for aggregating expert judgment data—in particular, for aggregating e' (median) values elicited from experts. The “classical model” (Cooke, 1991) is a calibration-based approach to aggregation, and is discussed first. Under this model, the predicted median based on aggregation is a weighted average of the elicited e' values. The raw weight given each expert is a product of a calibration score and an information score on “seed items” for which the true realizations are known; the raw weights are normalized by dividing by their sum so that the

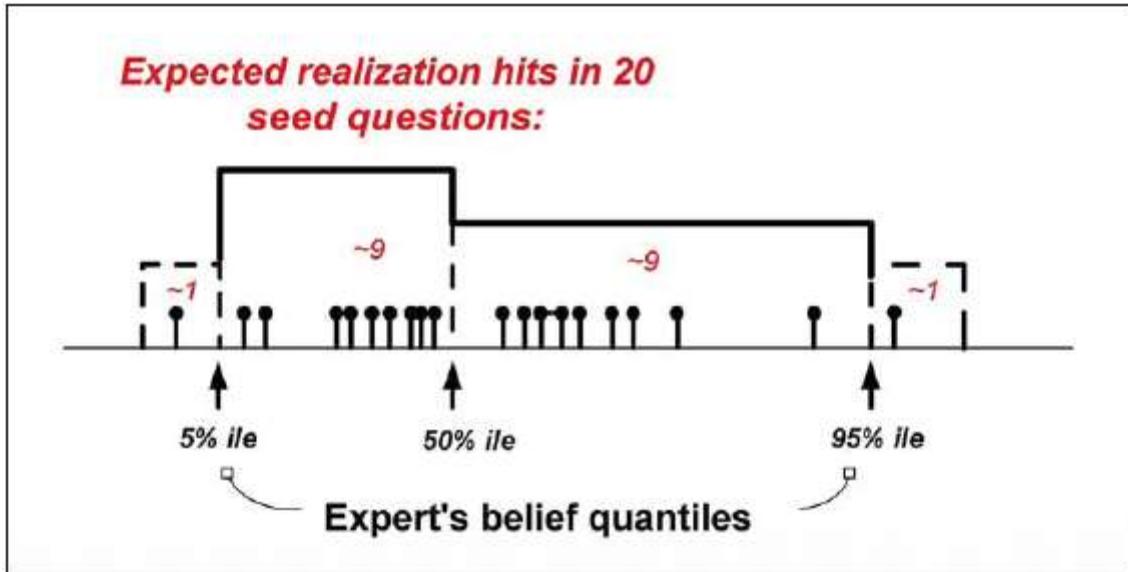
weights sum to unity. The classical model requires elicited values for percentiles such as 5th and 95th, in addition to the median, from each expert.

4.4.1 Calibration Score

A well-calibrated expert's responses to, for example, twenty seed questions, will be distributed such that roughly one (5%) of the realizations will fall below his 5th percentile; approximately nine (45%) of the realizations will fall between the expert's 5th percentile and his or median; approximately nine will fall between the expert's median and his 95th percentile; and approximately one will fall above the 95th percentile.

Figure 11: Schematic depiction of seed item realizations for a well-calibrated expert reproduces Figure 1 of Aspinall (2008). The caption in the latter reads “Schematic depiction of seed item realizations in relation to the inter-quantile ranges of a well-calibrated expert: the realization values should be distributed within the inter-quantile ranges in close agreement to the proportions $\{0.05, 0.45, 0.45, 0.05\}$ ”.

Figure 11: Schematic depiction of seed item realizations for a well-calibrated expert



Note: Adapted from Aspinall, (2008), Figure 1

For a given expert, the realization hits define a sample frequency distribution, s over the four bins corresponding to the expert's inter-quantile ranges. If there are N (in this example, $N=20$) seed questions with corresponding realization values. And, if the leftmost bin contains a single hit, then the corresponding sample frequency for the expert is $s_1 \equiv 1/N=0.05$. Similarly, if the second bin contains, for example, eight hits, then its sample frequency is $s_2 \equiv 8/N=0.40$. The four sample frequencies s_1, s_2, s_3 , and s_4 comprising distribution S , are compared to the “theoretical” frequencies comprising distribution P — p_1, p_2, p_3 , and $p_4 = (0.05, 0.45, 0.45, 0.05)$ respectively— using the Kullback-Leiber (K-L) divergence formula:

$$D \equiv \sum_{i=1}^4 s_i \cdot \ln(s_i/p_i) \quad (\text{Equation 4})$$

Since the information conveyed in an item or symbol is proportional to the log of its probability of occurrence, D can be interpreted as a measure of surprise: the expected difference in information conveyed when items which were thought to

have probabilities of occurrence specified by distribution P actually follow distribution S. Cooke (1991) stated that as the number of seed questions $N \rightarrow \infty$, the statistic $2N \cdot D$ converges to a chi-square distribution with three degrees of freedom (assuming three percentiles are elicited); eight to ten observations are considered sufficiently large that the convergence represents a good approximation. The classical model uses the right tail of the cumulative distribution evaluated at the point $2N \cdot D$ as the raw calibration score.

4.4.2 Information score

The information score for each expert is the average of his or her information scores with respect to each seed variable. Before the information score can be computed for a particular seed variable, a background information distribution must be specified for that variable. This background distribution is common to all experts, and is defined over the variable's "intrinsic range". To compute this range, first take the minimum, L, of the 5th percentile points for the variable elicited from all of the experts. Next, take the maximum, H, of the elicited 95th percentile points for the variable. Multiply the interval length H-L by an overshoot factor, k (typically, k=10%). The intrinsic range is [q_L,q_H] where q_L≡L-k(H-L), and q_H≡H+k(H-L). Cooke (1991) stated that, unlike the distribution used to compute the calibration score, the background information distribution is a "slow" function. Per Eggstaff et al. (2014), "The information score is considered a slow function in that even large modifications to an expert's assessment will only produce small changes in the score. In other words, each expert's weight is driven by the calibration score." (p. 74).

The EXCALIBUR application admits two scales for seed variables: uniform and lognormal. The background distribution is assumed to be uniform over $[q_L, q_H]$ if the seed variable is uniformly distributed. Otherwise, each elicited percentile point is replaced by its natural logarithm, after which the procedure for uniform variables is applied.

The remaining steps for calculating the information score will be illustrated via an example. The example assumes two experts, A and B, and two uniformly distributed seed variables, Var_1 and Var_2 . Let A's 5th, 50th and 95th elicited percentile values for Var_1 and Var_2 be represented by the triplets (40,60,200) and (50,70,100), respectively. Let the corresponding triplets for expert B be (30,60,150) and (40,90,300), respectively. By inspection, $L=30$ for Var_1 and 40 for Var_2 . Also by inspection, $H=200$ for Var_1 and 300 for Var_2 . $H-L=170$ and 260 for Var_1 and Var_2 , respectively. The overshoot lengths are 17 and 26, respectively, yielding $q_L \equiv L - k(H-L) = 30 - 17 = 13$, $q_H \equiv H + k(H-L) = 200 + 17 = 217$ for Var_1 . Similar computations yield $(q_L, q_H) = (40 - 26 = 14, 300 + 26 = 326)$ for Var_2 . The background densities for Var_1 and Var_2 are 204^{-1} over $[13, 217]$ and 312^{-1} over $[14, 326]$, respectively.

Expert A's information score with respect to Var_1 is now computed. By applying the Kullback-Leiber divergence formula, each "sample" probability s_i is the delta between elicited percentile cumulative probabilities, while each "theoretical" probability (p_i) is the integral of the uniform background density over the corresponding interval. Thus, $s_2 = 0.45$, the difference between the 5th percentile and the 50th percentile. The distance between the corresponding values

equals $60-40=20$, and p_2 =uniform density mass over this interval = $20 \cdot 204^{-1} \approx 0.10$. The difference between q_L and the 5th percentile is $s_1=0.05$; the corresponding interval length is $40-13=27$, and the uniform density has mass $p_1=27 \cdot 204^{-1} \approx 0.13$ over this interval. Similar computations yield $s_3=0.45$, $s_4=0.05$, $p_3 \approx 0.69$, $p_4 \approx 0.08$. As before, $D \equiv \sum_{i=1}^4 s_i \cdot \ln(s_i/p_i)$, approximately ≈ 0.42 , given the preceding values for s_i and p_i . This is the information score with respect to Var_1 : D is not multiplied by $2N$, nor is the right-tail mass of a chi-square distribution.

Next, Expert A's information score with respect to Var_2 is computed. As before, $s_1=0.05$. The corresponding interval length is $50-14=36$; the uniform density has mass $p_1=36 \cdot 312^{-1} \approx 0.115$ over this interval. After the remaining s_i and p_i have been computed by analogous computations, $D \equiv \sum_{i=1}^4 s_i \cdot \ln(s_i/p_i)$ is found ≈ 1.40 . The final information score for expert A is found by taking the average of D over the seed variables, approximately $(0.42+1.40)/2 \approx 0.91$.

If Expert B had tried to game the calibration computation by using a very wide interval between elicited percentiles for one of the seed variables, his p_i over that wide interval would have been greater than Expert A's for the same percentile difference s_i , causing the ratio (s_i/p_i) —and thus his information score with respect to this variable, D —to be smaller for B than for A, *ceteris paribus*. The raw score for expert A=0.054; the product of the final information score: 0.91; and, the calibration score: 0.059.

The information scores for Expert B with respect to the two variables are computed analogously, and yield a final information score for Expert B of 0.31.

This is multiplied by the calibration score for B, 0.03 to produce a raw score for expert B of 0.009. The raw scores for Experts A and B are normalized by dividing by their sum, 0.063 ($=0.054+0.009$) to produce final weights of 0.86 ($=0.054/0.063$) and 0.14 ($=0.009/0.063$) for A and B, respectively. The normalized weights are then applied to the elicited medians of the experts to yield predicted medians for each variable.

A refinement given in Cooke (1991) involves an artificial expert, the decision maker —DM—defined by applying normalized weights to the elicited 5th, 50th and 95th percentiles, where the raw scores are normalized over a subset of experts having calibration scores not less than α , for some value of α . Calibration and information scores are computed for the DM in the usual manner, and a raw score obtained. The refinement is to set α so as to maximize the raw score for the DM. The accuracy of the classical model is compared to that of several other aggregation methods in this Chapter.

4.4.3 Alpha-stable distribution

Another linear-pooling technique that does not rely on calibration—corresponding to the still often-observed situation in the field - where seed variables are not used, and only point estimates are elicited - is based on the properties of the Alpha-Stable distribution.

Section 4.3.4 observed that stable laws are the only possible limit distributions for properly normalized and centered sums of independent, identically distributed random variables (Borak et al., 2005). Although, in general, closed form PDF and CDF do not exist for the alpha-stable distribution, it

can always be specified by its characteristic function. Using Nolan's S0 parameterization (Nolan, 2009, p. 8, Equation 1.4) this is:

$$E\exp(iuX) =$$

$$\begin{cases} \exp(-\gamma^\alpha |u|^\alpha [1 + i\beta(\tan^{\pi\alpha}_2)(\text{sign } u)(|\gamma\mu|^{1-\alpha} - 1)] + i\delta u), & \alpha \neq 1 \\ \exp(-\gamma|u|[1 + i\beta_2^2(\text{sign } u)\log(\gamma|u|)] + i\delta u), & \alpha = 1. \end{cases}$$

(Equation 5)

In the characteristic function, α controls the rate at which the tail density decays.

For values of α between 1 and 2, as $x \rightarrow \infty$, the density $f(x)$ decays according to $x^{-(\alpha+1)}$ (Nolan, p. 14, Theorem 1.12). However, in the special case $\alpha=2$, where the stable density reduces to a Gaussian distribution, the tail decays according to $\exp(-x^2/2)$. β is a skew parameter in $[-1,1]$; if negative, the distribution is left-skewed; if positive, it is right skewed.

Nolan (2009) stated that under S0, S0 parameterization? “ γ and δ determine scale and location in the standard way” (p. 9) subtracting δ and dividing by γ yields a stable distribution with decay, skew, scale and location parameter values $(\alpha, \beta, 1, 0)$. Nolan (2009, p. 15, Proposition 1.13) stated that the mean, $\mu = \delta - \gamma \cdot \beta \cdot \tan(\pi\alpha/2)$. (Proposition 1.13, p. 15). In the special case where $\alpha=2$, the variance does exist; per Nolan (2009, p. 9), it is $\sigma^2 = 2\gamma^2$. Further, per Nolan (2009, p. 20, equation 1.9), the sum of n independent Alpha-Stable variates is itself Alpha-Stable, with decay, skew, scale, and location parameters $(\alpha, \beta, \gamma \cdot n^{1/\alpha}, d)$, where $d = n \cdot \delta + \gamma \cdot \beta \cdot (n^{1/\alpha} - n) \cdot \tan(\pi\alpha/2)$. All of the above assume $\alpha \neq 1$, which is the case for the EJE data.

4.4.4 Gaussian mixture

As stated in this chapter, simple combination methods such as arithmetic, harmonic, and geometric means were considered. A Gaussian mixture approach for obtaining bounds around the geometric mean was explored. Setting $FOM \equiv e/e'$ and $\ln(FOM) = \ln(e) - \ln(e')$ as before, $x \equiv \ln(FOM)$ was modeled as a Gaussian mixture, G , consisting of two normal distributions $N_1 \equiv N(\mu_1, \sigma_1)$ occurring with probability p and $N_2 \equiv N(\mu_2, \sigma_2)$ occurring with probability $q \equiv 1-p$. The following section describes how the bounds are developed, given mixture parameters.

Consider a given SeqID, (a record identifier as explained in Chapter 2) having n observations e'_1, e'_2, \dots, e'_n . In what follows, assume these observations have been \ln -transformed. Let $x_1 \equiv \ln(e) - e'_1, x_2 \equiv \ln(e) - e'_2, \dots, x_n \equiv \ln(e) - e'_n$ be independent draws from the Gaussian mixture. If the e'_i are aggregated into an estimate \hat{e} using the geometric mean, then $\ln(\hat{e}) = \sum e'_i/n = \ln(e) - S/n$, where $S \equiv \sum x_i$, implying

$$n \cdot \ln(e/\hat{e}) = S \quad (\text{Equation 6})$$

Given the mixture parameters for G , the distribution of the right hand side of Equation 6 is known:

$F_S(t) \equiv \Pr\{S \leq t\} = \sum_{k=0}^n p_k \Pr\{S \leq t | k \text{ drawn from } N_1, n-k \text{ from } N_2\} = \sum_{k=0}^n p_k \Phi((t - \mu_k)/\sigma_k)$, where
 $p_k \equiv \Pr\{k \text{ of the } x_i \text{ are drawn from } N_1; \text{ and } n-k \text{ are drawn from } N_2\}; p_k = \binom{n}{k} p^k q^{n-k};$
 $\mu_k = k\mu_1 + (n-k)\mu_2; \sigma_k = [\sum_{i=1}^k \sigma_1^2 + (n-k)\sigma_2^2]^{0.5}$; and Φ denotes the CDF of $N(0,1)$.

(Equation 7)

Note that the sum of $k x_i$ drawn from $N_1 \sim N(k\mu_1, \sqrt{k}\sigma_1)$; the sum of $(n-k) x_i$ drawn from N_2 is distributed analogously.

Since Φ is monotonic in t , so is $F_S(t)$; hence quantile points $q_{0.1} \equiv F_S^{-1}(0.1)$ and $q_{0.9} \equiv F_S^{-1}(0.9)$ can be found by binary search. Since $S^{1/n}$ is monotonic in S ,

Equation 6 implies 80 and 90 percent bounding intervals for e are, respectively:

$$[\hat{e} \cdot (\exp(q_{0.1}))^{1/n}, \hat{e} \cdot (\exp(q_{0.9}))^{1/n}] \text{ and } [\hat{e} \cdot (\exp(q_{0.05}))^{1/n}, \hat{e} \cdot (\exp(q_{0.95}))^{1/n}]$$

(Equation 8)

4.4.5 Median

The median, an additional simple aggregation method was also explored. Bounds were obtained using quantile regression applied to the median. Quantile regression will be discussed in the context of the likelihood function for a Bayesian aggregation technique discussed in Section 4.5.1.

4.4.6 Bayesian aggregation

The equations provided in this section are derived from Mosleh and Apostolakis (1986) and Shirazi (2009). According to Bayes' Theorem, given an initial prior distribution for an unknown quantity u $\pi_0(u)$; evidence u' ; and a likelihood function $L(u'|u)$; the posterior distribution for the quantity $\pi(u|u')$ can be obtained via

$$\pi(u|u') = L(u'|u)\pi_0(u)/\int L(u'|u)\pi_0(u)du \quad (\text{Equation 9})$$

This is consistent with the formulation in Shirazi (2009, Eq. 2, p. 42). The constant k can be regarded as a factor required to normalize the posterior density so that it integrates to one.

4.4.6.1 Multiplicative Lognormal Error Model

A Bayesian aggregation scheme presented in Eq. 11 of Mosleh and Apostolakis (1986) assumes a multiplicative error model, in which an expert's prediction, u' is assumed to equal the product of the true value, u and an error or bias factor, b independent of u :

$$u' = ub \quad (\text{Equation 10})$$

Assuming b follows a lognormal distribution, $\ln(b) \sim N(\mu_1, \sigma_1^2)$ for the first expert; the density function is

$$f_b(x) = \frac{1}{\sigma_1 \sqrt{2\pi} x} e^{-[\ln(x) - \mu_1]^2 / [2\sigma_1^2]} \quad (\text{Equation 11})$$

Letting F denote a cumulative distribution function,

$$F_{u'|u}(x) \equiv \Pr\{u' < x | u\} = \Pr\{ub < x | u\} = \Pr\{b < x/u | u\} \equiv F_{b|u}(x/u) \quad (\text{Equation 12})$$

Taking $\frac{\partial}{\partial x}$ of both sides,

$$f_{u'|u}(x) = f_{b|u}(x/u) \cdot (1/u) = f_b(x/u) \cdot (1/u) \quad (\text{Equation 13})$$

The left hand side of this equation is the likelihood function $L(u'|u)$, evaluated at $u' = x$. Substitute (x/u) for x in Equation 11, and multiply by $(1/u)$:

$$L(x|u) = \frac{1}{x\sigma_1 \sqrt{2\pi}} e^{-[\ln(x/u) - \mu_1]^2 / [2\sigma_1^2]} = \frac{1}{x\sigma_1 \sqrt{2\pi}} e^{-[\ln(u) - M_1]^2 / [2\sigma_1^2]}, \quad \text{where } M_1 \equiv \ln(u') - \mu_1. \quad (\text{Equation 14})$$

Assuming the prior is also lognormal, $\ln(u) \sim N(\mu_0, \sigma_0^2)$ and the prior density

$$\pi_0(u) = \frac{1}{\sigma_0 \sqrt{2\pi u}} e^{-[\ln(u) - \mu_0]^2/[2\sigma_0^2]} \quad (\text{Equation 15}).$$

Then, ignoring constants, Eq. 9 implies that

$$\pi(u|u') \propto e^{-[\ln(u) - M_1]^2/[2\sigma_1^2]} \cdot \pi_0(u) = \left(\frac{1}{u}\right) e^{-\{[\ln(u) - M_1]^2/[\sigma_1^2] + [\ln(u) - \mu_0]^2/[\sigma_0^2]\}/2} \quad (\text{Equation 16})$$

Ignoring constants,

$$\{ \} = [\ln(u)]^2 \cdot [1/\sigma_1^2 + 1/\sigma_0^2] - 2[\ln(u)] \cdot [M_1/\sigma_1^2 + \mu_0/\sigma_0^2] = [\ln(u) - \mu_p]^2/\sigma_p^2 + c, \text{ where}$$

$$1/\sigma_p^2 = 1/\sigma_0^2 + 1/\sigma_1^2 \text{ and } \mu_p/\sigma_p^2 = M_1/\sigma_1^2 + \mu_0/\sigma_0^2 \quad (\text{Equation 17})$$

The constant c used to complete the square equals $-\mu_p^2/\sigma_p^2$ and can be ignored, as it is incorporated into the constant k^{-1} required to make the integral of the posterior density equal to one.

Accordingly, the posterior density $X_p \equiv \pi(u|u')$ is recognized as lognormal, with

$$\ln(X_p) \sim N(\mu_p, \sigma_p^2). \text{ Letting } \tau \text{ denote precision, } \tau_0 \equiv 1/\sigma_0^2, \tau_1 \equiv 1/\sigma_1^2, \tau_p \equiv 1/\sigma_p^2; \mu_p \tau_p = M_1 \tau_1 + \mu_0 \tau_0.$$

Let weights $\omega_1 \equiv \tau_1/\tau_p, \omega_0 \equiv \tau_0/\tau_p$; then $\mu_p = M_1 \omega_1 + \mu_0 \omega_0$, where the weights sum to one as $\tau_p = \tau_0 + \tau_1$.

Let X_{50} and μ_{50} denote the medians of the posterior and prior distributions, respectively. Let b_{50} denote the median of the expert's bias, b . Then $b_{50} = e^{\mu_1}, \mu_{50} = e^{\mu_0}$, and

$$X_{50} = e^{\mu_p} = (e^{M_1})^{\omega_1} \cdot (e^{\mu_0})^{\omega_0} = (u'/b_{50})^{\omega_1} \cdot (\mu_{50})^{\omega_0} \quad (\text{Equation 18})$$

The posterior median equals the geometric mean of the first expert's prediction adjusted for bias, and the prior, with the weights proportional to the precision. It follows by induction that for an arbitrary number of independent experts having lognormal error factors, the posterior median equals the geometric mean of the adjusted experts' predictions and the prior, with weights as before

proportional to each expert's share of the cumulative? precision. As previously stated, this result is consistent with the formulae in Mosleh and Apostolakis (1986), provided zero precision is assumed for percentiles other than the median. The probability associated with a credible interval can be constructed by integrating the posterior distribution over the corresponding interval. Assuming zero precision for the prior, equal weights w_i , and no bias, X_{50} reduces to the geometric mean. The geometric mean is one of the aggregation methods used in this research.

As an exercise, the effect on the geometric mean of naively estimating bias using a “leave-one-out” procedure, was explored. As an example of the method, consider a theme such as ACNEXPTS, with ten variables and seven experts. The bias of the i -th expert, for purposes of adjusting the prediction against the first variable, is estimated by taking the geometric mean of the ratios of realized value to prediction of the i -th expert, e'_i , for each of the remaining nine variables. This factor is then applied to the prediction against the first variable. The procedure is repeated for each of the other variables and experts, in turn. This bias adjustment was not performed for themes having mixed numbers of variables.

4.4.6.2 Lognormal Absolute Percentage Error, APE

Forrester (2005) also employed a Bayesian framework, using absolute percentage error, APE as the metric of expertise. $APE \equiv |u' - u|/u \cdot 100$, where u' and u are an expert's estimated value, and the actual value, respectively. Forrester (2005) used meta-data to fit an exponential distribution to 58 APE values, E_1, E_2, \dots, E_{58} representing as many studies. A result of 11.059 was obtained for the mean parameter, β , via the maximum likelihood estimate (MLE) $\hat{\beta} = \sum E_i/N$, where $N=58$. Forrester noted that a rate parameter, $\lambda \equiv 1/\beta$ could be estimated in the Bayesian framework. A gamma prior with parameters a and b was assumed: $\pi_0(\lambda) = b^a (\lambda)^{a-1} e^{-b\lambda} / \Gamma(a)$. The

likelihood function is exponential in the rate parameter: $L(E)=\lambda e^{-\lambda E}$. Since the gamma is the conjugate prior for the exponential, the posterior distribution was also gamma, with parameters $a + N$ and $b + \sum E_i$ and mean $(a + N)/(b + \sum E_i)$.

However, Forrester (2005) found in “initial validation exercises” that the exponential distribution yielded “very poor fits for the data sets.” (pp. 101–102). Forrester (2005) also noted that the exponential “discards the lack of error values at and near zero”, assuming “experts are most likely to achieve zero error, and larger errors with decreasing probability” (p. 34). Accordingly, a lognormal distribution $\frac{1}{E\sigma_E\sqrt{2\pi}}e^{-[\ln(E)-\ln(\mu_E)]^2/[2\sigma_E^2]}$ having “a low probability of error at and near zero” (p. 34) was also fitted to the data.

Since the \ln of each E_i is normally distributed with mean and variance parameters $\ln(\mu_E)$ and σ_E^2 , respectively, the MLE for the former is given by the sample mean: $\ln \hat{\mu}_E = \sum_{i=1}^{58} \ln(E_i)/58$; the MLE for the latter is given by the sample variance: $\sigma_E^2 = [\sum_{i=1}^{58} (\ln(E_i) - \ln \hat{\mu}_E)^2]/58$. This yielded $\mu_E = \exp(\ln \mu_E) = 5.104$ and $\sigma_E = 1.64$. These empirical values were subsequently applied by Forrester (2005) in the Bayesian framework. Forrester (2005) noted that the formula for APE implies $f(u') = (100/u)f(E)$, for a given value of u . Accordingly, the lognormal likelihood function in $E, f(E)$ can be expressed as a likelihood function in u' , $f(u')$. The change of scale factor, $100/u$ can be absorbed into the prior $\pi_o(u)$ in the equation for the posterior density

$$\pi(u|u') = k^{-1} L(u'|u) \pi_0(u), \text{ where } k = \int L(u'|u) \pi_0(u) du. \quad (\text{Equation 19})$$

In this case, letting Y denote $\ln[|u' - u|/u \cdot 100]$ and ignoring multiplicative constants, the posterior distribution is equivalent to $e^Y e^{-[Y - \ln \mu_E]^2/[2\sigma_E^2]}$. This can be recognized as a normal distribution

centered about $Y = \sigma_E^2 + \ln(\mu_E)$. As previously stated, APE will not be used as a metric in this research, since it treats large multiplicative underestimates, e.g., by factors of 10^3 or 10^6 as essentially identical, whereas in reality, they could be associated with very different economic consequences.

4.4.6.3 Generic and Case-specific Distributions for Homogeneous and Nonhomogenous Data

Shirazi (2009) used a larger set of meta-data in a Bayesian framework, to update expert judgment. The principal metric used was relative error, u'/u , where u' and u are the expert's estimate and the realized value of the quantity estimated, respectively. The basic Bayes' equation for the posterior distribution $\pi(u|u')$ was given:

$$\pi(u|u') = L(u'|u)\pi_0(u)/ \int L(u'|u)\pi_0(u)du. \quad (\text{Equation 20})$$

Further, Shirazi (2009) used the assumption discussed previously, of a multiplicative error model, under which the estimate u' equals the product of the realized value, u and an error factor E . As before, the density of E , if lognormal, is given by

$$f(E) = \frac{1}{E\sigma_E\sqrt{2\pi}} e^{-[\ln(E) - \ln(E_{50})]^2/[2\sigma_E^2]} \quad (\text{Equation 21})$$

Shirazi (2009) noted that given u , $f(u')du' = f(E)dE$, hence $f(u') = f(E)dE/du' = f(E)/u$, yielding—after substituting $\ln(u') - \ln(u)$ for $\ln(E)$, u'/u for E —likelihood function

$$L(u'|u) = \frac{1}{u'\sigma_E\sqrt{2\pi}} e^{-[\ln(u') - \ln(u) - \ln(E_{50})]^2/[2\sigma_E^2]} \quad (\text{Equation 22})$$

Shirazi (2009) noted that the parameters E_{50} and σ_E comprised a parameter set $\underline{\theta}$, and that if the latter had a random distribution $g(\underline{\theta})$ due to epistemic uncertainty, it could be estimated given a

prior for $\underline{\theta}$, $\pi_0(\underline{\theta})$, and observed relative error evidence, $\underline{E} = E_1, E_2, \dots, E_n$: $g(\underline{\theta}|\underline{E}) = k^{-1} L(\underline{E}|\underline{\theta})\pi_0(\underline{\theta})$, where $k = \int_{\underline{\theta}} L(\underline{E}|\underline{\theta})\pi_0(\underline{\theta})d\underline{\theta}$. In this expression, $L(E|\theta)$ is the product of the likelihoods, with individual L_i given by L in Eq. 12 above, where $L(u'|u)$ is replaced by $L(u'|u, \theta)$. The posterior density for u given the evidence and the most recent estimate, u' is:

$$\pi(u|u', \underline{E}) = k^{-1} a \int_{\underline{\theta}} L(u'|u, \underline{\theta}) g(\underline{\theta}|\underline{E}) d\underline{\theta} \pi_0(u), \quad (\text{Equation 23})$$

$$\text{where } k = \int_u \int_{\underline{\theta}} L(u'|u, \underline{\theta}) g(\underline{\theta}|\underline{E}) d\underline{\theta} \pi_0(u) du.$$

Shirazi (2009) compared the median and mean of the posterior with the true value u , to determine the extent of reduction in relative error. This formula was applied to a “homogeneous case”, having a single true value, u . Shirazi (2009) also developed formal expressions for the posterior density in the case of multiple possible true values u_1, u_2, \dots, u_n (the “nonhomogeneous case”). In this situation, the distribution of θ , $g(\theta)$ is conditional on “hyperparameters”, $\underline{\omega}$: $\underline{\omega} = (\omega_1, \omega_2, \dots, \omega_n)$; $g(\theta) = g(\theta|\underline{\omega})$. The likelihood function for each estimate u'_i , observed true value u_i and consequent relative error, E_i is expressed in terms of the $\underline{\omega}$:

$$L(E_i|\underline{\omega}) = \int_{\theta} L(E_i|\theta) g(\theta|\underline{\omega}) d\theta \quad (\text{Equation 24})$$

As before, combined likelihood is the product of the L_i :

$$L(\underline{E}|\underline{\omega}) = \prod_{i=1}^n [\int_{\theta} L(E_i|\theta) g(\theta|\underline{\omega}) d\theta] \quad (\text{Equation 25})$$

The posterior density of the $\underline{\omega}$ can be estimated given a prior and evidence: $\pi(\underline{\omega}|\underline{E}) = k^{-1} L(\underline{E}|\underline{\omega})\pi_0(\underline{\omega})$

$$(\text{Equation 26})$$

$$\text{where } k = \int_{\underline{\omega}} L(\underline{E}|\underline{\omega}) \pi_0(\underline{\omega}) d\underline{\omega}.$$

At this point, the expected distribution of g can be computed:

$$\bar{g}(\theta|E) = \int_{\omega} \omega g(\theta|\omega) \pi(\omega|E) d\omega \quad (\text{Equation 27})$$

This eliminates the dependence on the $\underline{\omega}$. The posterior distribution for u given u' and the evidence E can now be obtained as before:

$$\pi(u|u', E) = k^{-1} \int_{\theta} \bar{g}(\theta|E) \cdot L(u'|u, \theta) d\theta \pi_0(u) \quad (\text{Equation 28})$$

where $k = \int_u \int_{\theta} \bar{g}(\theta|E) \cdot L(u'|u, \theta) d\theta \pi_0(u) du$.

Finally, Shirazi developed formal expressions to cover the case of a hybrid pool, where M_k repeated estimates of u_k are given. Analogous reasoning yields a density for the hyperparameters:

$$\pi(\omega|E) = k^{-1} \pi_0(\omega) \cdot \prod_{k=1}^N \left(\prod_{i=1}^{M_k} \int_{\theta} L(E_{ik}|\theta) g(\theta|\omega) d\theta \right) \quad (\text{Equation 29})$$

where $k = \int_{\omega} \pi_0(\omega) \cdot \prod_{k=1}^N \left(\prod_{i=1}^{M_k} \int_{\theta} L(E_{ik}|\theta) g(\theta|\omega) d\theta \right) d\omega$.

Average density $\bar{g}(\theta|E) = \int_{\omega} \omega g(\theta|\omega) \pi(\omega|E) d\omega$ can now be obtained, following which the posterior distribution is calculated as before:

$$\pi(u|u', E) = k^{-1} \int_{\theta} \bar{g}(\theta|E) \cdot L(u'|u, \theta) d\theta \pi_0(u) \quad (\text{Equation 30})$$

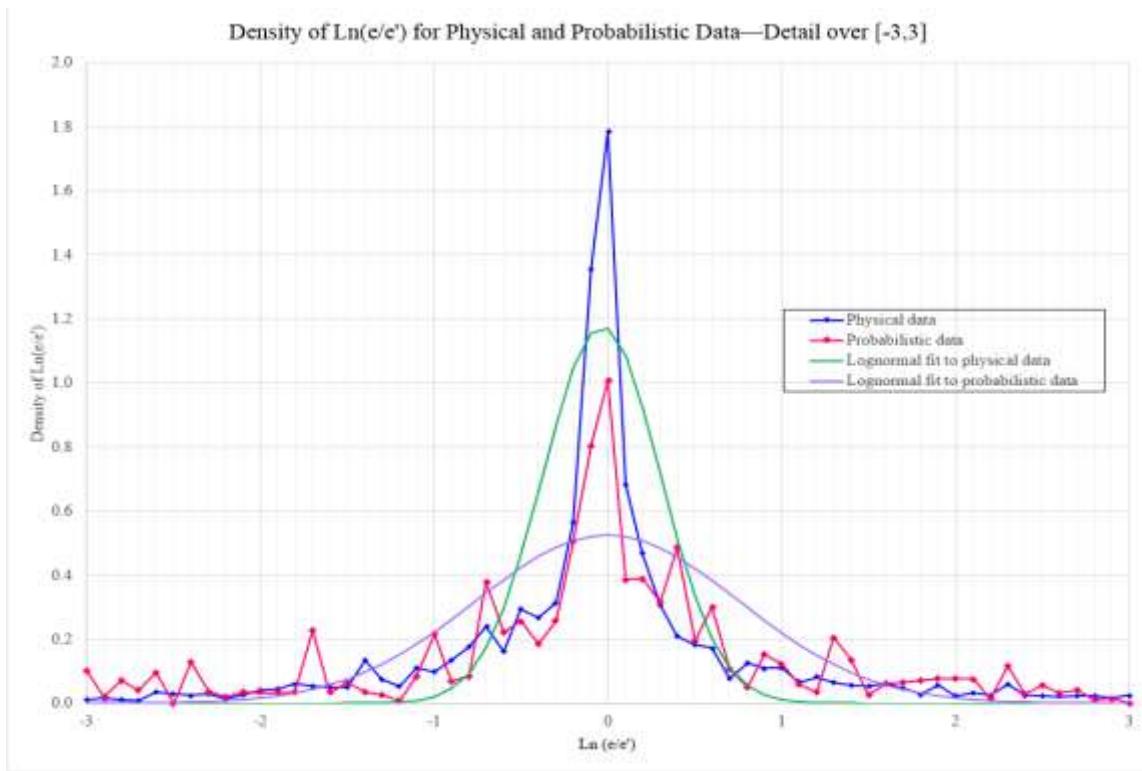
where $k = \int_u \int_{\theta} \bar{g}(\theta|E) \cdot L(u'|u, \theta) d\theta \pi_0(u) du$.

4.5 Research Question 2 Analysis and Results

The EJE dataset includes predictions ranging from 10^{-9} to 10^{10} , with multiplicative excursions exceeding 10^6 . As can be seen by inspection of

Figure 12: Density of $\ln(e/e')$ for Physical and Probabilistic Data—Detail over [-3,3] , the empirical distribution of e/e' is not lognormal for either data type.

Figure 12: Density of $\ln(e/e')$ for Physical and Probabilistic Data—Detail over [-3,3]



Additionally, sets of e' values for individual records can be bimodal -these records were rejected by the Shapiro-Wilk test for normality, whether log-transformed or not; reference Appendix B: Shapiro-Wilk Analysis.

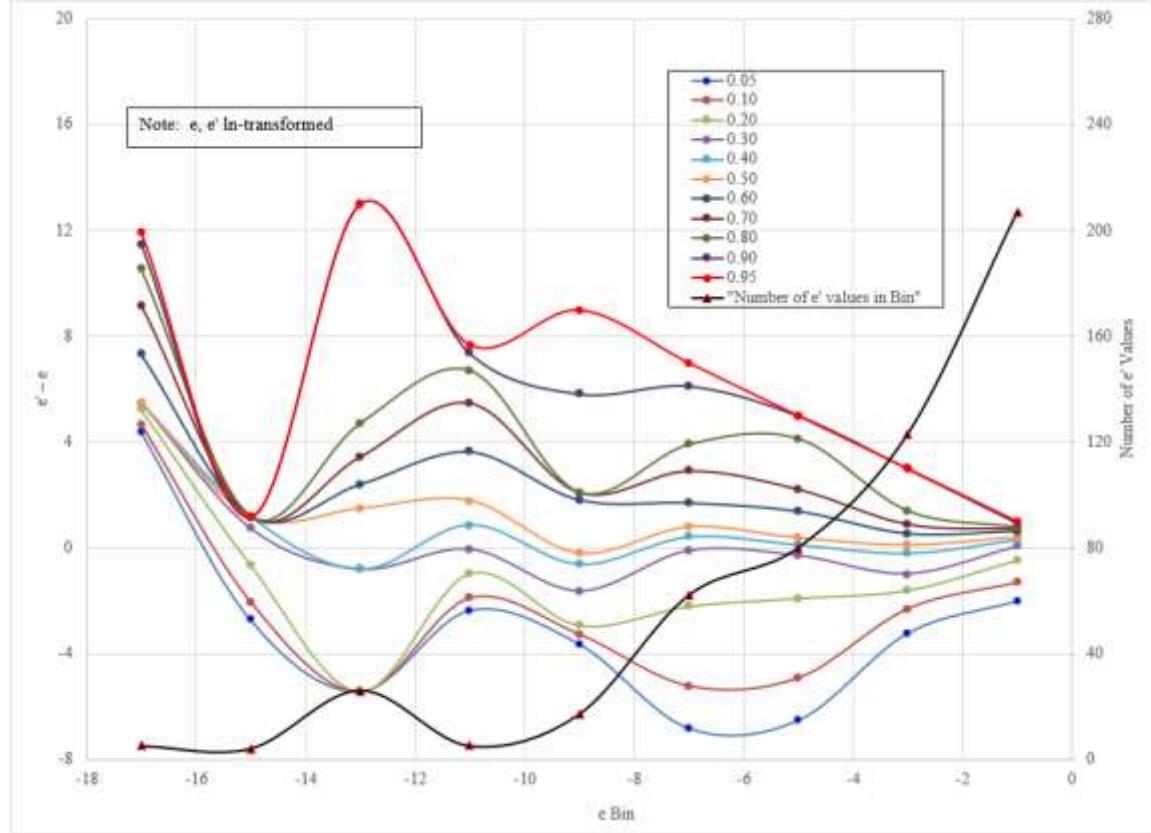
Except for the simple case of aggregating by geometric mean, a lognormal likelihood Bayesian scheme was not further considered. Other simple combination schemes for point estimates of \hat{e} given e' considered included arithmetic and harmonic means, along with the median of the e' predictions.

4.5.1 Bayesian Likelihood Function

To develop a sense of the form to be used for the Bayesian likelihood function, the empirical conditional distribution of $e'|e$ was considered. Percentiles of $\ln(e/e')$ versus binned $\ln(e)$ were plotted for probabilistic data, using bins of width two, ranging from -18 to zero. Log-

transforming the data accommodated the wide range of both e and e' ; see Figure 13: Percentiles of $e' - e \mid e$ for Probabilistic Data. The figure also shows the number of points in each bin; note that the two leftmost bins represent less than two percent of the data points.

Figure 13: Percentiles of $e' - e \mid e$ for Probabilistic Data



In general, this figure illustrates that the quantiles of e' broadly diverge from e as the latter decreases (again, the behavior of quantiles for bins centered on -17 and -15 can be ignored, as they reflect four and five e' values, respectively, out of more than 500 e' values). The e' tend to overestimate e at small values of e , consistent with results reported in the literature. As well, they tend to underestimate e as e approaches one. To allow for curvature of the quantiles while avoiding over-parameterizing the problem, a quadratic model was considered.

Quantile regression was applied to fit a second-order polynomial predicting e' given each e , to the meta-data. Quantile regression is based on minimizing the tilted absolute value function

$Tilted_Abs(\rho, x) = x \cdot (\rho - 1_x)$, where $1_x = 1$ if $x < 0$, 0 otherwise; and ρ is the quantile, e.g., 0.9.

(Equation 31)

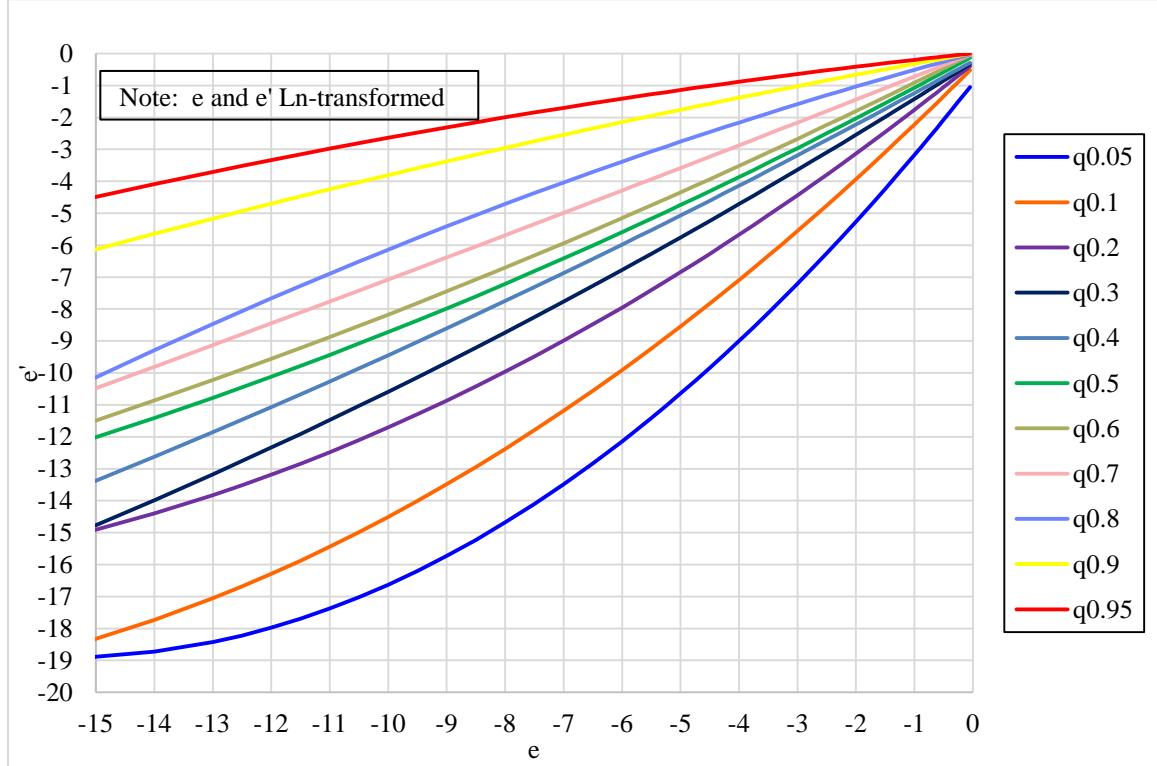
According to Koenker and Hallock (2001), the tilted absolute value function “asymmetrically [weights] absolute residuals—simply giving differing weights to positive and negative residuals” between observed and predicted. In order to incorporate record weights into the estimated quantiles, the tilted absolute value function was multiplied by the weight, w_e associated with the (e, e') pair. This weight is equal to the weight associated with the corresponding record containing realized value e and observation e' , divided by n , the number of observations in the record. The Python function `fmin` was used to solve for the second-order polynomial coefficients minimizing

$\sum Tilted_Abs(\rho, e' - polyval(coef, e)) \cdot w_e$ (Equation 32)

over all pairs of log-transformed (e, e') , where $polyval(coef, x) = a_0 + a_1 x + a_2 x^2$. The resulting quantile curves (5^{th} , 10^{th} , 20^{th} , 30^{th} , ..., 90^{th} , and 95^{th}) are plotted for probabilistic data in

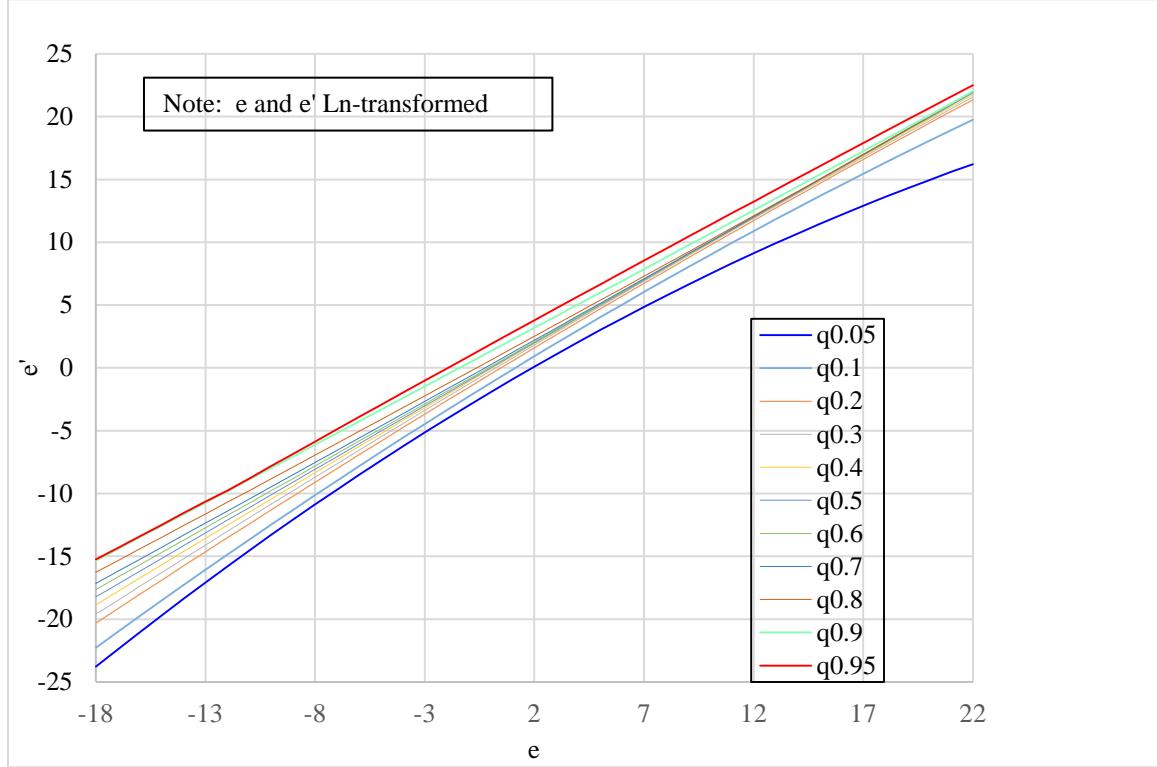
Figure 14: Quantiles of e' versus e for **Probabilistic Data**.

Figure 14: Quantiles of e' versus e for Probabilistic Data



A similar quadratic fit was applied to physical data, as well. (Straight line approximations were used for the 70th, 80th, and 90th, and 95th quantiles in each one percent tail of the distribution of e , in order to avert crossings.) The resulting quantile curves are shown in Figure 15: Quantiles of e' versus e for Physical Data. It will be observed that there is less curvature and smaller variation in the distance between 5th and 95th quantiles, versus e , for physical data than for probabilistic data.

Figure 15: Quantiles of e' versus e for Physical Data



Given the quantiles, the conditional CDF of $e'|e$ was obtained by linearly interpolating between them, which also produced the conditional density. Beyond the 95th - and below the 5th- percentiles, data was considered too sparse to apply quantile regression. Accordingly, a variant of a secondary power law approximation to the empirical distribution of $x=e/e'$, discussed subsequently in connection with fitting an Alpha-Stable distribution to the empirical distribution of x , was employed. The definition of x uses e and e' without ln-transform. It will be seen that an Alpha-Stable distribution is a reasonable fit to x after it has been transformed first by taking $\ln(x)$ and then by defining a new variable z as follows:

$$z = -[|\ln(x)|^{p_{lt}}] \text{ if } \ln(x) < 0; \text{ else } z = \ln(x)^{p_{rt}} \quad (\text{Equation 33})$$

where p_{lt} and p_{rt} are constants solved for by optimization.

The decay parameter α of the fitted Alpha-Stable was close to two (1.92 for probabilistic, 2.0 for physical), at which point the Alpha-Stable reduces to the normal distribution. The variance of the corresponding normal is twice the parameter γ of the Alpha-Stable; this produced σ values of 0.83 and 1.15 for the normal distributions associated with physical and probabilistic data, respectively. The approximation substituted a value of one for σ , and 0.5 for p_{lt} and p_{rt} , for both distributions. (The fit had produced values of 0.5 for each, for physical data, and 0.51 and 0.49 for p_{lt} and p_{rt} , respectively, for probabilistic data.). This resulted in $z = -[|\ln(x)|^{0.5}]$ if $\ln(x) < 0$; else $z = [\ln(x)]^{0.5}$ being treated as a standard normal distribution. With this approximation, the conditional density $L(e'|e)$ for values of e' above the 95th percentile point $y_{.95}$ at e (all quantities already ln-transformed) was calculated as follows.

The conditioning on e will be suppressed for notational convenience. To evaluate $L(e')$ for values of $e' > y_{.95} \equiv$ the 95th quantile, the following approach was taken:

$$s|e - e'|^{0.5} \sim N(0,1), \text{ where } s \equiv \text{sgn}(e - e') \quad (\text{Equation 34})$$

$$\Pr\{e' > y > y_{.95}\} = \Pr\{e' - e > y - e\} = \Pr\{e - e' \leq e - y\} \quad (\text{Equation 35})$$

$$= \Pr\{-|e - e'|^{0.5} \leq -|e - y|^{0.5}\} = \Pr\{N(0,1) \leq -|e - y|^{0.5}\} = \phi(-|e - y|^{0.5})$$

$$(\text{Equation 36})$$

Since the 0.95 quantile curve exceeds e , and $y > y_{.95}$, $|e - y| = y - e$.

$$\text{Let } F_e(y) \equiv \Pr\{e' \leq y\}; \text{ we have } 1 - F_e(y) = \phi(-(y - e)^{0.5}) \quad (\text{Equation 37})$$

Taking $\frac{d}{dy}$ of Eq. 37 yields: $-f_e(y) = -0.5\varphi(-(y-e)^{0.5}) (y-e)^{-0.5}$, hence

$$f_e(y) = \frac{1}{2\sqrt{2\pi}} \exp(-(y-e)/2) \frac{1}{\sqrt{y-e}}$$

Thus, $L(y) \sim \chi^2_{(1)} (y-e)/2$, $y > y_{0.95}$, where $\chi^2_{(1)}$ denotes the chi-square distribution with one degree of freedom. A final approximation was applied so that the conditional density beyond $y_{0.95}$ would integrate to 0.05: $L(y)$ was set equal to $\chi^2_{(1)}((y-e)/s)/[2s]$, where $s=(y_{0.95}-e)/[\Phi^{-1}(0.95)]^2$, $y > y_{0.95}$. Analogous computations yield $L(y) = \chi^2_{(1)}((e-y)/s)/[2s]$, where $s=(e-y_{0.05})/[\Phi^{-1}(0.95)]^2$, $y < y_{0.05}$. The conditional density discussed previously for y within $[y_{0.05}, y_{0.95}]$, along with L for y outside these limits, was used for the Bayesian likelihood function.

With the likelihood function defined, the Bayes formula in Equation 9 of Section 4.4 to the metadata set. Replace u , u' , and L by e , e' and L , respectively; then the posterior density $k\pi(e|e') = L(e'|e)\pi_0(e)$, where k represents a normalizing constant as before. The evidence (elicited estimate) e' is linked to the true but unknown value, e through the likelihood function $L(e'|e)$. If e , in turn, is distributed according to a parameter set θ , then we can partition on e :

$$k\pi\{\theta|e'\} = \Pr\{e'|\theta\}\pi_o(\theta) \quad (\text{Equation 38})$$

where \Pr denotes “density”

$$= \int_{\theta} \Pr\{e',e|\theta\}\pi_o(\theta)de = \int_{\theta} \Pr\{e'|\theta\}\Pr\{e|\theta\}\pi_o(\theta)de.$$

Replace $\Pr\{e'|\theta\}$ by $L(e'|e)$. Let $\theta=(\mu,\sigma)$. Assume e and e' have already been log-transformed. Suppose for the moment that e depends on θ through $e \sim N(\mu,\sigma)$. (It will be shown shortly that this is not, in fact, the case.)

Then the posterior density of θ , to within a normalizing constant k , is given by:

$$\pi(\mu, \sigma) \cdot k = \int_e L(e'|e) \frac{1}{\sigma\sqrt{2\pi}} \exp(-(e - \mu)^2 / 2\sigma^2) \pi_o(\mu, \sigma) d\mu \quad (\text{Equation 39})$$

For multiple e' values, $L(e'|e)$ is replaced by $\prod L(e'_i|e)$, $i=1,\dots,n$.

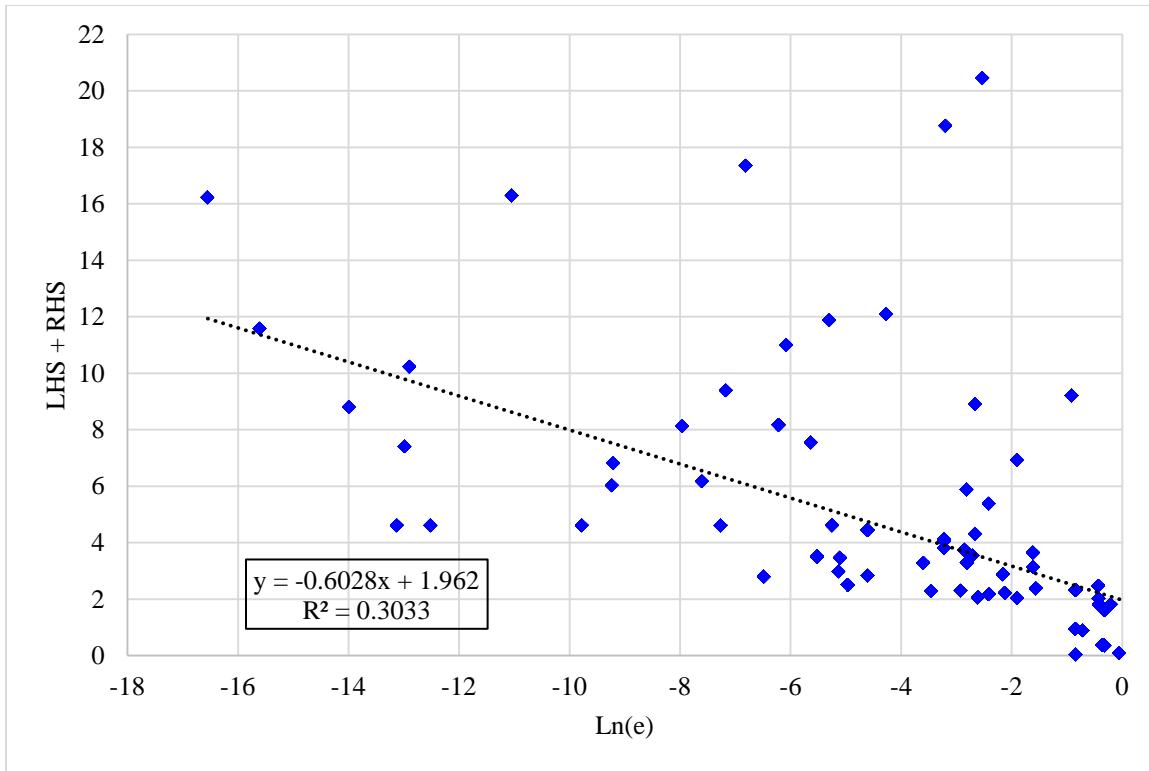
Given (μ, σ) , the posterior density for e would be obtained via

$$\pi(e) = \int_{\mu} \int_{\sigma} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(e - \mu)^2 / 2\sigma^2) \pi(\mu, \sigma) d\sigma d\mu. \quad (\text{Equation 40})$$

Given $\pi(e)$, a credible interval would be obtained by integrating it over appropriate limits.

The posterior density $\pi(e)$ was obtained assuming that e depended on θ through $e \sim N(\mu, \sigma)$, i.e., normal in these parameters. Is e , in fact, normally distributed? Since multiple observations of the realized value, e were not available for each variable in the EJE database the classical model performance weighted aggregated estimates of the 5th, 50th and 95th percentiles against each probabilistic and physical variable for which this was available were considered as proxies. After log-transforming the predictions, it was found that for probabilistic variables, the median of the ratios LHS/RHS over the individual records was approximately 1.43, where LHS represents the “left”-hand-side distance between the aggregated 5th and 50th percentiles, and RHS, the “right” hand side distance between the 50th and 95th percentiles. The calculation incorporated the weights associated with each variable. The lack of symmetry suggests that the normal is not an appropriate model. Further, the scatterplot of $d = \text{LHS} + \text{RHS}$ versus location, e (in log space) shown in Figure 16: Distance between Ln-transformed 95th and 5th Percentiles versus Ln(e) had an R^2 of 30% using a linear fit: $d = -0.6028 * e + 1.962$. The residuals were run through MathWave EasyFit to find a distribution which fit them.

Figure 16: Distance between Ln-transformed 95th and 5th Percentiles versus Ln(e)



MathWave EasyFit reports three measures of goodness of fit: Kolmogorov Smirnov (K-S), Anderson Darling (A-D), and Chi-Squared. The five-parameter Wakeby distribution ranked highest (1,1,1) against all three measures, but was discarded as over-parameterized. The four-parameter Burr distribution ranked (6,2,2) against these measures, respectively. The simpler two-parameter Cauchy distribution ranked (2,3,3) against these measures, respectively, and was selected as the fitting distribution for the residual spread. The two fitted Cauchy parameters were $C_\mu = -0.83964$ and $C_\sigma = 0.92965$. The Cauchy distribution has CDF given by:

$$\text{CDF}_{\text{Cauchy}}(x) = 0.5 + \frac{1}{\pi} \tan^{-1}((x - C_\mu)/C_\sigma) \quad (\text{Equation 41})$$

To find a distribution for the prior for location, μ , for probabilistic data, the values of e were run through MathWave EasyFit. The four-parameter Johnson SB distribution ranked (4,1,2) against the three goodness of fit measures, and was selected over three other distributions:

Kumaraswamy (2,5, N/A); Generalized extreme value distribution (5,2,1); and Generalized Pareto (3,37, and N/A).

The fitted Johnson SB has parameters $J_\gamma, J_\delta, J_\lambda, J_\xi$, where $J_\gamma = -1.7003$ and $J_\delta = 0.76428$ are shape parameters, and $J_\lambda = 22.355$ and $J_\xi = -22.095$ are scale and location parameters, respectively. The Johnson SB distribution has CDF given by:

$$\text{CDF}_{\text{Johnson SB}}(x) = \Phi(J_\gamma + J_\delta \cdot \ln(z/(1-z))), \quad (\text{Equation 42})$$

where $z = (x - J_\xi)/J_\lambda$, and Φ is the standard normal.

The total distance LHS_{tot} between the minimum (0th percentile) and mu, in Ln-space is assumed to equal (50/45) times the distance between the 5th and 50th percentiles, assuming e follows a uniform distribution over the left hand side. Such a distribution (log-uniform) has been used in EXCALIBUR. Similar reasoning applies on the total distance on the right hand side, RHS_{tot} .

Incorporating the usual normalizing constant k, the posterior density at a particular value of mu and residual distance, d is given by the integral over e of the likelihood function $L(e'|e)$, multiplied by the density of e given the prior for mu and d:

$$\pi(\mu, d) \cdot k = \int_{e \leq \mu} L(e'|e) \cdot 0.5/LHS_{\text{tot}}(\mu, d) \cdot \pi_o(\mu, d) de + \int_{e > \mu} L(e'|e) \cdot 0.5/RHS_{\text{tot}}(\mu, d) \cdot \pi_o(\mu, d) de$$

(Equation 43)

Given $\pi(\mu, d)$, the posterior density for e is obtained via

$$\pi(e) = \int_{\mu \geq e} \int_d 0.5/LHS_{\text{tot}}(\mu, d) \cdot \pi(\mu, d) ddd\mu + \int_{\mu < e} \int_d 0.5/RHS_{\text{tot}}(\mu, d) \cdot \pi(\mu, d) ddd\mu$$

(Equation 44)

The priors for , and d are as given above, for probabilistic data. Numerical integration was employed in Python using manual steps; the limits of integration were elo = -16.6; ehi = -0.05 (all parameters in Ln-space); the number of steps, ecount=10,000; mu ranged over these limits in

$\text{mucount} = 480$ steps. At each value of μ , the spread d is equal to the previously given linear fit, plus the residual drawn from the previously given Cauchy distribution, truncated between limits of -5.3 and 7.7 in $\text{rd0count} = 240$ steps. The total distance is constrained to be not less than 0.015 .

For physical variables, the median ratio RHS/LHS over the weighted individual records was equal to 0.91 ; a scatterplot of distance versus $\ln(e)$ had an R^2 of approximately zero (0.001%). The distribution of the spread, d between 5^{th} and 95^{th} obtained from the classical model subset of the physical data, was fit via MathWave EasyFit to a generalized Pareto distribution with parameters shape, scale and location parameters $GP_k=0.0740$, $GP_\sigma=2.8154$, and $GP_\mu=0.01588$, respectively. This distribution ranked (2,4,3); distributions which ranked higher on one of the goodness of fit measures ranked worse than 12 on others; and the five-parameter Wakeby, which ranked (1,5,2) was excluded. The generalized Pareto has distribution function given by:

$$\text{CDF}_{\text{Generalized Pareto}}(x)=1-[1+kz]^{-1/k}, \text{ where } z=(x-\mu)/\sigma; \text{ for } k=0, \text{CDF}=1-\exp(-z).$$

(Equation 45)

The prior for the location parameter for physical data was fit to the observed data via MathWave as a Laplace distribution with inverse scale and location parameters $L_\lambda=0.25207$ and $L_\mu=3.5894$, respectively. The Laplace has distribution function given by:

$$\text{CDF}_{\text{Laplace}}(x)=\frac{1}{2}\exp(-\lambda(\mu-x)), x \leq \mu; \text{ else } 1-\frac{1}{2}\exp(-\lambda(x-\mu)), x > \mu.$$

(Equation 46)

The Laplace ranked highest against all criteria in MathWave. A sample of the MathWave output for this case is shown in Appendix C: Sample of MathWave EasyFit Outputs.

As for probabilistic data, numerical integration in manual steps was employed for physical data, with limits of integration $\text{elo} = -18$, $\text{ehi} = 22$; $\text{ecount} = 5000$ steps, μ ranging over these

limits in mucount = 300 steps, spread d truncated between limits of 0.02 and 15.32 in rd0count = 230 steps. These parameter settings were chosen to stabilize 5th, 50th and 95th estimated percentiles of the posterior distribution of e, to within an average of 0.002 error in Ln-space. The initial coverage of the resulting 90% bounds was too large: approximately 97%.

One of the factors accounting for this excessive coverage is that the TUD subset upon which the physical spread was based, comprised 410 records out of the 1721 data records (56% by weight), and included some very large spreads, such as 10^{11} and $2.8 \cdot 10^9$, compared to the excursions between e' and e for the remainder of the physical data. All of the latter were less than 10^6 . By contrast, for probabilistic data, 89% by weight were TUD records. Additionally, for probabilistic data, aggregation incorporated a regression fit relating the spread to the magnitude of e. This tended to match larger spreads to smaller values of e, and smaller values to larger values of e approaching one. No such fit was available for physical data. This meant that at any magnitude of e within the domain of the prior, a large spread was equally likely to occur. Moreover, the prior for physical data ranged over approximately 2.5 times the width of its counterpart for probabilistic data, in log-space. In order to reduce the coverage percentage to approximately 90% for physical data, an additional “distance scaling factor”, dsf was applied to the spread distribution. A value of dsf=0.3 produced approximate 90% coverage over the 1721 data records.

MathWave EasyFit was applied to deweighted samples of Ln(e) and spreads, d between 5th and 95th in Ln-space, for each data type. For probabilistic data, the number of repetitions of each record in a theme was given by 864/[number of records in the theme]. This resulted in a sample of size 7773, with an average discrepancy of 0.1% between the weight given a record in the deweighted set, and the weight assigned it in the EJE based on the number of themes and the

number of records in each theme. For physical data records, the deweighting used $2880/[\text{number of records in each theme}]$. This resulted in a sample of size 123,765, with an average discrepancy of 0.15%. A single observation, e'3 in theme "Space Flight Risk", variable "ESTEC-1 Item 8" was removed, as an outlier. It had an MME exceeding 10^8 , and was smaller than any of the other three e' predictions for the record by a factor of 10^6 . Additionally, a single probabilistic record having four elicited values each exceeding e by a factor of 10^{11} , was excluded from the EJE at the start.

4.5.2 Aggregation via Alpha-Stable Distribution

A computer program in Python was written to fit an alpha-stable distribution to each data type (physical, probabilistic). First, an unweighted set of observations was created from each data set by replicating each (e, e') pair in each theme n times, where $n = [3200/m]$ and m is the total number of observations for the theme. (Using as numerator the least common multiple of the number of data pairs in each theme would have been exact, but would have required excessive run time; the discrepancy results averaged less than two percent. Runs using 1600, 1800 and 2400/m showed small differences in RMS deviation). This replication maintained approximate equal weighting across the themes, which was used in computing the empirical CDF. Next, $\ln(x)$ was computed for each observation, where $x = e/e'$. This transformation produced an approximately symmetric distribution centered near zero. Fitting an alpha-stable directly to this data produced a poor fit; the RMS error associated with the corresponding quantile-quantile plot (hereinafter, Q-Q plot) was approximately one. Therefore, a secondary power law transformation was applied to each $\ln(x)$ value as follows: if $\ln(x) < 0$, $|\ln(x)|$ was raised to a power, p_lt, and the result multiplied by negative one. Otherwise, $\ln(x)$ was raised to a different power, p_rt.

The following steps were performed inside a loop that tested these alternative secondary power law transformations. Note: in what follows, `levy.____` refers to a function in the “PyLevy” Python package used with α -alpha distributions.

1. A variate z was obtained for each transformation, defined by a pair of constants p_{lt} and p_{rt} as follows:

$$z = -[|\ln(x)|^{p_{lt}}] \text{ if } \ln(x) < 0; \text{ else } z = \ln(x)^{p_{rt}} \quad (\text{Equation 47})$$

2. An α -stable distribution was fit to the transformed observations via `levy.fit_levy`, which returned the parameters α , β , γ , and δ . Python returns a value of β having opposite sign from the S0 parameterization; the latter was used.
3. Given the parameters, an artificial sample of size 300,000 was computed from the resulting distribution via $\gamma\cdot\text{levy.alpha_stable_random}(\alpha,\beta,\text{shape}=300,000) + \delta$
4. The quantiles of both data sets—the variates resulting from the secondary transformation and the artificial sample—were computed along with the RMS deviation.

A Nelder-Mead downhill simplex algorithm was employed via the Python function `optimize.fmin`, to cycle through this loop to find the (p_{lt}, p_{rt}) pair minimizing the deviation. The resulting minimizing parameters using the Nolan (2009) S0 parameterization are provided in Table 5: Alpha-Stable parameters.

Table 5: Alpha-Stable parameters

Parameter	Probabilistic Data	Physical Data
α	2	1.9167
β	0.0017	0.4572
γ	0.8147	0.5861
δ	0.0001	-0.0045
p_{lt}	0.5125	0.5

Parameter	Probabilistic Data	Physical Data
p_rt	0.4938	0.5
RMS error	0.084	0.034
μ	0.0001	0.0308

The corresponding quantile-quantile (Q-Q) plots, parameterized in probability steps of 0.005 for the probabilistic and physical data sets are shown below in

Figure 17: Q-Q Plot of Probabilistic Data Alpha-Stable fit and

Figure 18: Q-Q Plot of Physical Data **Alpha-Stable fit** respectively.

Figure 17: Q-Q Plot of Probabilistic Data Alpha-Stable fit

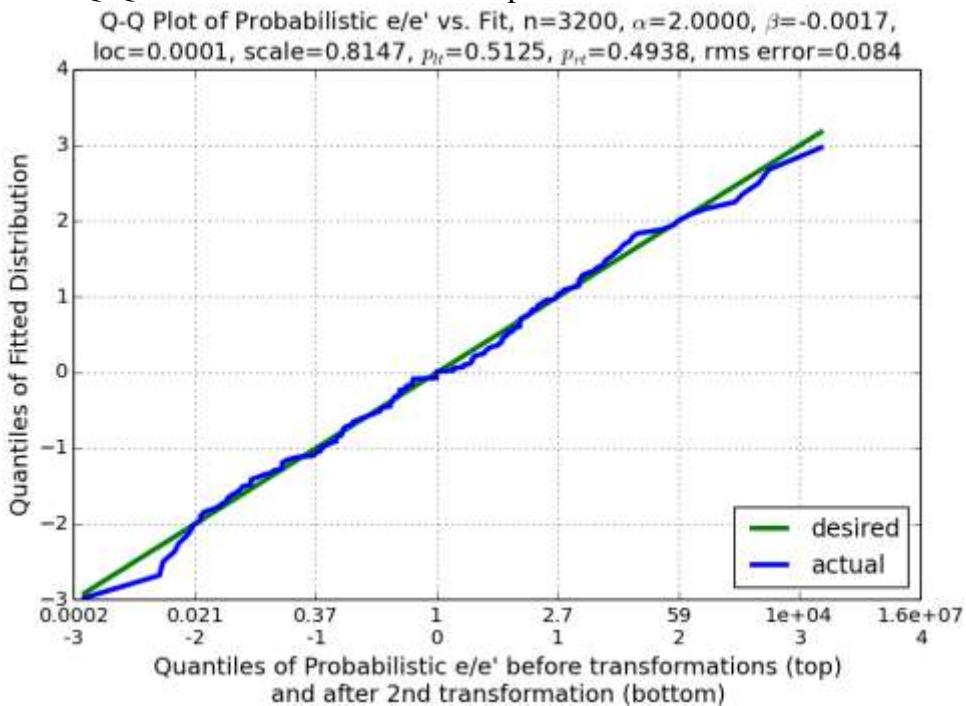


Figure 18: Q-Q Plot of Physical Data Alpha-Stable fit

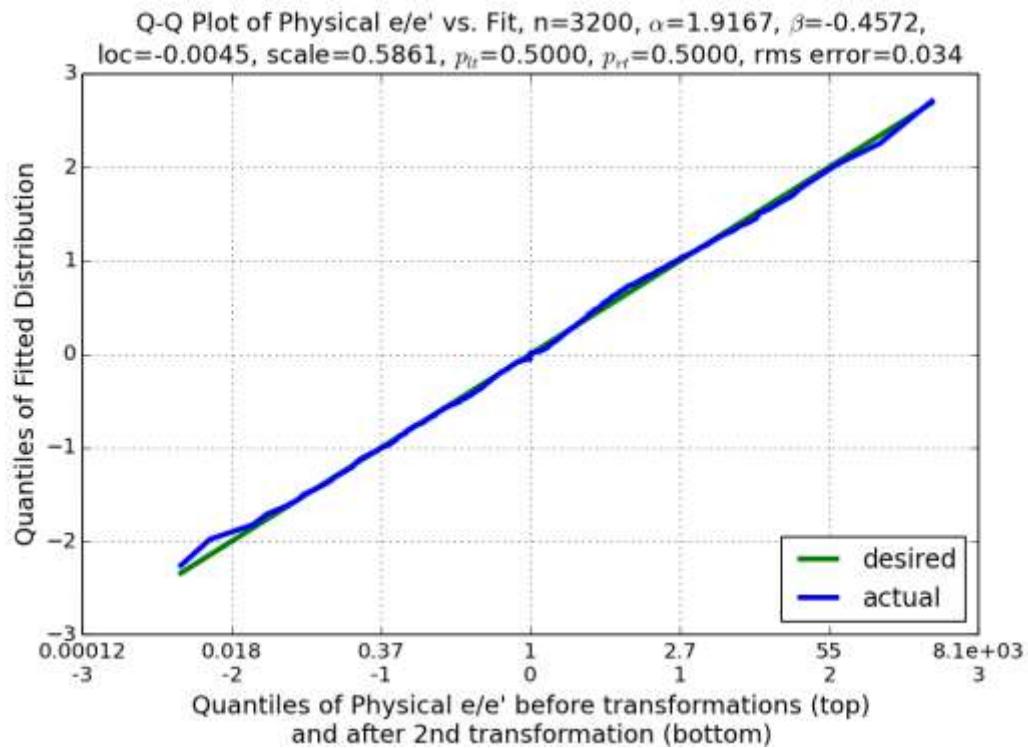


Figure 17 and

Figure 18 illustrate that the Alpha-Stable is a reasonable fit to the empirical distribution of the power-law transformed ratios e/e' . The first x-axis scale gives the ratio of e/e' ; the scale below it gives the ratio following the power law transformation. The $n e'$ values associated with an

individual EJE record of a given type, physical or probabilistic, can be aggregated into an estimate, \hat{e} using the following steps:

1. Set $x_k \equiv \ln(e^*/e_k)$, $k=1,2,\dots,n$, where e^* is defined below.
 2. Let $s_k \equiv \text{sign}(x_k)$.
 3. Let $q_k \equiv \{p_{\text{lt}} \text{ if } s_k < 0; \text{ else } p_{\text{rt}}\}$.
 4. Let $v_k \equiv (|x_k|^{\wedge} q_k) \cdot s_k$
 5. Let $S \equiv \sum v_k$
 6. Set \hat{e} equal to that value of e^* which causes S to equal $n \cdot \mu$; e^* can be found by binary search since S is monotonic in e^* . This follows from the fact that for $e^* \geq e_k$, $x_k > 0$ and $s_k = 1$, therefore increasing e^* causes v_k and S to increase. For $e^* < e_k$, increasing e^* causes $\ln(\hat{e}/e_k)$ to increase algebraically, therefore decrease in magnitude, since it is negative. After multiplication by $s_k = -1$, v_k increases algebraically, causing S to increase as well.
- The value of μ is given in Table 5 above according to data type.

Bounds can also be obtained around \hat{e} by exploiting the fact that sums of Alpha-Stable variates are themselves Alpha-Stable. The same steps 1–6 as above are applied, except that for a given bound, say the 90th, in step 6, \hat{e} is set equal to that value of e^* which causes S to equal S^* , where S^* is the 90th percentile of an Alpha-Stable distribution having the decay, skew, scale, and location parameters given previously in section 4.4 for a sum of n variates: $(\alpha, \beta, \gamma \cdot n^{1/\alpha}, d)$, respectively, where $d = n \cdot \delta + \gamma \cdot \beta \cdot (n^{1/\alpha} - n) \cdot \tan(\pi\alpha/2)$. Given that Python has a function which computes the CDF of an Alpha-Stable, both S^* and \hat{e} can be obtained via binary search.

4.5.3 Gaussian Mixture

The software package Dell Statistica 13 was used on deweighted sets of $\ln(e/e')$ data for each data type (physical and probabilistic) to obtain the Gaussian mixture parameters. The method was not useful for SeqIDs with small numbers of observations, n. In particular, for n=1 and probabilistic data, the one-sided multiplicative factor around \hat{e} for a 90% bound is approximately 64. This decreases to factors of approximately 7 and 5 at n=4 and n=6, respectively. The mixture parameters for physical and probabilistic data are shown in Table 6: Gaussian Mixture Parameters for Physical and Probabilistic Data.

Table 6: Gaussian Mixture Parameters for Physical and Probabilistic Data

Mixture parameters:	μ_1	σ_1	μ_2	σ_2	p	q
Physical data:	0.01353	0.17187	0.10815	2.09144	0.52574	0.47426
Probabilistic data:	0.02943	0.42617	0.08231	3.36037	0.53581	0.46419

4.5.4 Rule of Thumb (ROT)

A variant of the Alpha-Stable aggregation technique using $\alpha=2$ (Gaussian), $\mu=0$ and $p_{lt}=p_{rt}=0.5$ is denoted as “the rule of thumb” (ROT) approximation. It represents an additional aggregation method developed for this research. Results for ROT were similar to those for the Alpha-Stable.

4.5.5 Maximum Likelihood Estimate

The previously developed likelihood function was manually stepped through 10^6 iterations per record to find the location at which it reached a maximum. There was an average 0.02%

discrepancy between results at 10^6 versus 10^5 iterations. This constituted an additional aggregation method, although it did not yield associated bounds.

4.5.6 Classical model

The EXCALIBUR model was configured to produce 5th, 50th and 95th percentile estimates for both physical and probabilistic records. These represented the bounds along with the estimated median, \hat{e} .

4.5.7 Median, with bounds via quantile regression

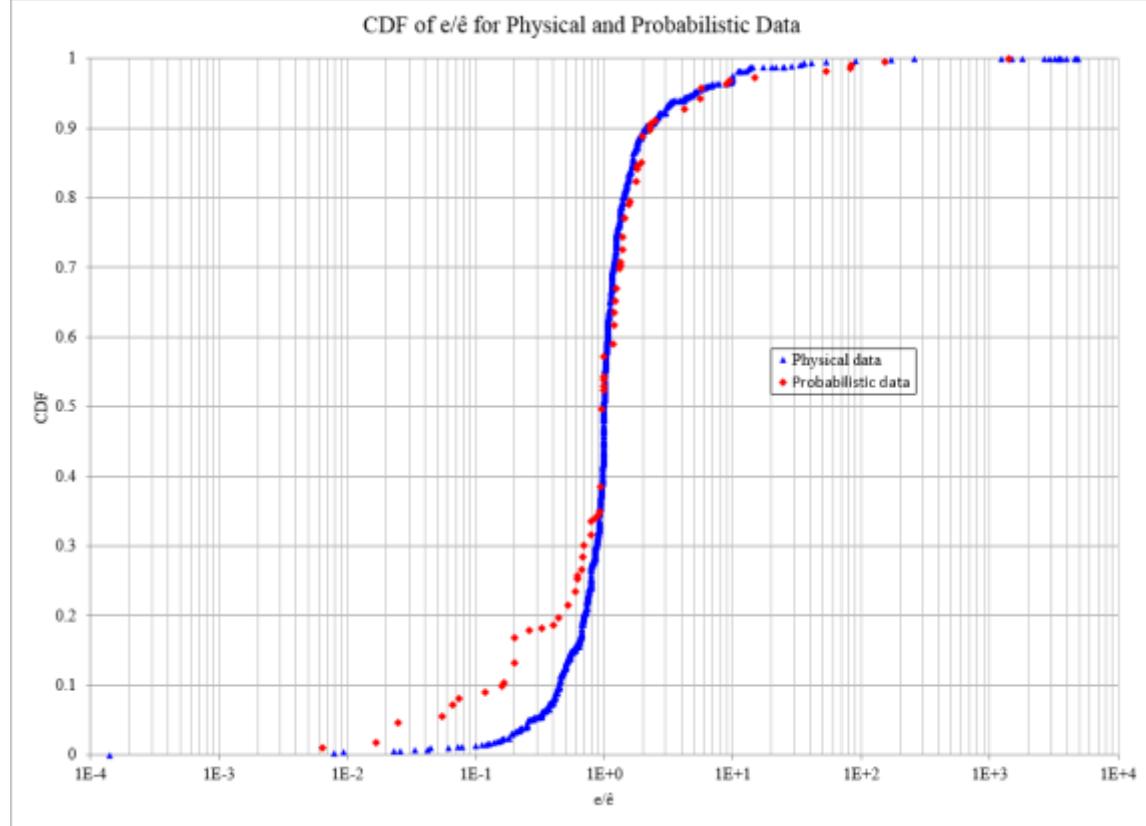
A final aggregation technique set \hat{e} equal to the median of the e' for each record. A quantile regression procedure paralleling that used to obtain quantiles of e' given e , was used to obtain quantiles of e given \hat{e} . A quadratic model was used as before, with the difference that \hat{e} was regarded as the independent variable, and the weight applied to each (\hat{e}, e) pair was the entire record weight instead of the record weight divided by the number of observations in the record. The resulting 5th and 95th quantiles were calculated at \hat{e} to obtain 90% confidence bounds. 10th and 90th quantiles were used to obtain 80% confidence bounds. Chapter 7 discusses bounds obtained by fitting simple parametric distributions to the \hat{e} values obtained from the above aggregation methods.

4.5.8 Accuracy of aggregated results

The computation of the CDF of e/\hat{e} parallels the approach discussed in Chapter 3 for the CDF of e/e' observations. Thus, $\Pr\{e/\hat{e} \leq x\}$ for a given value of x and method of aggregation applied to EJE physical data is equal to the sum of weights $\sum w_k$ over all physical records for which $e/\hat{e} \leq x$, where the mass w_k assigned to a record in a given theme equals $[n_{\text{themes}} \cdot n_{\text{vars in theme}}]^{-1}$. The CDF of e/\hat{e} for a given method of aggregation applied to probabilistic data is computed analogously, substituting nine (number of themes) for 43. For each data type, the CDF of the MME

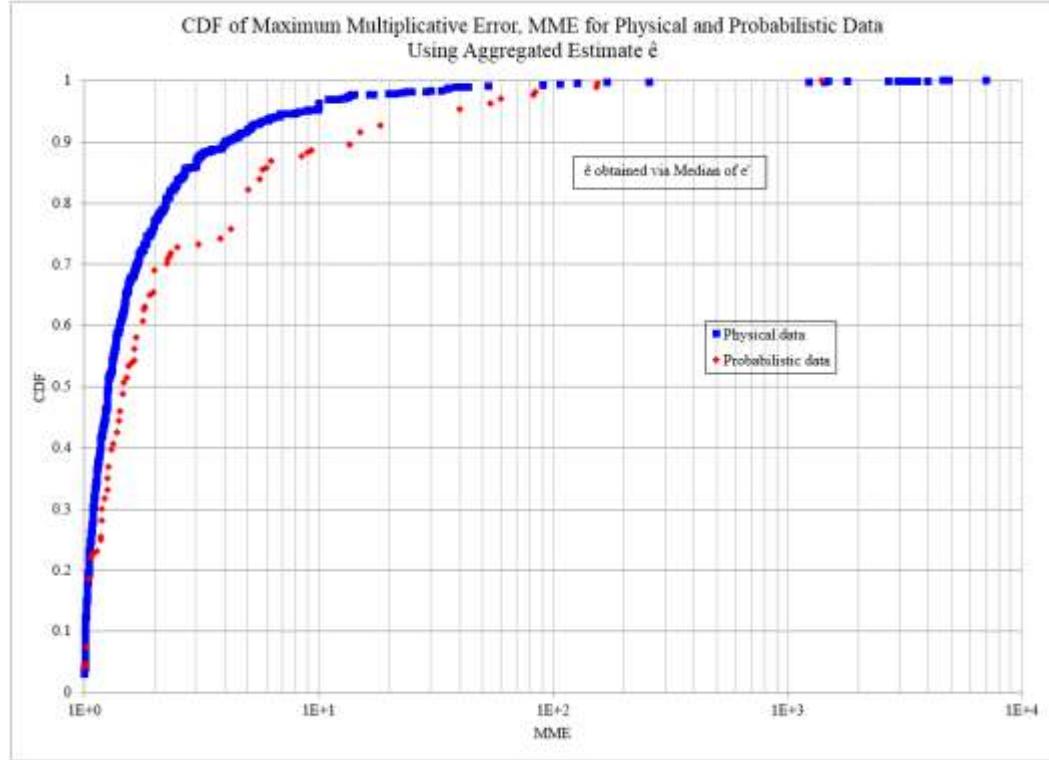
for e/\hat{e} parallels that used for the CDF of the MME of e/e' ; the exceedance probabilities are also computed analogously. Figure 19: CDF of e/\hat{e} for Physical and Probabilistic Data gives the CDF of e/\hat{e} based on one of the aggregation methods, the Median.

Figure 19: CDF of e/\hat{e} for Physical and Probabilistic Data



The CDF of the corresponding MME for this aggregation method is shown in Figure 20: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data Using Aggregated Estimate, \hat{e} .

Figure 20: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data Using Aggregated Estimate, \hat{e}



As discussed in Chapter 3, the chances of multiplicative excursions exceeding 2, 5, 10, 100, or 1000 decreased by factors of between 1.2 and 10, using this method of aggregation, versus relying on individual e' values, without any aggregation. Note that no excursion exceeds 10^4 , whereas when individual e' values were used, MMEs could reach $8 \cdot 10^7$.

4.5.9 Absolute distances between coverage percentages and 90% Over Each Theme

The ABSDist criterion defined in 4.1 is used as the metric of calibration quality of the bounds around \hat{e} . The methodology for computing ABSDist will be illustrated by way of an example. 90% bounds were obtained from the classical model for records in 24 themes. One of those themes, “ACNEXPTS”, has ten records, corresponding to SeqIDs EJEPHYS9–18, inclusive. Of these ten records, the 5th and 95th percentile points obtained from the classical model contained

the realized value, e in eight of them. The realized value fell outside of the bounds for two records in this theme. Therefore, the coverage of the classical model for this theme was eight out of ten, or 80%. The delta from the 90% coverage objective is -0.1 , and its absolute value, ABSDist equals 0.1.

For the Bayesian aggregation method applied to that same theme, nine of ten records fell within its 90% bounds, therefore the coverage was 90%, and ABSDist equals zero. ABSDist will be reported, as “ABSDist (Coverage–0.9)” for each aggregation method. “ABSDist (Coverage–0.8)” will also be reported, for 80% bounds. Only 90th percentile bounds were available for the classical method; “ABSDist (Coverage–0.9)” will be the second criterion against which the methods are ranked. Note that an aggregation method which achieves 90% coverage by covering half of its themes at 80%, and the other half at 100%, would have twice the ABSDist of a competing method which covered each theme at $90\% \pm 5\%$.

4.5.10 Sensitivity to Outliers

Per NIST (2010) for a given record with observations e'_i and median \bar{e} , let

$M_i = 0.6745(e'_i - \bar{e})/\text{MAD}$, where MAD≡Median Absolute Deviation and M_i is considered a modified Z-score. (MAD is the median of the $|e'_i - \bar{e}|$.) If $|M_i| > 3.5$, declare e'_i to be an outlier. In order to bring the e'_i closer to normality before applying the outlier test, they are first transformed via the power law approximation associated with the Alpha-Stable, $|\ln(e/e'_i)|^{0.5} \cdot \text{sgn}(\ln(e/e'_i))/\sigma$, where $\sigma=0.83$ and 1.15 for physical and probabilistic data, respectively. This resulted in the identification of 21 and 17 records containing outliers for the two data types, respectively. (Classical model results were not available for three of the outlier records, therefore MMEs were averaged over the remaining 18 records for this method.) As an example of a physical record containing outliers,

SeqID EJEPHYS1719 (magma production rate in Holocene, 1000s tons per year), true value $e=3600$, there were 45 e^t observations. The first 37 of these ranged from 1.1 to 10,000 and averaged 1481. The last eight ranged from 30,000 to 10^8 and were considered outliers. SeqID EJEPHYS1720 had four outliers, all others had two or less. Of the probabilistic outlier records, one had three outliers, and all others had one or two only.

Table 7: MME Scorecard for EJE Physical Data and Table 8: MME Scorecard for EJE Probabilistic Data provide MME ranking information for EJE Physical and Probabilistic Data, respectively. Table 9 includes the averages and median MMEs computed for the Classical method. The minimum MMEs for the 410 and 409 record sets differ by less than 10%. The difference for the maximum MMEs is approximately 42%. Dropping the record with the highest MME did not affect the rank of the six best aggregation methods; the Arithmetic Mean and Classical methods interchanged second and third-worst rank places. Harmonic Mean performed worst. The largest percentage decrease between the two sets of MMEs is associated with the Classical Method (46%) followed by the Harmonic Mean. The Median aggregation method dominated both record sets, followed by the Alpha Stable. For practical purposes, the Median, Alpha Stable, ROT, and Classical methods tied for the highest rank with respect to the median (approximately 1.4). The difference between the maximum and minimum median MME across methods is about 15%.

In common with Table 9, Table 10 includes the Classical method. The average MME for 66 records ranges from 16 for the Median aggregation method, to 132,809 for the Classical. When the record with highest MME for any given aggregation method is dropped (65 records), the average MME ranges from 8 for the Bayesian method, to 3028 for the Harmonic Mean. The largest percentage decrease between the two sets of MMEs is associated with the Classical Method, followed by the Arithmetic Mean. With respect to the median, for practical purposes, the MLE and Median methods tied for highest rank at approximately 1.62. The difference between the maximum and minimum median MME across methods is about 30%, twice its counterpart for physical TUD data.

A comparison of the two sets of physical data (Table 7 and Table 9) show that with the exception of the Arithmetic Mean, the common methods of aggregation yielded smaller average MMEs for the smaller data set. The same trend was identified when the record with the highest MME for any given aggregation method was dropped. The Harmonic Mean had highest MME; Classical and Arithmetic Mean method had the second and third worst performances with respect to average MME. The Median aggregation method performed best; the Bayesian takes second rank for the 1721 and 1720 data sets, whereas this rank is attributable to the Alpha Stable for the 410 and 409 record sets.

A comparison of the two sets of probabilistic data (Table 8and Table 10) shows that the ranks of the various aggregation methods based on average MME are identical for the two sets; the Classical performs worst against the latter. When the highest MME record is dropped, the Classical has the largest reduction, and moves up to 7th place. This reflects the contribution to the weighted average of a single record having a very large MME. The Bayesian and Median aggregation methods have the best performance; dropping the worst record reduced the average MMEs for these two methods by factors of three and two, respectively. The Harmonic Mean performed worst or second worst in both cases. With respect to the median, the Classical ranked in fourth place, approximately 7% higher than the two best methods (MLE and Median aggregation).

The average MMEs for the common methods of aggregation are consistently higher in the TUD set for the 66 and 65 record sets compared with the respective EJE 67 and 66 record sets. These two datasets differ by a single theme (SeqID EJEPROB64) having a relatively low MME, which is not present in the TUD set.

Table 9: MME Scorecard for EJE Physical Data- TUD Records and Table 10: MME Scorecard for EJE Probabilistic Data- TUD Records provide the rankings for their TUD counterparts, respectively. The ranking was performed with the objective of obtaining a relative measure of accuracy among the aggregation methods. Within a given data set and aggregation method, the ranks for the average (weighted sum product) and the median (50th percentile) based on all records in the data set were computed, and then the average was recomputed when the highest MME was dropped from each aggregation method.

With respect to Table 7, there was no difference in the ranks for either average computation. The Median aggregation method dominated in the sense of having lowest MME, and the Harmonic Mean came in last. However; the percentage difference between the Harmonic Mean taken over 1,721 records compared with 1720 records was the largest decrease, ~60% compared with all the other methods, which were all under 3%. With the exception of the Arithmetic Mean and Harmonic Mean, the average MMEs for the other methods ranged from 7.88 to 9.42 for 1,721 records, and from 7.65 to 9.18 for 1,720 records, i.e., approximately 16% difference between best and worst aggregation method in each record set. The medians ranged from 1.25 (a first-place tie was incurred for three methods) to 1.32, a difference of approximately 5%.

Table 8 shows a different trend from Table 7. The same MME rankings are not preserved for the 67 records and the 66 records. The Median and Bayesian rank 1 and 2 respectively for 67 records, but the ranks are interchanged for the 66 records. Notably, the respective percentage decreases are 46% and 72%. For 67

records, the MMEs range from 14 to 7889 across aggregation methods; for 66 records, they range from 7 to 2669. Thus, overall, dropping the record containing the highest MME for any particular aggregation method caused the overall minimum and maximum MMEs to drop by factors of two and three, respectively. The medians ranged from 1.46 to 2.01, resulting in a difference of over 25%. The aggregation methods having the best performance with respect to this statistic were Median and MLE, at 1.46 and 1.48, respectively.

Table 7: MME Scorecard for EJE Physical Data

MME Physical	No. Records	Arithmetic Mean	Geometric Mean	Harmonic Mean	Median	Alpha-Stable	Rule of Thumb	MLE	Bayesian
Average	1,721	27.75	8.82	202.61	7.88	8.43	8.66	9.42	8.37
Average Rank	1,721	7	5	8	1	3	4	6	2
Average*	1,720	27.52	8.59	82.65	7.65	8.21	8.43	9.18	8.17
Average*Rank	1,720	7	5	8	1	3	4	6	2
Median	1,721	1.319	1.277	1.318	1.26	1.25	1.25	1.25	1.28
Median Rank	1,721	8	5	7	4	3	3	3	6
* Max MME eliminated									

Table 8: MME Scorecard for EJE Probabilistic Data

MME Probabilistic	No. Records	Arithmetic Mean	Geometric Mean	Harmonic Mean	Median	Alpha-Stable	Rule of Thumb	MLE	Bayesian
Average	67	327	31.19	7888.72	14.04	26.68	28.95	115.3	25.82
Average Rank	67	7	5	8	1	3	4	6	2
Average*	66	12.97	18.49	2668.76	7.56	15.76	15.59	52.17	7.22
Average*Rank	66	3	6	8	2	5	4	7	1
Median	67	1.59	1.89	2.01	1.46	1.62	1.62	1.48	1.62
Median Rank	67	3	7	8	1	5	5	2	4
* Max MME eliminated									

Table 9 includes the averages and median MMEs computed for the Classical method. The minimum MMEs for the 410 and 409 record sets differ by less than 10%. The difference for the maximum MMEs is approximately 42%. Dropping the record with the highest MME did not affect the rank of the six best aggregation methods; the Arithmetic Mean and Classical methods interchanged second and third-worst rank places. Harmonic Mean performed worst. The largest percentage decrease between the two sets of MMEs is associated with the Classical Method (46%) followed by the Harmonic Mean. The Median aggregation method dominated both record sets, followed by the Alpha Stable. For practical purposes, the Median, Alpha Stable, ROT, and Classical methods tied for the highest rank with respect to the median (approximately 1.4). The difference between the maximum and minimum median MME across methods is about 15%.

In common with Table 9, Table 10 includes the Classical method. The average MME for 66 records ranges from 16 for the Median aggregation method, to 132,809 for the Classical. When the record with highest MME for any given aggregation method is dropped (65 records), the average MME ranges from 8 for the Bayesian method, to 3028 for the Harmonic Mean. The largest percentage decrease between the two sets of MMEs is associated with the Classical Method, followed by the Arithmetic Mean. With respect to the median, for practical purposes, the MLE and Median methods tied for highest rank at approximately 1.62. The difference between the maximum and minimum median MME across methods is about 30%, twice its counterpart for physical TUD data.

A comparison of the two sets of physical data (Table 7 and Table 9) show that with the exception of the Arithmetic Mean, the common methods of aggregation yielded smaller

average MMEs for the smaller data set. The same trend was identified when the record with the highest MME for any given aggregation method was dropped. The Harmonic Mean had highest MME; Classical and Arithmetic Mean method had the second and third worst performances with respect to average MME. The Median aggregation method performed best; the Bayesian takes second rank for the 1721 and 1720 data sets, whereas this rank is attributable to the Alpha Stable for the 410 and 409 record sets.

A comparison of the two sets of probabilistic data (Table 8and Table 10) shows that the ranks of the various aggregation methods based on average MME are identical for the two sets; the Classical performs worst against the latter. When the highest MME record is dropped, the Classical has the largest reduction, and moves up to 7th place. This reflects the contribution to the weighted average of a single record having a very large MME. The Bayesian and Median aggregation methods have the best performance; dropping the worst record reduced the average MMEs for these two methods by factors of three and two, respectively. The Harmonic Mean performed worst or second worst in both cases. With respect to the median, the Classical ranked in fourth place, approximately 7% higher than the two best methods (MLE and Median aggregation).

The average MMEs for the common methods of aggregation are consistently higher in the TUD set for the 66 and 65 record sets compared with the respective EJE 67 and 66 record sets. These two datasets differ by a single theme (SeqID EJEPROB64) having a relatively low MME, which is not present in the TUD set.

Table 9: MME Scorecard for EJE Physical Data- TUD Records

MME Physical	No. Records	Arithmetic Mean	Geometric Mean	Harmonic Mean	Median	Alpha-Stable	Rule of Thumb	MLE	Bayesian	Classical
Average	410	40.93	5.58	133.32	4.53	5.39	5.62	7.82	6.95	44.32
Average Rank	410	7	3	9	1	2	4	6	5	8
Average*	409	29.95	4.14	77.89	4.08	4.13	4.23	4.67	4.26	23.81
Average* Rank	409	8	3	9	1	2	4	6	5	7
Median	410	1.591	1.444	1.617	1.4	1.398	1.398	1.413	1.44	1.399
Median Rank	410	8	7	9	4	1	2	5	6	3
* Max MME eliminated										

Table 10: MME Scorecard for EJE Probabilistic Data- TUD Records

MME Probabilistic	No. Records	Arithmetic Mean	Geometric Mean	Harmonic Mean	Median	Alpha-Stable	Rule of Thumb	MLE	Bayesian	Classical
Average	66	367.75	34.96	8874.68	15.66	29.89	32.44	129.5	28.87	132808.9
Average Rank	66	7	5	8	1	3	4	6	2	9
Average*	65	14.62	20.78	3017.95	8.42	17.69	17.5	58.6	7.96	28.04
Average* Rank	65	3	6	9	2	5	4	8	1	7
Median	66	1.677	2.018	2.23	1.63	1.86	1.86	1.62	1.78	1.74
Median Rank	66	3	8	9	2	7	6	1	5	4
* Max MME eliminated										

As stated in the introduction to the chapter, calibration quality is reflected in the second ranking criterion, average absolute value of the difference between the coverage percentage over each theme, and 0.9: ABSDist (Coverage–0.9). ABSDist (Coverage–0.8) is also reported.

Tables 11-16 provide scorecards for the ABSDist criterion, for both physical (24 and 43) and probabilistic (8) themes. (One of the nine EJE probabilistic data themes consists of a single record and is not included in the TUD probabilistic data; all methods cover this record, therefore ABSDist is reported and ranked for the eight TUD themes only.) For both physical and probabilistic TUD data, the Bayesian had the best performance, i.e. smallest ABSDist value, with respect to both 90% as well as 80% bounds. The Classical (where applicable, i.e. for TUD data against 90% bounds) and the Median aggregation methods came in second and third place, respectively in all of the tables. (Where the Classical method did not apply, the Median came in second place.) Aggregation via Alpha stable, ROT, and Geometric Mean yielded the worst results, in terms of highest ABSDist values.

For probabilistic data, the Bayesian average coverage with respect to 90% and 80% bounds was 93% and 83%, respectively. Its ABSDist(Coverage–0.9) value was 0.09; Classical and Median aggregation both had 0.11 ABSDist values. Their coverage with respect to 90% bounds was 88% and 85%, respectively. Geometric, Alpha stable and ROT ranked 4,5, and 6, respectively, against the ABSDist criterion; the latter's values for these three methods were three times larger than that for the Bayesian. Coverages

were relatively low: approximately 60% and 50% against 90% and 80% bounds, respectively.

Similar results were obtained for the physical data set. The Bayesian average coverage with respect to 90% and 80% bounds for the 24-theme TUD data was 90% and 82%, respectively. Classical and Median coverages for this data with respect to 90% bounds were 86% and 87%, respectively. ABSDist(Coverage=0.9) value was 0.095 for the Bayesian; 0.101 and 0.11 for the Classical and Median, respectively. ABSDist values were approximately twice as large for the Geometric, Alpha stable and ROT methods, which again had relatively low coverages of approximately 60% and 50% against 90% and 80% bounds. Similar values and rankings applied over the full 43 theme physical data set.

It must be noted that although the Bayesian had the smallest ABSDist value, it had approximately twice the median one-sided multiplicative bounds width of the Classical model, for both physical and probabilistic data. The Median aggregation method had a width approximately 20% greater than the Classical, while the Alpha stable and ROT widths were approximately half that of the Classical. However, the relatively narrow bounds associated with the Alpha stable and ROT yielded relatively low coverage, as discussed above. The Geometric Mean represented an intermediate case; its widths ranged from three quarters to 90% of their Classical counterparts; its average coverage was approximately 80% of the Classical value. However, this coverage was more uneven: the ABSDist values for the Geometric Mean were at least twice their Classical counterparts.

Table 11: EJE Physical Data TUD Records - 90% Coverage by Aggregation Method

Physical Data TUD 24 themes: 90% Bounds Coverage by Method	Geometric	Median	Alpha-Stable	ROT	Bayesian	Classical
ABSDist (Coverage=0.9)	0.24	0.11	0.36	0.33	0.095	0.101
Rank	4	3	6	5	1	2
Average Coverage Over All Themes	0.70	0.87	0.57	0.61	0.90	0.86
Minimum	0	0.44	0	0	0.63	0.5
Maximum	1	1	1	1	1	1
Median One-sided Multiplicative Bounds Width	2.46	3.43	1.62	1.74	5.24	2.74
Rank	3	5	1	2	6	4

Table 12: EJE Physical Data - 90% Coverage by Aggregation Method

Physical Data 43 themes: 90% Bounds Coverage by Method	Geometric	Median	Alpha-Stable	ROT	Bayesian	Classical
ABSDist (Coverage=0.9)	0.19	0.10	0.27	0.25	0.093	N/A
Rank	3	2	5	4	1	
Average Coverage Over All Themes	0.80	0.90	0.71	0.73	0.93	
Minimum	0	0.44	0	0	0.63	
Maximum	1	1	1	1	1	
Median One-sided Multiplicative Bounds Width	2.65	3.36	1.84	2.02	5.37	
Rank	3	4	1	2	5	

Table 13: EJE Physical Data TUD Records – 80% Coverage by Aggregation Method

Physical Data TUD 24 Themes: 80% Bounds Coverage by Method	Geometric	Median	Alpha-Stable	ROT	Bayesian	Classical
ABSDist(Coverage-0.8)	0.24	0.19	0.36	0.34	0.14	N/A
Rank	3	2	5	4	1	N/A
Average Coverage Over All Themes	0.65	0.73	0.50	0.52	0.82	N/A
Minimum	0.00	0.25	0.00	0.00	0.25	N/A
Maximum	1.00	1.00	1.00	1.00	1.00	N/A

Table 14: EJE Physical Data - 80% Coverage by Aggregation Method

Physical Data 43 themes: 80% Bounds Coverage by Method	Geometric	Median	Alpha-Stable	ROT	Bayesian	Classical
ABSDist (Coverage-0.8)	0.21	0.16	0.30	0.29	0.15	N/A
Rank	3	2	5	4	1	
Average Coverage Over All Themes	0.76	0.80	0.64	0.66	0.86	
Minimum	0	0.25	0	0	0.25	
Maximum	1	1	1	1	1	

Table 15: EJE Probabilistic TUD Data - 90% Coverage by Aggregation Method

Probabilistic Data TUD 8 Themes: 90% Bounds Coverage by Method	Geometric	Median	Alpha-Stable	ROT	Bayesian	Classical
ABSDist (Coverage-0.9)	0.32	0.11	0.38	0.45	0.09	0.11
Rank	4	3	5	6	1	2
Average Coverage Over All Themes	0.66	0.85	0.57	0.45	0.93	0.88
Minimum	0.33	0.50	0.17	0.17	0.75	0.67
Maximum	1	1	1	0.75	1	1
Median One-sided Multiplicative Bounds Width	3.14	4.97	2.13	1.96	9.28	4.20
Rank	3	5	2	1	6	4

Table 16: EJE Probabilistic TUD Data - 80% Coverage by Aggregation Method

Probabilistic Data TUD 8 Themes: 80% Bounds Coverage by Method	Geometric	Median	Alpha-Stable	ROT	Bayesian	Classical
ABSDist (Coverage-0.8)	0.34	0.15	0.35	0.36	0.14	N/A
Rank	3	2	4	5	1	
Average Coverage Over All Themes	0.61	0.73	0.45	0.44	0.83	
Minimum	0.25	0.50	0.17	0.17	0.50	
Maximum	1	1	0.75	0.75	1	

The fourth criterion is sensitivity to outliers. The tables below provide the average of MMEs for different aggregation methods over those records containing outliers (21 physical and 17 probabilistic data records). The rankings are not particularly informative for probabilistic data: for all methods except aggregation via median, average MMEs over the outliers are the same or less than their counterparts over the full data set. Even for aggregation via the median, the difference between its average over outliers and its average over all probabilistic records is an artifact arising from a single outlier record, SeqID EJEPROB45, with true value $e=0.00077$, where the outlier (0.001) is both the largest as well as the closest e' value to e , and is ignored by the median, which reflects five other e' values ranging from 10^{-4} to 10^{-6} . It may be noted that for this particular record, the other small e' values “pulled” the harmonic mean further away from e , since it gives larger weight to small values; however, the resulting MME for this record of 232, while larger than the 154 obtained for the median, was masked by even larger MMEs in the full data set.

The Arithmetic Mean was minimally affected by the small values: it is not particularly sensitive to them, except to increase the count of observations by which the sum of the values must be divided. Its MME was 4 for this record. The peak of the likelihood function was pulled away from e by the numerous small values, and had the largest MME for this record, at 752. However, this did not stand out compared to other large MMEs in the full data set.

Bayesian aggregation, while incorporating the likelihood function, does not simply use its peak, but integrates it in conjunction with priors for location and spread. The JohnsonSB prior for location has a PDF five times higher at 0.001 than at 10^{-6} ; this damped the

impact of the small values on the likelihood function, and the resulting MME for the Bayesian was 26 for this record. The Alpha-Stable, rule of thumb, and geometric mean had intermediate MME values between 50 and 100 for this record.

The classical method had by far the best performance against this record, with an MME of 1.02: its prediction reflected a weighting of 75% on the expert whose prediction was closest. On the other hand, the classical model has a very large average MME of 132808.86 over the full probabilistic data set, arising from a single record, SeqID EJEPROB51, with $e=0.08$, and a single outlier e' value of 10^{-9} . The other three e' values ranged between 0.001 and 0.1. However, the distribution associated with the expert predicting 10^{-9} was given the majority of the weight by the classical model, based on calibration against other variables. The value of 10^{-9} was subsequently discarded as an outlier before performing quantile regression; this is why this record does not appear as an outlier. It is also the reason for showing median as well as average MME excluding the largest MME in the preceding MME scorecard tables.

For physical data outliers, the harmonic and arithmetic means had the largest average as well as median MME values against outliers. The harmonic mean's largest single MME of 9600 occurred at SeqID EJEPHYS1178, with true value $e=7.9 \cdot 10^7$. All of the 12 e' observations for this record ranged between 10^6 and 10^9 , except for a single relatively extremely small e' value of 750, which had predominant influence on the harmonic mean and gave rise to the large MME value. The arithmetic mean had large MME values of 1667 and 1867 at SeqIDs EJEPHYS1187 and EJEPHYS1720, respectively. For the first of these, the mean was pulled away from the realized e value of 500 by a single outlier $e'=5 \cdot 10^6$; all other e' values ranged between 10 and 1000 for this

record. For the second of these outlier records, the mean was pulled off of the e value of 12 by a single large $e'=10^6$; no other e' value exceeded 3000, and some values were as small as 0.01. As stated above, the harmonic and arithmetic means had the largest median MME values against outliers: approximately 26 and 18, respectively. Medians for other aggregation methods were confined to a narrower band between approximately 3 and 5.5. Of the other aggregation methods, the classical model and the MLE had the highest average MME values; the Geometric Mean and Alpha-Stable, the lowest. The median MME against outliers was also lowest for the Geometric Mean and Alpha-Stable

Table 17: Average MME by Aggregation Method for Record with Outliers – Physical Data

MME Physical	Arithmetic	Harmonic	Median	Bayesian	ROT	Alpha-Stable	MLE	Geometric Mean	Classical
Average	265.49	559.68	10.32	13.90	7.54	7.32	16.86	6.60	36.65
Rank	8	9	4	5	3	2	6	1	7
Median	17.62	26.47	4.55	5.43	3.41	3.60	5.01	2.98	4.64
Max	1,867.10	9,609.49	43.65	81.53	34.34	33.26	95.62	30.13	509.55

Table 18: Average MME by Aggregation Method for Records with Outliers – Probabilistic Data

MME Probabilistic	Arithmetic	Harmonic	Median	Bayesian	ROT	Alpha-Stable	MLE	Geometric Mean	Classical
Average	16.58	44.00	23.65	10.01	18.82	18.71	52.46	16.75	34.50
Rank	2	8	6	1	5	4	9	3	7
Median	3.23	2.23	5.00	5.33	4.72	4.73	5.03	3.97	1.88
Max	81.95	262.50	154.00	36.97	99.87	97.66	752.27	125.61	455.04

4.6 Summary

Table 19: Complexity Rating, Table 20: Physical Data Scorecard, and Table 21: Probabilistic Data Scorecard are used to summarize Research Question 3.

provides the complexity rating for the methods of aggregation used in this research.

Complexity is used to represent comparative ease of computation. These ratings are included in Table 20 and Table 21.

Table 20 and Table 21 provide the rankings for six methods of aggregation applied to the TUD physical and probabilistic data for the following five criteria: accuracy, calibration, informativeness; sensitivity to outliers, and complexity. The Arithmetic Mean, Harmonic Mean, and MLE were dropped from consideration at this stage since they generally under-performed.

The Average* and Median MME on a record-weighted basis (per Chapter 3) are used to represent the accuracy criterion. The Average* represents the reweighted average after the largest MME is eliminated, since such a value could dominate the average. In terms of accuracy measured by the Average*, the Median and the Bayesian ranked first for the physical and probabilistic data respectively, and in terms of the Median, the Alpha-Stable and the ROT tied for first place for the physical data and the Median ranked first for the probabilistic data.

The calibration is represented by the absolute distance, ABSDist, between the coverage and 0.9, which was averaged over all the themes associated with a given data set, resulting in the metric, denoted as ABSDist(Coverage–0.9). The Bayesian ranked first in each data set.

Informativeness is represented by the one-sided multiplicative bounds width. The Alpha Stable and ROT ranked in the first and second places, respectively for the physical data and exchanged ranks for the probabilistic data. All other aggregation methods resulted in the same ranks for the data sets.

Table 19: Complexity Rating

Rank	Explanation	Aggregation Methods
1	Aggregation Method is provided as-is in office automation software.	Arithmetic Mean, Geometric Mean, Harmonic Mean, and Median.
2	Aggregation technique requires code that is reused by other aggregation techniques, e.g., Python and/or is less resource-intensive than development required for Rank 3.	Rule of Thumb, Maximum Likelihood Expectation, and Alpha-Stable.
3	Aggregation technique requires code leveraged from other techniques and additional development or requires a stand-alone development effort that is MORE resource-intensive than development required for Rank 2. The classical technique requires setting a cutoff parameter for expert calibration so as to optimize the product of a calibration and information score for a synthetic decision maker.	Bayesian, Cooke

Table 20: Physical Data Scorecard

Physical TUD	Geometric Mean	Median	Alpha-Stable	Rule of Thumb	Bayesian	Classical
Average * Rank	3	1	2	4	5	6
Median Rank	6	4	1	1	5	3
ABSDist (Coverage=0.9) Rank	4	3	6	5	1	2
Median One-sided Multiplicative Bounds Width Rank	3	5	1	2	6	4
Outliers	1	4	2	3	5	6
Complexity	1	1	2	2	3	3
Σ Ranks	18	18	14	17	25	24
Placement	3	3	1	2	6	5

Table 21: Probabilistic Data Scorecard

Probabilistic	Geometric Mean	Median	Alpha-Stable	Rule of Thumb	Bayesian	Classical
Average * Rank	5	2	4	3	1	6
Median Rank	6	1	4	4	3	2
ABSDist (Coverage=0.9) Rank	4	2	5	6	1	2
Median One-sided Multiplicative Bounds Width Rank	3	5	2	1	6	4
Outliers	2	5	3	4	1	6
Complexity	1	1	2	2	3	3
Σ Ranks	21	16	20	20	15	23
Placement	5	2	3	3	1	6

Chapter 5: Research Question 3 (Magnitude of e)

5.1 Introduction

Does the magnitude of the quantity estimated impact the range of multiplicative error? How does the MME change between estimates of infrequent events, and estimates of frequent events? Does one aggregation method perform better than others for values of e falling in a certain region, e.g. for small values of e representing relatively infrequent events?

Figure 14 and Figure 15 showed quantiles of e' versus e for probabilistic and physical data, respectively. The figures demonstrate that the spread between quantiles changes with magnitude of e .

Figure 14 showed a spread of approximately 15 units between the 0.05 and 0.95 quantiles of e' given e , at $e=-15$ (both e and e' Ln-transformed), for probabilistic data. Approximately two thirds of the spread is above e ; one third below, at this point. This is consistent with the known tendency of elicitees to overestimate small values of e for probabilistic data. The spread narrows to approximately one unit at $e=-0.05$, with the

bulk of the spread necessarily below e at this point. Narrower spreads are observed for physical data, but over a wider range of e values: nine units at $e=-18$ (with two thirds of the spread below the value of e); decreasing to four units, split approximately evenly above and below e at $e=2$; then increasing to six units at $e=22$, almost all of which is below e at this point.

The varying distances between quantiles impact the chances of over- or underestimating e by given factors, depending on the magnitude of e . Table 3 and Table 4 respectively, gave over- or underestimation chances by various factors, both before and after aggregation (in this case, aggregation via the median of the e' values), but without considering the impact of the magnitude of e on these chances. They were simply averaged over all occurrences of e in the EJE probabilistic and physical data sets.

As will be demonstrated, the chances of over- or-underestimating e by various factors, varies significantly with the magnitude l of e , both for individual e' predictions as well as for aggregated estimates \hat{e} obtained using the median of the e' . Aggregation via the median was used since it had the best or second best performance in terms of smallest average MME for both probabilistic and physical EJE data. The chapter also presents summary metrics—average, median, and maximum MME—for various aggregation methods, against binned magnitude of e , as a preliminary investigation of whether certain methods work best over certain regions of values of e , as opposed to applying one method group-wide.

5.2 Research Question 3 Significance

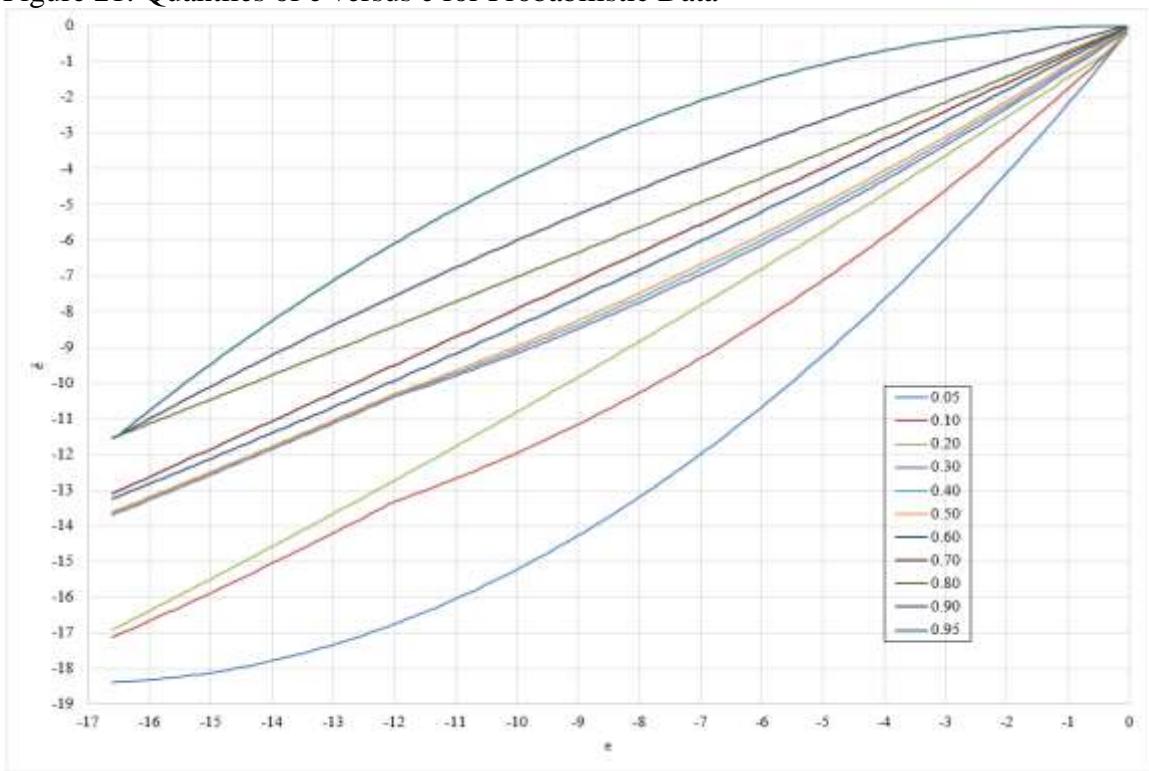
Knowing the impact of the magnitude of e on over-and under estimation chances may be useful to organizations if they have a rough idea of what magnitude of e may be applicable, prior to engaging experts. Comparing accuracy of different aggregation methods versus different levels of binned e values, may give insight into whether some methods perform better at low (e.g., relatively infrequent for probabilistic data) levels of e , while other methods perform better at higher levels of e .

5.3 Research Question 3 Methodology and Results

The quantile regression technique discussed in Chapter 4 was employed to estimate the chances that \hat{e} over- or underestimates e by given factors, depending on the magnitude of e . For this purpose, a quadratic model and quantile regression were employed, in which e was treated as the independent variable, and \hat{e} as the dependent variable. The record weights, w associated with each SeqID were applied to each (e, \hat{e}) pair in the regression. The Python function `fmin` was used to solve for the second-order polynomial coefficients minimizing $\sum \text{Tilted_Abs}(\rho, \hat{e} - \text{polyval}(\text{coef}, e)) \cdot w$ over all pairs of log-transformed (e, \hat{e}) , where $\text{polyval}(\text{coef}, x) = \text{polyval}(\text{coef}, x) = a_0 + a_1 x + a_2 x^2$, and ρ is the quantile, e.g. 0.9. The resulting quantile curves (5th, 10th, 20th, 30th, ..., 90th, and 95th) are plotted for probabilistic data in

Figure 21: Quantiles of \hat{e} versus e for Probabilistic Data.

Figure 21: Quantiles of \hat{e} versus e for Probabilistic Data

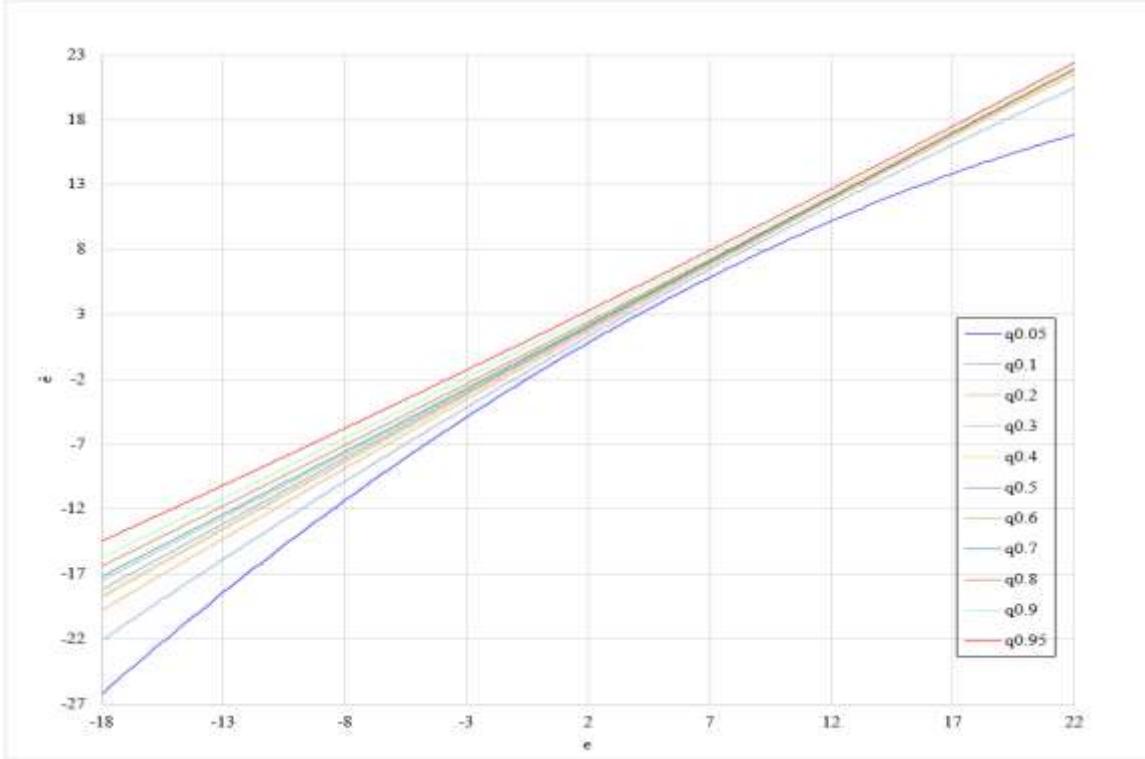


Straight line adjustments were used in the tails where necessary to prevent crossover of the quantile curves. An analogous process was applied to physical data as well to obtain a quadratic fit for \hat{e} given e .

The resulting quantile curves are shown in

Figure 22: Quantiles of \hat{e} versus e for Physical Data. As was the case with e' , so with \hat{e} : There is less curvature and smaller variation in the distance between 5th and 95th quantiles for physical data than for probabilistic data.

Figure 22: Quantiles of \hat{e} versus e for Physical Data

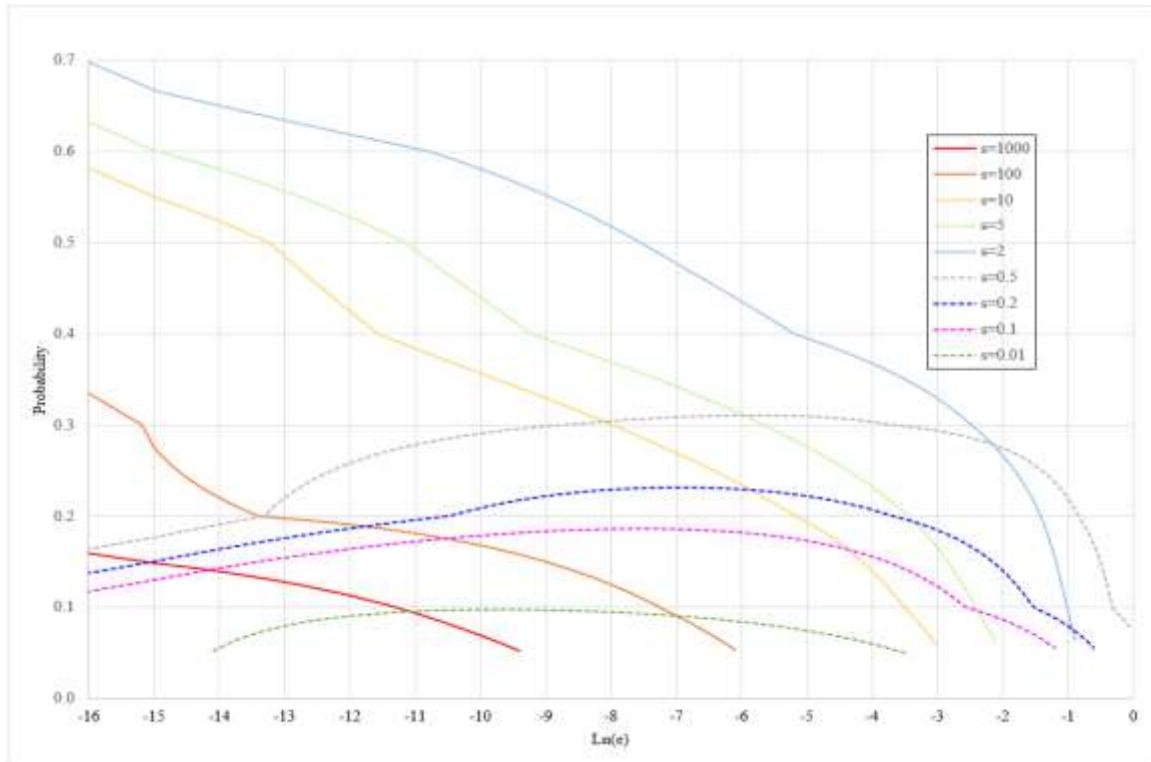


Given these quantile curves, over- or underestimation chances by various factors, both before and after aggregation via the median of the e' values can be estimated. The reason that these chances are estimated, and not exact, is that the quantiles were obtained using a model fit to the data via quantile regression. Additionally — and especially for probabilistic data — the points falling at or below the 0.05 quantile, or at or above the 0.95 quantile, are sparse. For example, when quantile regression is applied with respect to \hat{e} data, since there are only 67 probabilistic records, the 0.05 quantile represents approximately three records for which the \hat{e} value falls below this curve. For probabilistic data, the 5th, 10th, 25th, 50th, 75th and 90th empirical percentile points of e are 0.00001, 0.001, 0.01, 0.1, 0.65, and 0.75. In Ln-space, these are $-11.5, -6.9, -4.6, -2.3, -0.4$, and -0.3 , respectively. The minimum value of e is $6.50 \cdot 10^{-8}$; $\ln(e) = -16.5$. The few points falling at or below the 0.05 quantile, or at or above the 0.95 quantile, are

dispersed over this range, which exceeds sixteen units in Ln-space. Inaccuracies associated with the quadratic fit increase towards the sparse tails of the distribution of e .

Figure 23: Probability $\{e' \leq s \cdot e, s < 1; e' \geq s \cdot e, s > 1\}$ for specified values of s —Probabilistic Data shows the chances that e' over- or underestimates e by various factors.

Figure 23: Probability $\{e' \leq s \cdot e, s < 1; e' \geq s \cdot e, s > 1\}$ for specified values of s —Probabilistic Data

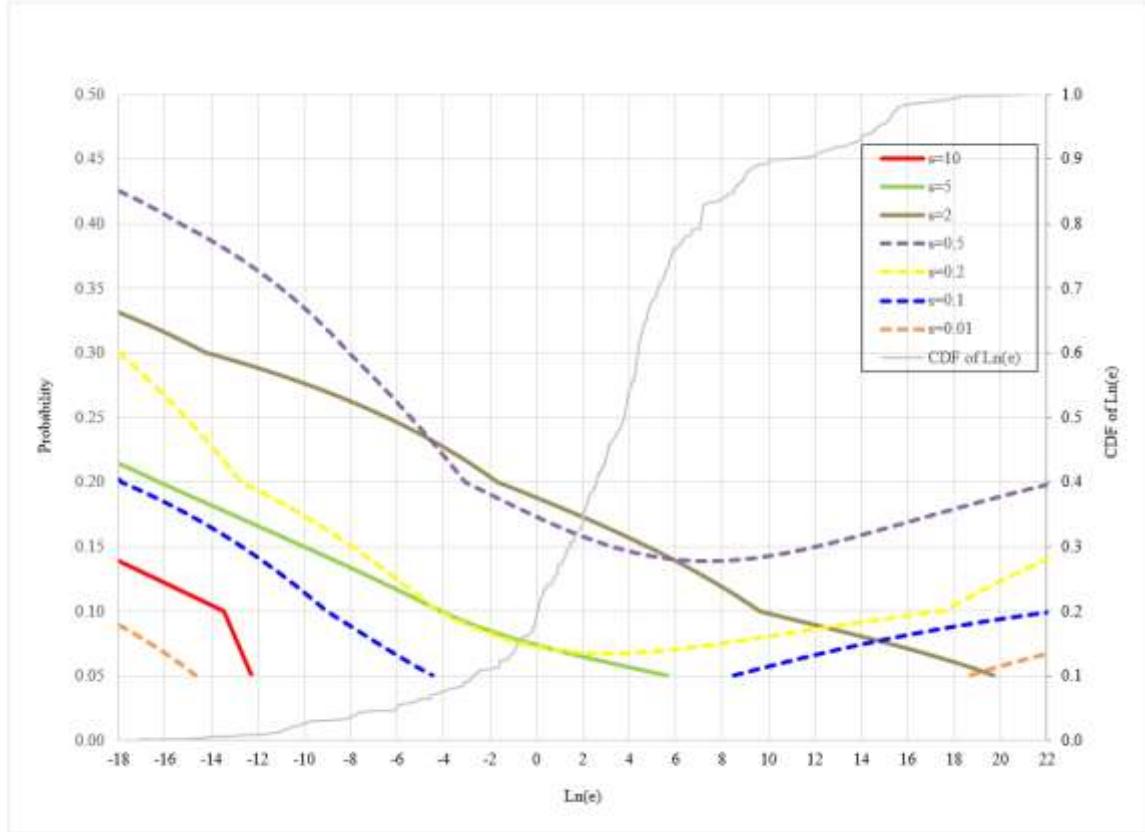


From Figure 23, it can be seen that the magnitude of e has a notable impact on multiplicative error. For example, e' is approximately twice as likely to over- rather than underestimate e by factors of 2,5, and 10 at $\text{Ln}(e) = -9$, corresponding to $e \approx 0.00012$, representing the left eight percent tail of the distribution of e . e' is approximately one and one half times as likely to over- rather than underestimate e by factors of 2,5, and 10, at $\text{Ln}(e) = -7$ ($e \approx 0.001$). As $\text{Ln}(e)$ continues to increase, the probabilities of over- and underestimating by factors of two, five and ten become approximately equal at $\text{Ln}(e) =$

-2.2 , -3.3 , and -4.3 , respectively. Since e is bounded by one, underestimation becomes more likely than overestimation as e continues to increase. For example, at $\ln(e) = -1$, underestimation by a factor of two is more than twice as likely to occur as overestimation. Approximately forty percent of the mass of e lies to the right of this point.

The analogous over- and underestimation probabilities for physical data, computed from the physical quantile curves are shown in Figure 24: Probability $\{e' \leq s \cdot e, s < 1; e' \geq s \cdot e, s > 1\}$ for specified values of s —Physical Data.

Figure 24: Probability $\{e' \leq s \cdot e, s < 1; e' \geq s \cdot e, s > 1\}$ for specified values of s —Physical Data



The CDF associated with $\ln(e)$ is shown on the right axis. For physical data, the 5th, 10th, 25th, 50th, 75th and 90th empirical percentile points of e are 0.0025, 0.07, 2.2, 44, 340 and 50,000. Corresponding $\ln(e)$ values are $-6.0, -2.7, 0.8, 3.8, 5.8, 10.8$,

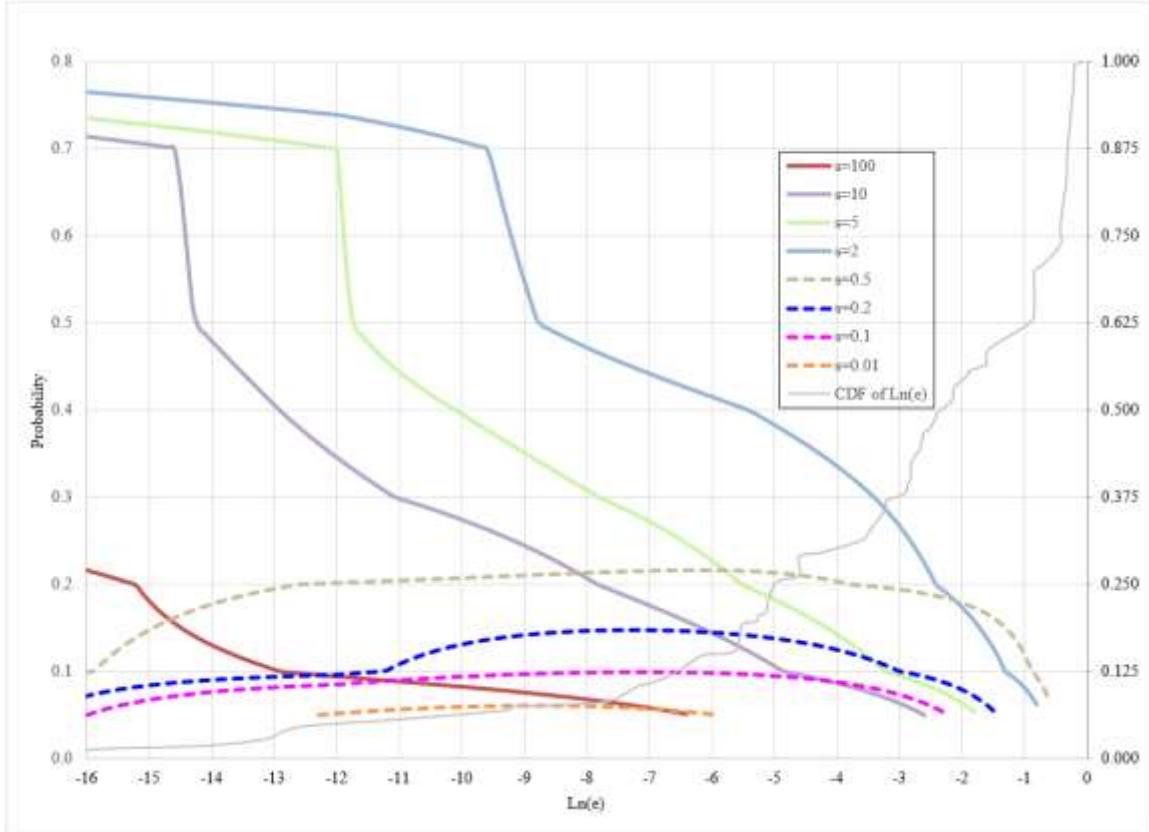
respectively. However, the full range of e values for physical data extends from ten units in $\ln(e)$ -space to either side of these limits: $3.6 \cdot 10^{-8}$, $\ln(e) = -17$; to $1.5 \cdot 10^9$, $\ln(e) = 21$.

From Figure 24 it can be seen that e' is approximately equally likely to over- as to underestimate e by factors of 2 and 5, over a broad region of $\ln(e)$ extending from approximately -4 (7% of e values lie to the left of this point) to $+5$ (one third of e values fall to the right of this point). As $\ln(e)$ decreases to the left of -4 , underestimation becomes more likely than overestimation. This trend is the opposite of the one observed for probabilistic data. Underestimation probabilities also increase as $\ln(e)$ increases beyond 5 ; underestimation by a factor of two is approximately twice as likely to occur as overestimation at $\ln(e) = 14$.

As previously stated, the chances that \hat{e} under- or overestimates e by given factors, as a function of the magnitude of e , were computed using the quantile curves for \hat{e} given e , obtained via quantile regression. For probabilistic data, these chances are shown in

Figure 25: Probability $\{\hat{e} \leq s \cdot e, s < 1; \hat{e} \geq s \cdot e, s > 1\}$ for specified values of s —
Probabilistic Data.

Figure 25: Probability $\{\hat{e} \leq s \cdot e, s < 1; \hat{e} \geq s \cdot e, s > 1\}$ for specified values of s —Probabilistic Data



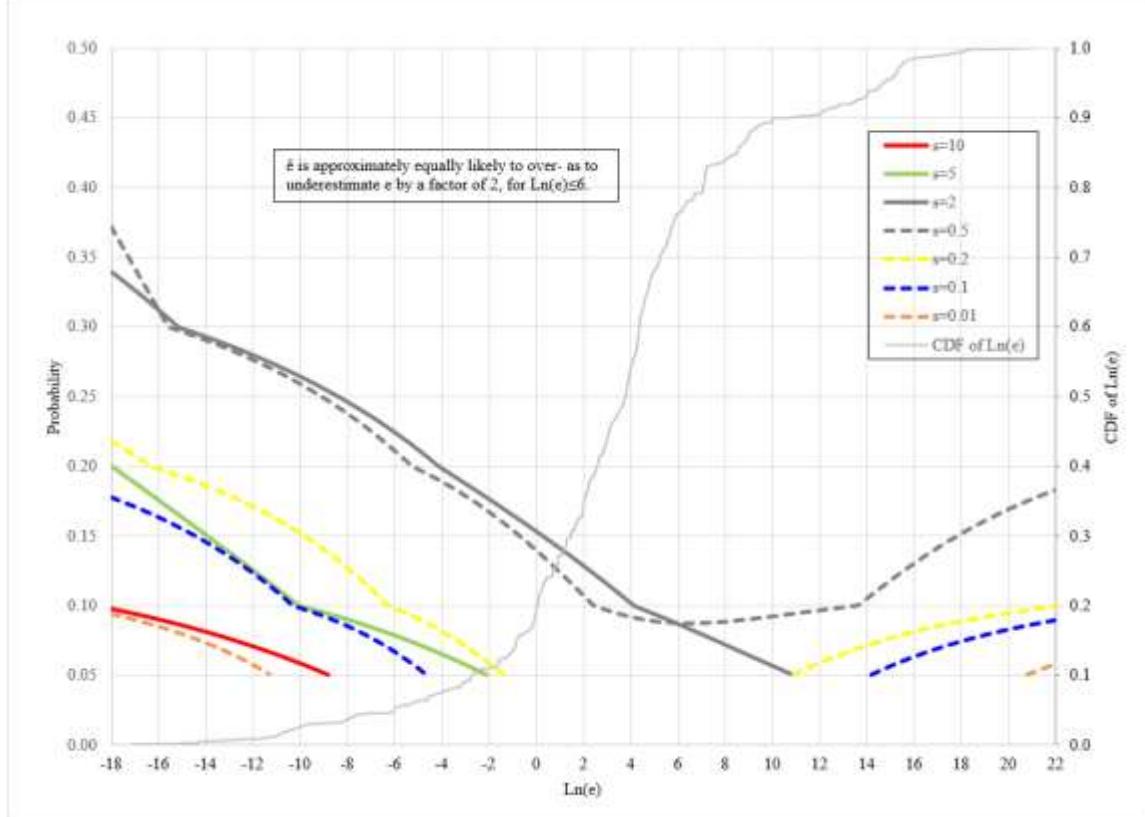
The CDF associated with $\ln(e)$ is shown on the right axis. The relatively flat portions of the overestimation probability curves for given factors, over the left tail of e , arise from the wide gaps between the 0.2 and 0.3 quantile curves for \hat{e} given e , observable in Figure 21. As e continues to increase, the probability of overestimation drops below 70%. At this point, the much narrower gaps between the 0.3 and 0.4 quantiles, or between the 0.4 and 0.5 quantiles in Figure 21, become applicable, and the corresponding probability of occurrence decreases precipitously. Once it drops below fifty percent, the wider gaps between quantiles above 0.5 in Figure 21 become applicable; as a result, the probability of occurrence decreases at a roughly constant rate in $\ln(e)$, until it drops below five percent and is no longer shown on the graph.

Aggregation considerably reduces the likelihood of overestimation error: instead of 15 e' points having MMEs exceeding 1000 (and in some cases, 10^5), ê for only one record, having 0.005 associated probability of occurrence, has an MME exceeding 1000 (the actual MME is 1400). At $\ln(e) = -6$ (corresponding to $e \approx 1/400$, $CDF \approx 0.15$), aggregation reduces the chances of overestimating by factors of two, five, or ten from 44, 31 and 23 percent, to 42, 23 and 14, percent, respectively. The corresponding chances of underestimation at this point are reduced by roughly one third, from 31, 23 and 18 percent, to 22, 14 and 10 percent. As before, underestimation becomes more likely than overestimation, as e approaches one.

For physical data, the analogous chances are shown in

Figure 26: Probability $\{\hat{e} \leq s \cdot e, s < 1; \hat{e} \geq s \cdot e, s > 1\}$ for specified values of s—**Physical Data.**

Figure 26: Probability $\{\hat{e} \leq s \cdot e, s < 1; \hat{e} \geq s \cdot e, s > 1\}$ for specified values of s —Physical Data



As before, the right axis shows the CDF of $\ln(e)$. If the curve associated with $s=10$ in this figure is compared to its analogue in Figure 24, it seems to show a higher probability of overestimating by a factor of ten at $\ln(e) = -12$ using an aggregated \hat{e} , than using an individual prediction e' . To check this anomalous result, quantile curves for both \hat{e} and e' were generated by applying quantile regression over the left three percent tail of e ($\ln(e) < -9$) only. The results revealed a greater spread between quantiles for e' than for \hat{e} . For example, at $\ln(e) = -16$, the 5th and 95th quantiles, respectively were -18 and -12 , for e' ; while they were -15 and -13 for \hat{e} . At $\ln(e) = -9$, the quantiles for e'

were -11 and -8 ; while they were -11.5 and -8.5 , for \hat{e} , implying a smaller probability that \hat{e} would overshoot e by a large factor. For this reason, the curves in Figure 26 associated with $s=10$ and $s=0.01$, corresponding to the left and right tails of e , are considered spurious artifacts, and are not compared to their counterparts in Figure 24. The same general trends of influence of e on over- and underestimation error observed in Figure 24 are observed in Figure 26 for aggregated \hat{e} : increasing e reduces the chance of overestimation approximately linearly in $\ln(e)$ until it falls below five percent and drops off the graph. The chances of factor-of-two over- or underestimation are approximately equal, declining to approximately nine percent at $\ln(e)=6$ ($CDF \approx 0.75$); thereafter, underestimation becomes more likely. By comparison with Figure 24, at this same point ($\ln(e)=6$), the chances of factor-of-two over- or underestimation were approximately 14% for individual e' , which represents a fifty percent greater likelihood.

This chapter also reports on a preliminary investigation of whether certain methods work best in terms of having the greatest accuracy, over certain regions of values of e , as opposed to applying one method group-wide. The specific behavior of the descriptive statistics over the subgroups may provide information which could otherwise be lost in a group-wide estimate. Table 22 through

Table 27 provide the information used in this analysis.

Table 22: EJE Probabilistic Data – Spearman ρ by Aggregation Method for Bin Mid-Point shows the overall rank correlation of accuracy with bin midpoint, for each method and for each summary descriptive measure of accuracy—average, median, and maximum MME—for binned EJE probabilistic data provided in Table 24: Stratification EJE Probabilistic Data. Table 23: EJE Physical Data- Spearman ρ by Aggregation

Method for Bin Mid-Point provides the analogous correlations based on the same set of statistics for binned physical data shown in Table 25: Stratification EJE Physical Data Bins [17.2, -11) – [1,2) and Table 26: Stratification EJE Physical Data Bins [4,4.3333) – [15,21.14). All eight methods of aggregation were used in all tables.

Each table was constructed as follows: Ln(e) values for realized values, e were assigned to bins. It was desired to have a minimum of five points in each bin, against which summary statistics could be computed. It was also desired to avoid using very large (greater than five in Ln-space) bin widths, to avoid large mass concentrations or numbers of points in individual bins, and to use constant widths for bins. It proved impossible to satisfy these considerations simultaneously.

For probabilistic data, six bins were used. These ranged in width from approximately 4.5 for the left two bins, with 7 and 6 points, respectively, and masses of approximately four percent each; to approximately 2.25 for the next three bins, with 10, 12, and 22 points, respectively, and masses ranging from 12 to 33 percent; to 0.75 for the last bin, with 10 points and 30 percent mass.

For physical data, 23 bins were used. Seven bins between -1 and 2, and between 5 and 9, had width one; bins beyond nine had width two; bins less than -1 had width four, except that the leftmost and rightmost bins had widths of approximately six, and included approximately one percent of the data, in 32 points at the left tail, and 4.5 percent of the data in 63 points at the right tail. It was necessary to split several bins between 2 and 5 into finer increments, to avoid excessive numbers of points. For example, one of these bins runs from 3.3333 to 3.6666, yet includes 249 points representing only two percent of the mass. These points were largely associated with a single theme involving Crop

Yields. Except for the leftmost bins, masses ranged between approximately two and eight percent.

For each bin with realized values falling in that bin, the descriptive statistics, specifically, average, median, and maximum were computed for the corresponding MMEs for each aggregation method. Table 24 through Table 26 also show the number of realized values in each bin and the associated mass for the bin. The analysis was developed as follows. The Spearman Correlation Coefficient (ρ) was computed for each aggregation method's statistic for a bin versus the bin's mid-point to ascertain an approximate trend between the bin limits series (increasing) and the representative MME values. The ρ values are provided in Table 22 and Table 23 for Table 24, and Table 25 and Table 26 respectively. However; the physical data has 23 bins as compared to six for probabilistic data, thereby allowing for firmer inferences. Table 24 through Table 26 were examined to determine whether any particular aggregation method prevailed or proved anomalous for sets of bins for any particular statistic. The MMEs were ranked by aggregation method within a bin for a specific statistic (see

Table 27: Ranks of EJE Physical Data Bins **Stratification**) to support detection of such a pattern.

Table 22: EJE Probabilistic Data – Spearman ρ by Aggregation Method for Bin Mid-Point

Statistic	Aggregation Method							
	Arithmetic	Geometric	Harmonic	Median	Alpha-Stable	Bayesian	Rule of Thumb	Maximum Likelihood
Average	-1.0	-0.3	0.5	-0.4	-0.4	-0.8	-0.4	-0.2
Median	-1.0	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8
Maximum	-1.0	-0.3	0.7	-0.4	-0.4	-0.5	-0.4	-0.2

Table 22 shows that with the exception of the Harmonic Mean for the average and maximum statistics, all statistics decreased with increasing bin limits. Inferencing is limited given that only six bins are available. However, the overall trend is tentatively aligned with the related Literature Review in Chapter 2, i.e., prediction accuracy decreases with decreasing numbers.

Table 23: EJE Physical Data- Spearman ρ by Aggregation Method for Bin Mid-Point

Statistic	Aggregation Method							
	Arithmetic	Geometric	Harmonic	Median	Alpha-Stable	Bayesian	Rule of Thumb	Maximum Likelihood
Average	-0.01	-0.49	-0.21	-0.53	-0.52	-0.45	-0.52	-0.45
Median	-0.86	-0.87	-0.87	-0.86	-0.86	-0.83	-0.87	-0.89
Maximum	-0.01	-0.32	-0.16	-0.32	-0.29	-0.21	-0.31	-0.35

Table 23 shows that all statistics decreased with increasing bin limits for all aggregation methods and statistics. The number of bins is 23. The highlighted ρ values are significant with $p<0.05$.

Table 24 shows that with respect to the average MME statistic for probabilistic data, the Arithmetic Mean had the lowest MMES for two consecutive bins and the Bayesian had the second lowest MMEs for these bins. The Alpha-Stable had the lowest spread of ranks (3, 4, 5) within this statistic. With respect to the median statistic, the Median method of aggregation demonstrated the lowest spread of ranks (1, 2, 3), and was approximately tied with the Alpha-Stable and ROT for the smallest difference (roughly 7.3) between the maximum and minimum MMEs across the bins. Regarding the maximum statistic, the Arithmetic Mean had the smallest MMEs for the three largest bin limits (3, -0.75, 0); the Harmonic Mean had the largest MMEs for these bins. The

Arithmetic Mean also had the largest maximum as well as average value for the smallest bin limit, by an order of magnitude.

For purposes of this analysis, Table 25 and Table 26 for physical data statistics are combined. The smallest MMEs for seven out of the eight aggregation methods are assigned to the [7,8) bin for the average and median statistics; the exceptions are the Arithmetic Mean for the average statistic and the MLE for the median statistic. In these cases, the representative MME was the second smallest by rank. Within the median statistic, seven out of the eight aggregation methods had their second smallest values in the [15,21.14) bin; the MLE had its smallest value there. Similarly, all third smallest MMEs are in the [13,15) bin.

As previously stated, Table 23 shows that per the Spearman coefficient, all the MMEs produced by each method of aggregation of physical data are negatively correlated with increasing bin limits. In the case of the median statistic, all Spearman coefficients are significant with $p < 0.05$, and this determination is applicable to the average statistic with the exception of the Harmonic and Arithmetic mean.

Table 27 highlights series of identical ranks within a given method of aggregation across three or more bins for the physical data, in order to gain visual perspective on possible patterns. Several of these streaks run for six bins. This appears to be statistically significant, since an upper bound on the probability that, by random chance, a particular method out of eight methods performs best for six consecutive bins is $(1/8)^6$ multiplied by eight possible methods, multiplied by 23 bins, or about 0.005. For

example, the average and maximum statistics for the Bayesian aggregation ranked 1 for the range of bins [2.5,3) to [4.3333 to 4.6666), i.e., six bins.

5.4 Summary

In summary, there is a general tendency for the representative MMEs for any given method of aggregation to decrease with increasing bin limits, consistent with the general trend of decreasing accuracy with decreasing numbers. The median statistic appears to be less subject to variation across the bins compared with the average and maximum statistics. No one method dominates the probabilistic bins, with respect to average MME. As discussed, for physical bins, with the exception of a streak of six bins in which the Bayesian had best performance, there is no consistent pattern to best performance of any particular aggregation method over bins.

Table 24: Stratification EJE Probabilistic Data

Data type: Probabilistic	Upper bin limit (Ln): 12)	-16.55,- 12)	-7.50	-5.24	-3.00	-0.75	0.00
	Number of obs:	7	6	10	12	22	10
	Mass:	0.046	0.037	0.122	0.166	0.328	0.301
Statistic	Aggregation method						
Average	Arithmetic	4869	134.9	11.7	9.5	2.1	1.3
Average	Geometric	190.3	7.0	10.7	273.5	13.4	1.6
Average	Harmonic	76.4	283.8	95.2	94415	25023	251.6
Average	Median	38.7	2.9	20.2	133.5	5.5	1.3
Average	Alpha-Stable	118.7	4.5	15.5	279.3	6.2	1.4
Average	Bayesian	308.3	5.4	8.8	48.8	4.2	1.5
Average	Rule of Thumb	113.1	4.4	15.8	326.4	6.3	1.4
Average	Maximum Likelihood	5.1	76.2	672	1357	12.5	1.2

Data type: Probabilistic	Upper bin limit (Ln):	[-16.55,-12)	-7.50	-5.24	-3.00	-0.75	0.00
	Number of obs:	7	6	10	12	22	10
	Mass:	0.046	0.037	0.122	0.166	0.328	0.301
Statistic	Aggregation method						
Median	Arithmetic Mean	37.3	12.9	4.3	2.8	1.6	1.4
Median	Geometric Mean	9.2	2.4	4.1	3.6	2.2	1.4
Median	Harmonic Mean	39.9	3.7	14.1	35.2	3.1	1.4
Median	Median	8.5	2.4	2.6	3.5	1.6	1.3
Median	Alpha-Stable	8.8	2.2	4.4	5.0	2.0	1.4
Median	Bayesian	21.4	2.8	3.6	4.0	1.6	1.4
Median	Rule of Thumb	8.7	2.2	4.4	5.0	2.0	1.4
Median	Maximum Likelihood	4.6	6.1	13.6	4.1	1.7	1.1
Maximum	Arithmetic Mean	33915	751	48.7	47.6	10.2	1.8
Maximum	Geometric Mean	1166	19.9	56.7	2742	114	3.2
Maximum	Harmonic Mean	263	1669	429	1127510	375008	2502
Maximum	Median	154	6.2	154	1400	54.0	1.8
Maximum	Alpha-Stable	703	9.9	97.7	2360	53.1	1.9
Maximum	Bayesian	2016	20.3	25.8	367.2	33.2	2.1
Maximum	Rule of Thumb	664	10.0	99.9	2887	53.5	1.9
Maximum	Maximum Likelihood	8.9	412	5690	13678	200	2.2

Table 25: Stratification EJE Physical Data Bins [17.2, -11) – [1,2)

Data type: Physical	Upper bin limit (Ln):	[-17.14,-11)	[-11,-7)	[-7,-3)	[-3,-1)	[-1,0)	[0,1)	[1,2)	[2,2.5)	[2.5,3)	[3,3.3333)	[3.3333,3.6666)	[3.6666,4)
	Number of obs:	32	38	31	42	51	64	43	43	88	105	249	216
	Mass:	0.013	0.032	0.044	0.045	0.057	0.081	0.067	0.054	0.047	0.027	0.021	0.052
Statistic	Aggregation method												
Average	Arithmetic Mean	5.4	6.1	231.89	3.04	3.42	6.63	1.58	51.25	69.85	28	71.14	86.89
Average	Geometric Mean	2.56	38.5	232.1	5.61	3.59	3.55	1.43	2.57	69.61	27.84	71	86.34
Average	Harmonic Mean	2.59	471.33	232.71	54.29	28.86	2.93	1.38	4.61	70.05	28.91	71.11	86.38
Average	Median	3.38	14.33	232.54	4.36	3.81	5.59	1.53	2.87	69.65	27.93	70.98	86.35
Average	Alpha stable	2.65	34.08	232.62	4.84	3.28	4.32	1.45	2.75	69.55	27.84	70.93	86.26
Average	Bayesian	4.06	9.42	211.63	4.49	3.58	5.47	1.56	2.92	49.09	20.05	49.95	60.68
Average	Rule Of Thumb	2.62	37.28	232.42	5	3.32	4.19	1.44	2.73	69.61	27.86	71	86.34
Average	MLE	3.99	12.87	241.86	6.68	4.84	6.48	1.72	3.23	56.95	23.18	57.97	70.42
Median	Arithmetic Mean	2.09	2.17	1.9	2.36	1.38	1.31	1.29	1.45	1.23	1.58	1.41	1.32
Median	Geometric Mean	1.81	1.92	1.78	2.63	1.39	1.32	1.25	1.34	1.25	1.56	1.41	1.32
Median	Harmonic Mean	1.72	2.04	2	2.82	1.46	1.32	1.23	1.34	1.24	1.58	1.41	1.31
Median	Median	1.79	2	1.62	2.49	1.45	1.3	1.25	1.34	1.26	1.58	1.41	1.32
Median	Alpha stable	1.83	2	1.7	2.63	1.4	1.32	1.25	1.35	1.25	1.57	1.41	1.3
Median	Bayesian	2.04	1.94	1.73	2.6	1.51	1.28	1.3	1.4	1.25	1.5	1.35	1.3
Median	Rule Of Thumb	1.81	2.02	1.68	2.68	1.37	1.32	1.24	1.34	1.25	1.58	1.41	1.29
Median	MLE	1.88	2.33	1.65	2.75	1.61	1.24	1.28	1.41	1.27	1.63	1.43	1.32
Maximum	Arithmetic Mean	101.5	79.78	7,080.00	6.96	38.75	143.1	7.83	1,867.10	1,800.00	2,700.00	3,600.00	4,900.00
Maximum	Geometric Mean	9.42	801.58	7,080.00	43.68	44.12	82.79	3.83	16.24	1,800.00	2,700.00	3,600.00	4,900.00
Maximum	Harmonic Mean	10.39	10,656.41	7,080.00	667.4	1,246.77	53.24	2.5	95.44	1,800.00	2,700.00	3,600.00	4,900.00

Data type: Physical	Upper bin limit (Ln):	[-17.14,-11)	[-11,-7)	[-7,-3)	[-3,-1)	[-1,0)	[0,1)	[1,2)	[2,2.5)	[2.5,3)	[3,3.3333)	[3.3333,3.6666)	[3.6666,4)
	Number of obs:	32	38	31	42	51	64	43	43	88	105	249	216
	Mass:	0.013	0.032	0.044	0.045	0.057	0.081	0.067	0.054	0.047	0.027	0.021	0.052
Statistic	Aggregation method												
Maximum	Median	30.14	256.97	7,080.00	28.57	38.75	126.9	7.83	24	1,800.00	2,700.00	3,600.00	4,900.00
Maximum	Alpha stable	10.66	699.88	7,086.70	30.33	27.52	92.35	4.11	20.99	1,798.30	2,697.45	3,596.59	4,895.36
Maximum	Bayesian	53.13	157.98	6435.41	34.62	33	126.3	6.46	25.12	1255.85	1883.78	2511.71	3418.72
Maximum	Rule Of Thumb	9.66	769.8	7,080.00	31.93	28.8	89.97	3.83	20.57	1,800.00	2,700.00	3,600.00	4,900.00
Maximum	MLE	40.64	177.83	7,330.58	95.62	74.81	153.5	14.62	27.56	1,461.14	2,191.71	2,922.28	3,977.55

Table 26: Stratification EJE Physical Data Bins [4,4.3333) – [15,21.14)

Data type: Physical	Upper bin limit (Ln):	[4,4.3333)	[4.3333,4.6666)	[4.6666,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,11)	[11,13)	[13,15)	[15,21.14)
	Number of obs:	154	111	33	124	39	58	33	26	22	56	63
	Mass:	0.052	0.055	0.034	0.081	0.029	0.047	0.036	0.026	0.018	0.036	0.045
Statistic	Aggregation method											
Average	Arithmetic Mean	2.07	27.57	1.39	1.28	89.83	1.28	38.9	164.48	42.19	16.12	1.86
Average	Geometric Mean	1.68	31.14	1.52	1.6	1.93	1.12	1.54	2.32	1.59	1.46	4.16
Average	Harmonic Mean	2.58	40.55	8.61	9.46	2.85	1.2	8.45	1.36	2.89	10.07	1,522.19
Average	Median	1.69	28.24	1.34	1.53	2.44	1.12	1.64	1.98	1.31	1.93	3.35
Average	Alpha stable	1.69	28.85	1.41	1.54	1.91	1.12	1.56	2.19	1.44	1.49	3.74
Average	Bayesian	1.62	24.68	1.43	1.54	2.42	1.14	2.61	41.08	1.5	2.82	2.06
Average	Rule Of Thumb	1.69	29.68	1.41	1.56	1.92	1.12	1.59	2.1	1.43	1.47	3.95
Average	MLE	1.71	26.6	1.3	1.61	2.28	1.12	2.54	47.78	1.48	2.83	2.01
Median	Arithmetic Mean	1.39	1.26	1.26	1.07	1.28	1.03	1.09	1.12	1.11	1.04	1.04
Median	Geometric Mean	1.38	1.25	1.26	1.07	1.21	1.03	1.09	1.16	1.11	1.04	1.04
Median	Harmonic Mean	1.4	1.26	1.24	1.06	1.2	1.03	1.09	1.15	1.1	1.04	1.04
Median	Median	1.39	1.26	1.24	1.07	1.26	1.03	1.09	1.12	1.11	1.04	1.04
Median	Alpha stable	1.38	1.26	1.24	1.07	1.27	1.03	1.09	1.15	1.11	1.04	1.03
Median	Bayesian	1.32	1.21	1.38	1.15	1.34	1.03	1.16	1.22	1.14	1.09	1.06
Median	Rule Of Thumb	1.38	1.25	1.24	1.06	1.24	1.03	1.09	1.15	1.11	1.04	1.04
Median	MLE	1.41	1.28	1.2	1.08	1.19	1.04	1.09	1.09	1.11	1.05	1.02
Maximum	Arithmetic Mean	55.2	2,622.95	3.05	5.02	1,700.79	10.76	1,243.73	4,245.16	903.54	663.67	23.83
Maximum	Geometric Mean	11.29	3,299.49	5.28	33.76	13.9	2.7	6.55	30.6	6.99	14.35	107.72
Maximum	Harmonic Mean	104.51	4,333.33	197.11	588.9	26.47	3.73	186.75	3.26	39.87	374.67	61,982.11

Data type: Physical	Upper bin limit (Ln):	[4,4.3333)	[4.3333,4.6666)	[4.6666,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,11)	[11,13)	[13,15)	[15,21.14)
	Number of obs:	154	111	33	124	39	58	33	26	22	56	63
	Mass:	0.052	0.055	0.034	0.081	0.029	0.047	0.036	0.026	0.018	0.036	0.045
Statistic	Aggregation method											
Maximum	Median	11.29	2,962.96	3	35.96	33.33	2.56	9	22.67	4.9	43.65	90.91
Maximum	Alpha stable	11.28	3,041.38	3.6	33.39	12.12	2.53	6.86	28.04	5.7	18.22	101.65
Maximum	Bayesian	9.79	2568.63	2.9	33.28	35.53	2.48	36.31	1036.74	8.27	87.59	36.08
Maximum	Rule Of Thumb	11.29	3,133.39	3.73	35.1	11.24	2.62	7.2	25.54	5.31	16.79	110.52
Maximum	MLE	10.55	2,775.06	2.35	43.98	35.28	2.25	34.87	1,213.88	8.63	81.66	35.01

Table 27: Ranks of EJE Physical Data Bins Stratification

Statistic	Aggregation	LB:	-	17.14	-11	-7	-3	-1	0	1	2	2.5	3	3.3333	3.6666	4	4.3333	4.667	5	6	7	8	9	11	13	15
		UB:	-11	-7	3	1	0	1	2	2.5	3	3.3333	3.6666	4	4.3333	4.6666	5	6	7	8	9	11	13	15	21.1	
Average	Arithmetic Mean		8	1	2	1	3	8	7	8	7	7	8	8	7	3	3	1	8	8	8	8	8	8	8	1
Average	Geometric Mean		1	7	3	6	5	2	2	1	4	3	5	4	2	7	7	6	3	1	1	5	6	1	7	
Average	Harmonic Mean		2	8	7	8	8	1	1	7	8	8	7	7	8	8	8	8	7	7	7	1	7	7	8	
Average	Median		5	4	5	2	6	6	5	4	6	6	4	6	3	4	2	2	6	1	4	2	1	4	4	
Average	Alpha stable		4	5	6	4	1	4	4	3	3	3	3	3	5	4	3	1	1	2	4	3	3	5		
Average	Bayesian		7	2	1	3	4	5	6	5	1	1	1	1	1	6	3	5	6	6	6	5	5	3		
Average	Rule Of Thumb		3	6	4	5	2	3	3	2	4	5	5	4	3	6	4	5	2	1	3	3	2	2	6	
Average	MLE		6	3	8	7	7	7	8	6	2	2	2	2	6	2	1	7	4	1	5	7	4	6	2	
Median	Arithmetic Mean		8	7	7	1	2	4	7	8	1	4	2	5	5	4	6	3	7	1	1	2	2	1	3	
Median	Geometric Mean		3	1	6	4	3	5	3	1	3	2	2	5	2	2	6	3	3	1	1	7	2	1	3	
Median	Harmonic Mean		1	6	8	8	6	5	1	1	2	4	2	4	7	4	2	1	2	1	1	4	1	1	3	
Median	Median		2	3	1	2	5	3	3	1	7	4	2	5	5	4	2	3	5	1	1	2	2	1	3	
Median	Alpha stable		5	3	4	4	4	5	3	5	3	3	2	2	2	4	2	3	6	1	1	4	2	1	2	
Median	Bayesian		7	2	5	3	7	2	8	6	3	1	1	2	1	1	8	8	8	1	8	8	8	8	8	
Median	Rule Of Thumb		3	5	3	6	1	5	2	1	3	4	2	1	2	2	2	1	4	1	1	4	2	1	3	
Median	MLE		6	8	2	7	8	1	6	7	8	8	8	8	5	8	8	1	7	1	8	1	1	2	7	1
Maximum	Arithmetic Mean		8	1	2	1	4	7	6	8	4	4	4	4	7	2	4	1	8	8	8	8	8	8	1	
Maximum	Geometric Mean		1	7	2	6	6	2	2	1	4	4	4	4	4	7	7	4	3	6	1	5	4	1	6	
Maximum	Harmonic Mean		3	8	2	8	8	1	1	7	4	4	4	4	8	8	8	8	4	7	7	1	7	7	8	
Maximum	Median		5	4	2	2	4	6	6	4	4	4	4	4	4	4	3	6	5	4	4	2	1	4	4	
Maximum	Alpha stable		4	5	7	3	1	4	4	3	3	3	3	3	3	5	5	3	2	3	2	4	3	3	5	
Maximum	Bayesian		7	2	1	5	3	5	5	5	1	1	1	1	1	1	2	2	7	2	6	6	5	6	3	
Maximum	Rule Of Thumb		2	6	2	4	2	3	2	2	4	4	4	4	4	6	6	5	1	5	3	3	2	2	7	
Maximum	MLE		6	3	8	7	7	8	8	6	2	2	2	2	2	3	1	7	6	1	5	7	6	5	2	

Chapter 6: Research Question 4 (Number of Experts)

6.1 Introduction

According to Winkler and Clemen (2004) “The motivation behind using multiple experts and/or multiple methods is simply to get additional information that can lead to more accurate forecasts or estimates and, ultimately, to better decisions” (p. 167). How does the quality of the aggregated estimate vary with the number of experts used? Does a larger number of experts imply a tighter credible interval, for the same level of probability?

6.2 Research Question 4 Significance

There are no frameworks or best practices that indicate how many experts are required for decision support in any particular domain; for example, despite the wide use of the Delphi technique, Rowe and Wright (2001) noted that “no firm rule governs the number of panelists” (p. 128). The authors however did recommend that the experts should be chosen such that their “combined knowledge and expertise reflects the full scope of the problem domain” (Rowe & Wright, 2001, p.128). Generally, according to Rowe and Wright (2001), multiple experts are preferred to a single expert since the reliability of an aggregated opinion is generally more reliable than a single opinion, and the error or bias in the individual opinions may be lowered. These considerations coupled with resource factors such as budget and time, render the response question more of a judgment call than a scientific determination.

6.3 Research Question 4 Literature Review

The U.S. Environmental Protection Agency (2011) task force on expert elicitation observed that time and costs associated with expert elicitation can be significant and that proper performance “can be relatively resource-intensive and time-consuming” (p. 56). Meyer and Booker (2001) claimed that paying experts for their time, generally, should be a “last resort” (p. 89) since it may bias their results. Sniezek, Schrah, and Dalal (2004) noted that “committing money for expert—but not novice—advice increases Judges’ use of advice and their subsequent estimation accuracy” (p. 173).

Costs include not only remuneration but organizational expenses such as travel, lodging, and meeting facilities. Devilee and Knol (2011) concluded that although software exists that support aspects of expert elicitation such as consensus building and characterization of uncertainties, thereby potentially reducing costs, software that “lowers the costs of expert elicitations in terms of travel, organizing and meeting time and consequently money” (p. 10) is not available.

Although such limitations must be considered in developing an expert elicitation study, the EFSA (2014) noted that “there may be diminishing returns on the number of experts used in an elicitation” (p. 159). For example, Aspinall (2010) observed that “My experience with more than 20 panels suggests that 8–15 experts is a viable number — getting more together will not change findings significantly, but will incur extra expense and time. However, this has not been rigorously tested” (p. 295). A structured elicitation survey regarding food-specific attribution for nine illnesses conducted in Canada used responses from between 10 and 35 experts. Although the authors of that survey recognized that the number of experts was large compared with similar studies, the extent

“of the disagreement with experts clustered in two distinct subgroups for certain pathogens was not expected”, (Davidson, Ravel, Nguyen, Fazil, & Ruzante, 2011, p. 990).

Clemen and Winkler (1985) noted that although information may be obtained from a number of different experts, “positive dependence among information sources can have a serious detrimental effect on the precision and value of the information” (p. 427). Cooke and Probst (2006) reported that based on panelists interviewed during an expert judgment policy symposium and workshop, the number of experts for most studies they conducted was “targeted to lie between 6 and 12”, (p. 16). The number of participants in these studies logically impacts costs. However, per capita information is not widely available.

In a presentation to the Nuclear Technical Waste Review Board of the US Department of Energy, Office of Civilian Radioactive Waste Management regarding the Yucca Mountain Site Characterization Project, Bjerstedt (1996) stated that “Formalized use of expert judgment is expensive and time-consuming” (Slide 30). Table 28: Costs of Expert Judgment Studies (Yucca Mountain Site Characterization Project) lists the study costs cited in the presentation.

Table 28: Costs of Expert Judgment Studies (Yucca Mountain Site Characterization Project)

Study Title	Year	Duration	Cost
Exploratory Studies Facility Alternatives Study	1990	13 months	\$25,000,000
Calico Hills Risk/Benefit Analysis	1990	13 months	\$5,000,000
Test Prioritization Task	1991	10 months	\$3,500,000
Early Site Suitability Evaluation (including peer review)	1992	10 months	\$3,500,000
Probabilistic Volcanic Hazard Assessment	1994	14 months	\$1,400,000
Probabilistic Seismic Hazard Assessment (Projected)		15 months	\$4,000,000

Note: Adapted from “Principles and Guidelines for Formal Use of Expert Judgment by Yucca Mountain Site Characterization Project”, by T.W. Bjerstedt (1996). Presentation retrieved on November 26, 2014 from <http://www.nwtrb.gov/meetings/1996/jan/bjerstedt.pdf>.

Cooke and Probst (2006) reported that expert judgment panelists engaged in supporting US government regulation “cost \$100,000–300,000 or more; studies in Europe tend to cost between one and three “person” months, or \$30,000–100,000, excluding experts’ time” (pp. 14-15). The U.S. Environmental Protection Agency (n.d.) funded an expert elicitation between October 2011 and December 2013 regarding uncertainty surrounding the market and non-market damages of climate change that cost \$63,553. According to Cooke and Kelly (2010) although the exact cost estimates for the U.S. Nuclear Regulatory Commission–European Union project estimating uncertainty of accident consequence codes for nuclear power plant between 1990 and 2000 “have not been retrieved, a ballpark estimate for the entire study, including expert remuneration (\$15,000 per expert) is US\$7 million (2010)” (p. 1).

Knol, Slottje, van der Sluijs, and Lebret (2010) discussed two expert elicitation studies regarding ultrafine particles (Knol et al. (2009); Hoek et al. (2010)), whose combined costs were 20,000 euros but did not include “preparation, analysis and

reporting” (p. 6). Twelve experts participated in these studies. To minimize costs, only experts based in Europe were invited. Knol et al. (2010) noted that a study for campylobacter transmission during broiler-chicken processing (see Van der Fels-Klerx, Cooke, Maarten, Goossens, and Havelaar (2005)) employed 12 experts and cost approximately 50,000 euros.

Two United States Department of Agriculture (USDA) expert elicitation studies were performed in 2007 and 2012 (see USDA (2007, draft); USDA (2012)). These studies did not use seeds to calibrate experts; they employed 17 and 10 experts respectively and offered honorariums of \$250 and \$750 respectively to each expert. Roman, Hammitt, Walsh, and Stieb (2012) elicited three experts regarding the benefits of air pollution regulations in reducing premature mortality; compensation was based on “a uniform competitive academic consulting rate” (p. 2137). Walker, Evans, and Mackintosh (2001) noted that the seven experts who participated in elicitation regarding characterization of personal exposure to benzene, were reimbursed for expenses and received “a modest stipend for their involvement” (p. 311); a similar practice was noted in EFSA (n.d.). Sweidan et al. (2010) reported that the 12 experts participating in an e-prescribing elicitation study were compensated for “their time and travel expenses according to the organisation's remuneration policy” (p. 3).

Examples of expert elicitation studies that identify the number of experts but do not cite costs include 21 experts elicited for probabilistic seismic hazard analysis for Swiss nuclear power plant sites (Abrahamson, et al., 2004). Eighteen experts were elicited regarding future module prices current and emerging photovoltaic (PV) technologies (Curtwright, Morgan, & Keith, 2008). A social cost of carbon expert

elicitation conducted on behalf of the UK Department for Environment, Food and Rural Affairs (2005) employed 14 experts. Usher and Strachan (2013) elicited 25 UK energy experts from academia, industry and government regarding climate, energy and economic uncertainties. Hora and Jensen (2005) reported that five experts served in the expert panel elicitation of seismicity following glaciation in Sweden.

The number of experts selected for elicitation needs to consider whether aggregation will be performed. Such a factor is not always explicitly stated. For example, Meyer and Booker (2001) stated that “Having less than five experts reduces the chances of providing adequate diversity or information to make inferences. (p. 88). Hora (2004) found that there was “little to be gained” (p. 603) from using more than ten experts and five to six experts provide the most benefit.

6.4 Research Question 4 Mathematical Formalism

Given a record containing n observations e' (hereinafter, “obs”), there are $M = \binom{n}{s}$ distinct subsets of size $s \leq n$ obs which can be drawn, without regard to order. For a given method of aggregation, data type (physical or probabilistic), value of s , and record containing $n \geq s$ obs, the average, A of the M MMEs is computed using brute force enumeration. A weight, w is then applied to A . If a record belonging to a theme T containing n_{var} records for which $n_{obs} \geq s$ and a total of n_{themes} containing at least one such record -- the weight w applied to A equals $[n_{themes} \cdot n_{var}]^{-1}$. The sum product of the $w_i \cdot A_i$ over all records having at least s obs, is reported as the overall MME associated with subsets of size s , for the particular method of aggregation.

It will be shown that as the number of experts (nheads) increases, i.e., as the subset size s increases, average MME is generally reduced. This reduction can be combined with the incremental cost of each additional expert engaged to participate in an expert judgment panel, to obtain a total cost function which varies with nheads. The function is developed as follows:

Without loss of generality, assume the penalty cost of a unit increase in MME is unity. Then the cost of incurring an MME which is (MME–1) units greater than unity equals MME–1. If the incremental cost of an additional expert is a fraction R of the unit penalty cost, then the cost of an expert is R. The cost of engaging s experts is 0.1s, and the total cost function, $C = MME - 1 + Rs$.

6.5 Research Question 4 Methodology and Results

Table 29: Distribution of Number of Predictions by Number of Records in EJE shows the distribution of number of predictions, where a prediction corresponds to an expert.

Table 29: Distribution of Number of Predictions by Number of Records in EJE

Number of Observations	Number of Records: Physical Data	Number of Records: Probabilistic Data
1	1141	
2	26	
3	5	
4	65	12
5	64	11
6	84	20
7	137	6
8	27	
9	13	
10	28	
11	49	9
12	8	2
13	8	5
17	47	

Number of Observations	Number of Records: Physical Data	Number of Records: Probabilistic Data
20	1	
31	9	1
45	9	1
Total Number of Records:	1,721	67

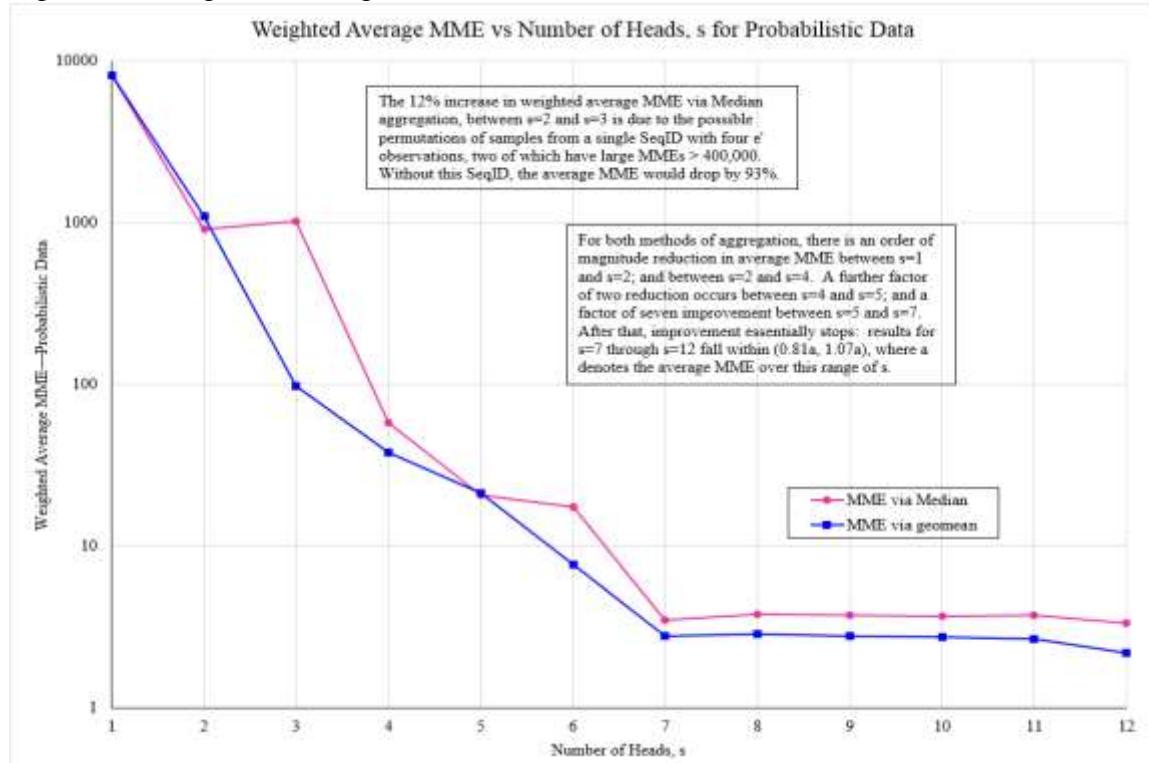
Subsets of size $s=1$ to 12 were considered. Two methods of aggregating them were considered: Median, and Geometric Mean. The Median was a top performer on the MME scorecards, and the geometric mean was generally within twenty percent of other methods (excluding the poorly performing arithmetic and harmonic means). Both of these methods are rated A in terms of ease of computation, see Table 19: Complexity Rating, the Median is computed efficiently by recursively cycling through the distinct possible combinations of e' values, once the latter have been initially sorted). Efficiency is important at $s=12$, as the number of distinct combinations in that case equals 28,760,021,745.

Results for aggregation of subsets of probabilistic data are shown in Figure 27: Weighted Average MME vs Number of Heads, s for Probabilistic Data where s ranges from 1 to 12. As shown in the figure, for both methods of aggregation, there is an order of magnitude reduction in average MME between $s=1$ and $s=2$, that is, when increasing the number of experts from one to two. There is an additional order of magnitude reduction between $s=2$ and $s=4$. A final order of magnitude reduction occurs between $s=4$ and $s=7$. After that, improvement essentially stops: results for $s=7$ through $s=12$ fall within $(0.81a, 1.07a)$, where a denotes the average MME over this range of s . It should be noted that these reductions occur over the entire set of probabilistic meta-data; a

concluding section in this chapter will discuss briefly what may occur when attention is confined to a particular theme or domain.

Figure 27 shows a seemingly anomalous increase in average MME versus s between $s=2$ and $s=3$, using median aggregation. The mechanism driving the anomaly is explained as follows. A single record in the probabilistic dataset had $n=45$ obs (SeqID EJEP67); at $s=45$, the MME is reduced by only two percent from the average value computed for this record at $s=12$. A second record in the dataset has $n=31$ obs (SeqID EJEP61); at $s=31$, the MME is decreased by 0.1% compared to the average value computed for this record at $s=12$ using geometric mean aggregation (an 0.4% increase was recorded for aggregation via median).

Figure 27: Weighted Average MME vs Number of Heads, s for Probabilistic Data



For aggregation via the Median, significant anomalous increases in average MME versus subset size s can occur. This occurs because of the different methods of computation of the median depending on whether s is even or odd. For s odd, the “middle” value of the sorted e' is used as the median; for s even, the average of the two “middle values” is used. The following notional example will show how these different methods of computation can cause average MME versus subset size to decrease from $s=1$ to $s=2$, then increase from $s=2$ to $s=3$.

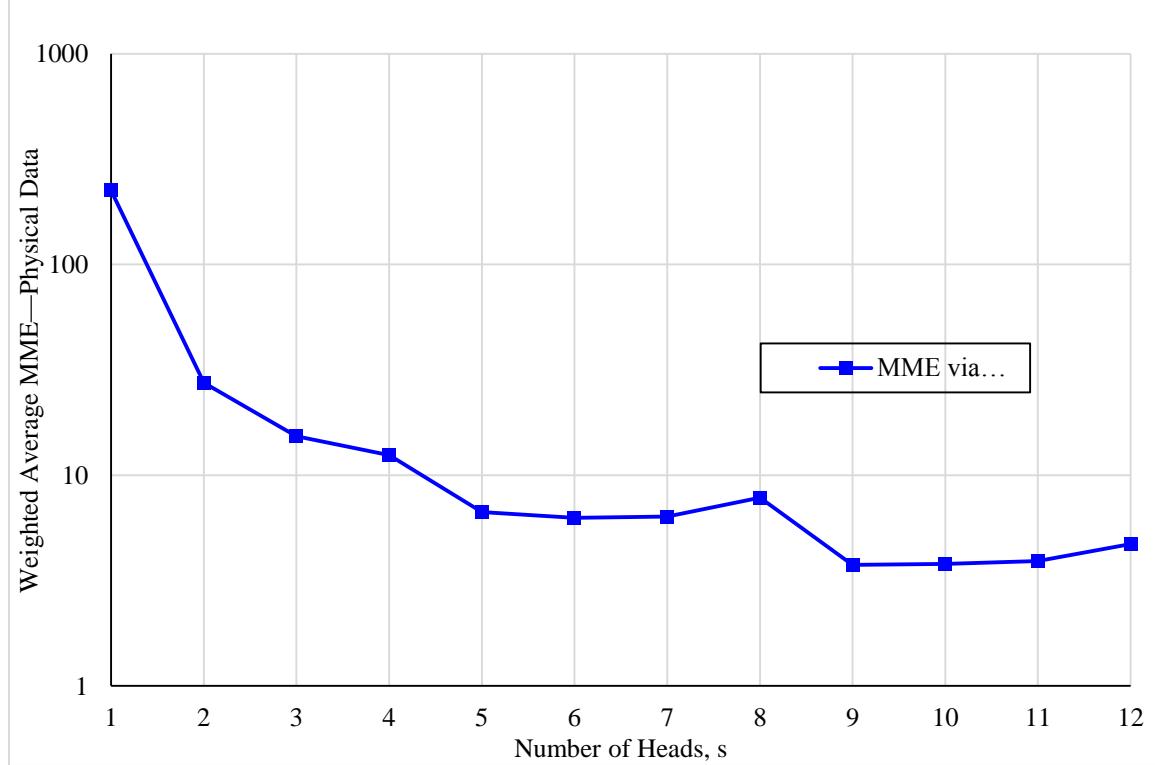
Consider a record containing $n=3$ obs $e' = 0.001, 0.01$ and 1 , and the true value of e equals 1 . At $s=1$, there are three possible values of MME depending on which e' is drawn: $1000, 100$ or 1 . Therefore, the average MME at $s=1$ will be 367 . At $s=3$, the middle value of e' must be chosen: 0.01 ; this yields an MME of 100 . However, at $s=2$, the median is found by taking the average of two values. There are three possibilities: 0.001 and 0.01 ; 0.001 and 1 , and 0.01 and 1 . These give rise to medians of 0.0055 , 0.5005 , and 0.505 , respectively. The corresponding MMEs are approximately 181.82 , 2.00 , and 1.98 , respectively. The average MME is, therefore, 61.93 . Thus for this example, there is a factor of six reduction in average MME between $s=1$ and $s=2$, followed by a 60% increase between $s=2$ and $s=3$.

For physical data, a similar phenomenon caused a 37% increase in average weighted MME, from 7.04 to 9.67 , between $s=6$ and $s=7$. The increase was associated with the different methods of computation and permutations of e' values selected in the subsets for two SeqIDs, EJEPHYS1131 and EJEPHYS1134 (both having $n=8$ obs, with all e' values less than e , in some cases, much less: by factors of 30,000). For this reason,

results for aggregation of subsets of physical data via geometric mean only are shown in

Figure 28: Weighted Average MME vs Number of Heads, s for Physical Data.

Figure 28: Weighted Average MME vs Number of Heads, s for Physical Data



Even using geometric mean only, an increase of 23% in average weighted MME was observed between $s=7$ and $s=8$. This was due to over 137 records with $n=7$ and somewhat lower MMEs for individual SeqIDs dropping out of the mix at $s=8$. The average MME for each SeqID present at $s=8$ (199 records) was compared to the average MME for the same SeqID at $s=7$. The ratio of the latter to the former ranged between 1 and 1.21, averaging 1.02.

The same general trend observed for probabilistic data can be seen for physical data: A large drop — an order of magnitude reduction in average MME — occurs between $s=1$ and $s=2$. A more moderate factor of two reduction is observed for physical data between $s=2$ and $s=4$; and a final factor of two reduction, to 6.7 is observed between

$s=4$ and $s=5$. Further increases in s are associated with fluctuations in average MME: the latter decreases to 6.3, then increases to 7.8 as s is increased from 5 to 8. A final factor of two reduction in average MME, to 3.8, is observed at $s=9$, followed by increases in average MME of over twenty percent—to 4.7—as s continues to increase, from 9 to 12.

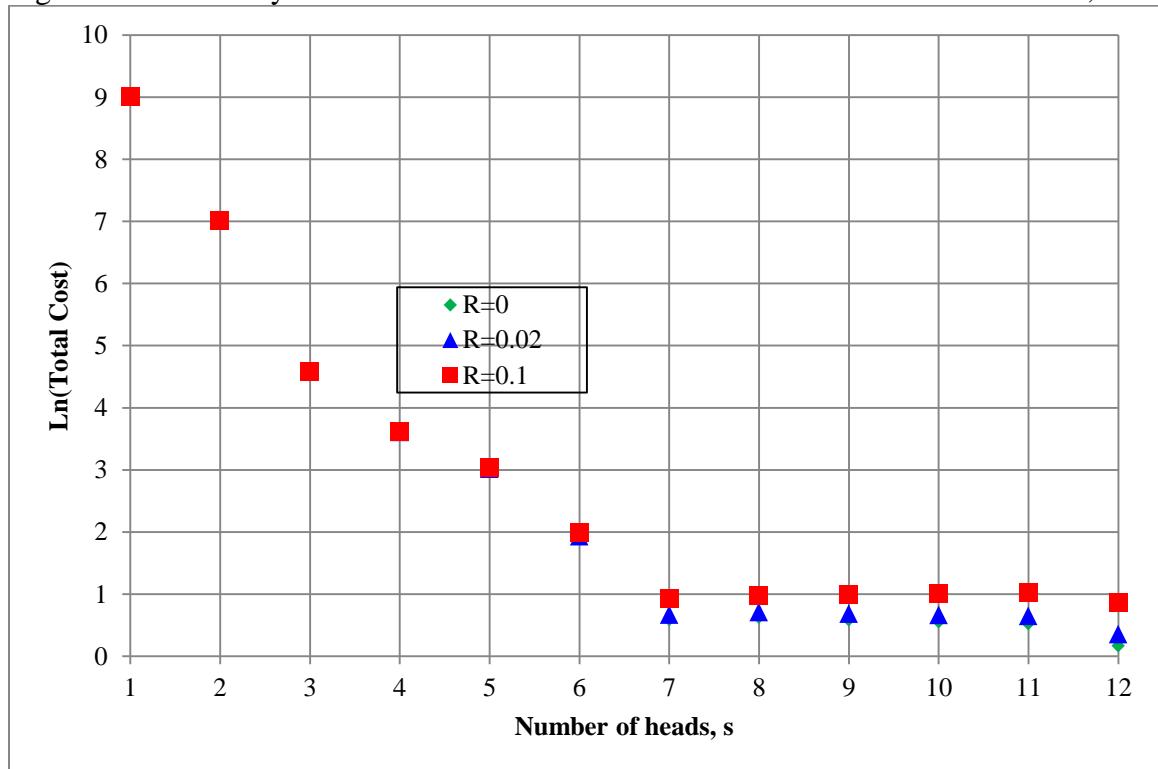
The reduction in MME associated with increased number of experts, or “nheads” discussed above, can be combined with the incremental cost of each additional expert engaged to participate in an expert judgment panel, to obtain a total cost function, $C(s)$ which varies with nheads, s .

For example, suppose an agency is considering how many experts to engage to estimate the incremental effectiveness of a proposed improvement to an air traffic control system, which would be applicable at a number of small airports. The projected pool of fatalities supposed to be addressed by the improvement over a nominal lifecycle, say ten years, is initially estimated by the agency to have a value of one. Applying a Department of Transportation policy value of approximately \$10M per statistical life saved, the pool is equivalent to \$10M times the number of fatalities. Supposing the true but unknown incremental effectiveness of the system is ten percent, the system would be worth \$1M. If the experts overestimated the effectiveness of the system by a factor of $MME=2$, this would represent a unit penalty cost of \$1M. This follows from the fact that they would incorrectly assume that the system effectiveness is twenty percent instead of ten percent, thus $0.2 * \$10M = \$2M$ safety benefits in avoided fatalities, instead of $0.1 * \$10M = \$1M$ benefits. If each expert consultant is engaged for two weeks at a nominal cost of \$10,000, then the ratio of expert cost to unit MME cost, $R=0.01$.

Figure 29: Sensitivity of Total Cost for a Probabilistic Variable to Number of Heads, s and

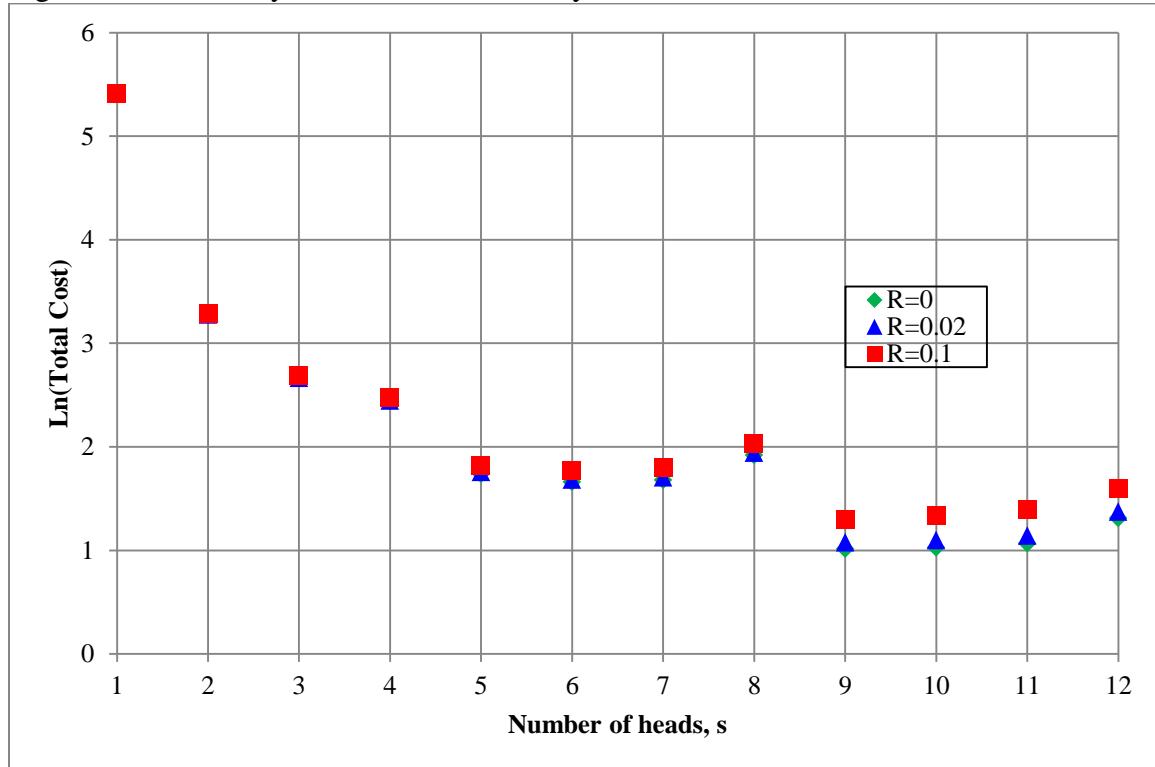
Figure 30: Sensitivity of Total Cost for a Physical Variable to Number of Heads, s shows the interplay of total cost with n_{heads} using geometric mean aggregation, for probabilistic and physical data, respectively, at R values of 0, 0.02 and 0.1.

Figure 29: Sensitivity of Total Cost for a Probabilistic Variable to Number of Heads, s



At all three levels of R , there is a large drop in the cost function between $s=1$ and $s=7$, after which a plateau is reached for $R=0.1$. The drop observed at $s=12$ for smaller values of R must be considered an artifact arising from the dropping out of half of the 18 records present at $s=11$. Those nine absent records had an average MME of 3.84, compared to the average MME of 2.95 for the remaining nine records.

Figure 30: Sensitivity of Total Cost for a Physical Variable to Number of Heads, s



For physical variables, there is a similar large drop between $s=1$ and $s=6$, after which the cost function fluctuates between increases and decreases of approximately 40% with increasing values of s . Note that the cost function increases as s moves beyond 9, towards 12.

6.5 Summary

It was shown in this chapter, for both probabilistic and physical meta-data, that there is an order-of-magnitude reduction in average MME as the number of experts is

increased from one to two, regardless of whether geometric mean or median aggregation is used. For probabilistic data, an additional order of magnitude reduction is observed as the number of experts increases from two to four. For physical data, the reduction is approximately a factor of two. As the number of experts increases beyond six, to approximately six, MME drops by roughly an additional thirty percent. Beyond this number of experts, improvement is minimal, and depending on the cost of experts, an overall cost function incorporating both the penalty associated with MME, as well as the cost of engaging experts, may actually increase.

It should be noted that the order-of-magnitude reduction in average MME is observed when the entire meta-data set of a given type, probabilistic or physical, is included. However, if attention is confined to a particular subset of the data, there may be much more modest gains from increasing the numbers of experts. Shanteau (2015) noted that for certain domains, such as those involving physical systems, or where stimuli were “relatively constant, ... judges were faced with stationary targets. In contrast, domains with poor performance involved dynamic stimuli, generally involving human behavior.” (p. 172). Judges performed more poorly in the latter types of domain. This phenomenon is present in the EJE database. For example, for the theme pm25, which involves particulate emissions, the ratio of maximum to minimum e' against each of the variables in this theme, averages less than 1.14. The average MME when only a single prediction from this theme is chosen, is 1.12. Increasing nheads to six decreases the average MME over this theme by one percent, to 1.11. By contrast, the theme INFOSEC (information security) represents a newer, dynamic field involving human behavior. For this theme, the ratio of maximum to minimum e' against each of its probabilistic variables

averages 30,000. The corresponding average MME is 684 when nheads equals one; it declines to 6.5 when nheads equals six. In conclusion, the applicable domain is highly significant as to the extent of improvement associated with increases in numbers of experts.

Chapter 7: Research Question 5 (Parametric Distributions for Bounds)

7.1 Introduction

Are there simple parametric distributions that perform reasonably well in yielding bounds around an estimate, \hat{e} obtained by aggregation? It will be shown in this chapter that with respect to physical data, there does indeed exist a simple distribution, the Cauchy, which performs nearly as well in yielding bounds around \hat{e} as the more complex methods discussed in Chapter 4. It will also be shown, however, that this is not the case for probabilistic data.

7.2 Research Question 5 Significance

If a simple parametric distribution can be found which performs reasonably well in yielding bounds around an estimate, it may represent an easy way to obtain confidence bounds, compared to more complex techniques involving Bayesian aggregation or quantile regression.

7.3 Research Question 5 Mathematical Formalism

Deweighted samples of $\ln(e/\hat{e})$ values for several methods of aggregation were run through MathWave EasyFit. The samples had size 123,765 and 7,773 for physical and probabilistic data, respectively, as discussed in Chapter 4. The methods of aggregation included arithmetic and geometric means, as well as median. Harmonic mean was not used, since for both data types, the MMEs associated with its point estimates exceeded by orders of magnitude their counterparts obtained via geometric mean or median aggregation.

For physical data, the Cauchy distribution was the best fit, ranking 1 or 2 on each of the three MathWave EasyFit ranking criteria: Kolmogorov-Smirnov, Anderson-Darling, and chi-square. For aggregation via the median, the Cauchy ranked 1 on all three criteria. Where a competitor distribution had a higher ranking against one of the above criteria, it had much lower ranking, e.g. 5th or 10th against other criteria. For aggregation of probabilistic data using the median or geometric mean, similar results applied. The arithmetic mean constituted an exception: instead of ranking 1 or 2 against all criteria, the Cauchy ranked (3,6,2); however, other distributions had higher sums of ranks. (The five-parameter Wakeby distribution ranked (1,3,4) and thus had a lower sum of ranks, but was excluded as not simple.)

The Cauchy has two parameters: location and scale, m and s , respectively. The quantile q corresponding to the p th percentile point is given by $q = m + s \cdot \tan(\pi \cdot (p - 0.5))$. Since \exp is a monotonic transformation, the corresponding bound on e is given by $\hat{e} \cdot \exp(q)$.

7.4 Research Question 5 Methodology and Results

For physical data, with \hat{e} obtained via the median, the Cauchy fit parameters m and s are zero and 0.22336, respectively. The corresponding (0.05, 0.95) factors for 90% bounds are 0.244 and 4.097. For 80% bounds, they are .503 and 1.989. The corresponding factors for \hat{e} obtained via the geometric mean differ from the preceding factors by less than one percent on average. The coverage of the 90% bounds around \hat{e}_{Median} obtained using the Cauchy is 0.90; and $\text{ABSDist}(\text{Cov} - 0.9)$ equals 0.106. These metrics represent coverage performance almost as good as that shown in

Table 12: EJE Physical Data - 90% Coverage by Aggregation Method for bounds around \hat{e}_{Median} obtained using quantile regression. The one-sided multiplicative bounds width of 4.1 is approximately twenty percent higher its counterpart in

Table 12. Similarly, the coverage of the 80% bounds obtained using the Cauchy is 76%, and ABSDist(Cov=0.8) is equal to 0.19. These results are within 5% and 22%, respectively of those reported for their counterparts in Table 14: EJE Physical Data - 80% Coverage by Aggregation Method for bounds around \hat{e}_{Median} obtained using quantile regression. In that table, the coverage is 80%, and ABSDist(Cov=0.8) is equal to 0.16. For Cauchy-based 80% bounds around $\hat{e}_{\text{Geometric Mean}}$, coverage and ABSDist values were a few percent worse than their counterparts for \hat{e}_{Median} . However, the Cauchy-based 90% bounds performed as well: they had 90% coverage, and an ABSDist(Cov=0.9) value equal to 0.101. Finally, for the arithmetic mean, the one-sided multiplicative bounds width was significantly wider, at 5.4. Since the median is a top performing aggregation method, the Cauchy-based simple parametric bounds of $(\hat{e}/4.1, \hat{e}\cdot 4.1)$ and $(\hat{e}/2, \hat{e}\cdot 2)$ seems a viable technique for obtaining 90% and 80% intervals around a point estimate \hat{e} obtained via the median.

Discussion of methodology and results are now presented for probabilistic data. For \hat{e} obtained via the median for probabilistic data, the Cauchy fit parameters are $m=0.0546$ and $s=0.39188$; $q_{0.05} = -2.420$; the 5th percentile bound = $0.089\hat{e}$; $q_{0.95} = 2.529$; the 95th percentile bound is $12.539\hat{e}$. The one-sided multiplicative bounds width associated with these 90% bounds of $(0.089, 12.539)$ is 11.9, which is more than twice the corresponding width of approximately 5.0 shown in Table 15 in Chapter 4 for 90% bounds around the median, obtained using the more complicated quantile-regression

technique. Eighty percent bounds are $(0.316\hat{e}, 3.528\hat{e})$. Where the 95th percentile bound exceeds one, the left end of the eighty percent bounding interval can be used to construct a 90% interval: $(0.316\hat{e}, 1)$. As an example, for SeqID EJEPROB17, where the median, $\hat{e}=0.012$, the parametric fit 90% bounds would be $(0.001, 0.15)$.

The Cauchy parameters for \hat{e} obtained via the geometric mean are $m=0.1249$ and $s=0.53947$, yielding ninety and eighty percent bounds of $(0.038, 34.159)\hat{e}$ and $(0.215, 5.961)\hat{e}$, respectively. Finally, the arithmetic mean has bounds factors of $(0.056, 19.048)\hat{e}$, which fall between those for the median and the geometric mean.) Location and scale parameters along with 5th and 95th percentile bounds factors for both physical and probabilistic data are shown in Table 30: Cauchy Parameters for Bounds. Coverage and ABSDist(Cov–0.9) values obtained by applying the Cauchy 90% bounds factors to \hat{e}_{Median} were 0.87 and 0.13, respectively. The coverage figure is slightly closer to 90% than its counterpart of 0.85 obtained via quantile regression, shown in Table 15: EJE Probabilistic TUD Data - 90% Coverage by Aggregation Method. The ABSDist(Cov–0.9) figure is approximately 15% higher than its counterpart in Table 15. However, the one-sided multiplicative bounds width of approximately 12 is more than twice the width shown in Table 15. For Cauchy-based bounds around geometric or arithmetic means, the one-sided widths are even greater: approximately 20 and 30, respectively. These are more than double the Bayesian-based bounds width of 9 shown in that table.

Table 30: Cauchy Parameters for Bounds

Parameter	Physical Data			Probabilistic Data		
	$\hat{e}_{\text{ArithmeticMean}}$	$\hat{e}_{\text{Geometric mean}}$	\hat{e}_{Median}	$\hat{e}_{\text{ArithmeticMean}}$	$\hat{e}_{\text{Geometric mean}}$	\hat{e}_{Median}
m	-0.034	0	0	0.035	0.125	0.055

	Physical Data			Probabilistic Data		
s	0.262	0.225	0.223	0.461	0.539	0.392
0.05 bound	0.185	0.241	0.244	0.056	0.038	0.089
0.95 bound	5.044	4.149	4.097	19.048	34.159	12.54

The principal disadvantage of the Cauchy-based simple bounds technique is that it is insensitive to the magnitude of e . Equal width bounds are always applied, irrespective of the aggregate point estimate \hat{e} and implied magnitude of e . This ignores the fact that for probabilistic data, there tends to be a wider spread around small \hat{e} values than around relatively large \hat{e} values approaching unity. For physical data, this presents less difficulty, as there is a smaller change in spread versus level of \hat{e} . These facts are illustrated in

Figure 31: Quantiles of $e|\hat{e}$ for **Probabilistic Data** and Figure 32: Quantiles of $e|\hat{e}$ for Physical Data, obtained by applying quantile regression and a quadratic model to fit weighted pairs of $(e, \hat{e}_{\text{Median}})$ data.

Figure 31: Quantiles of $e|\hat{e}$ for Probabilistic Data

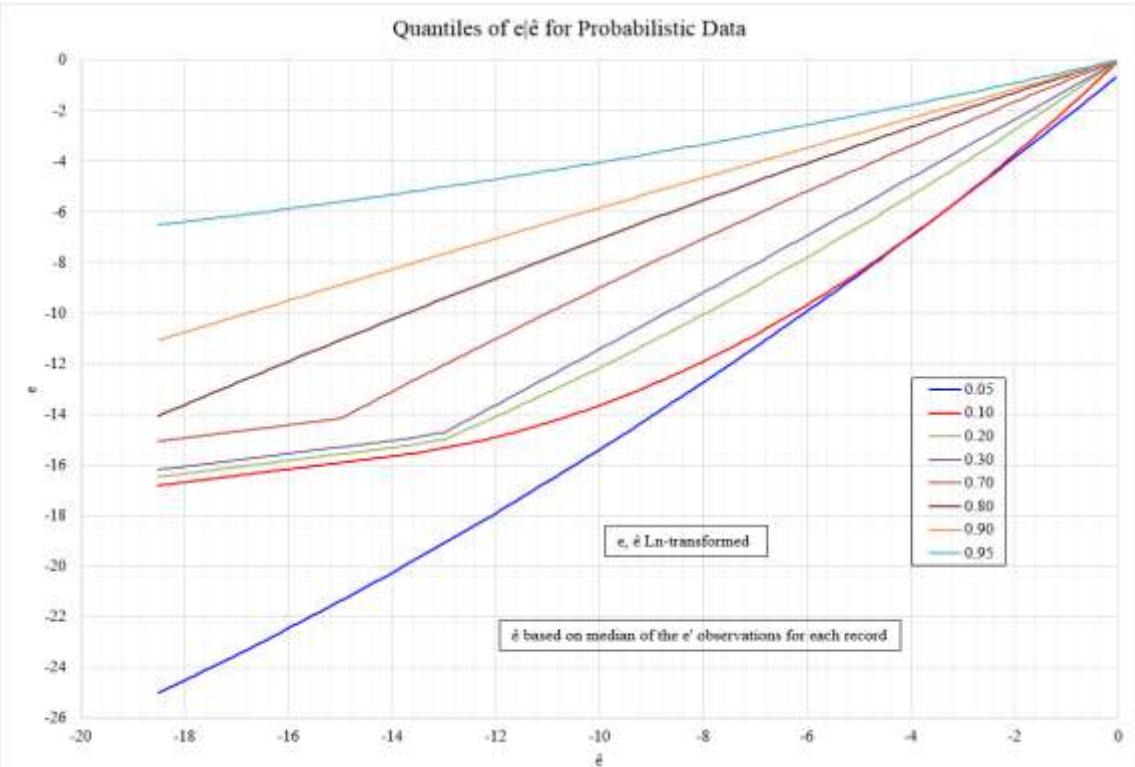
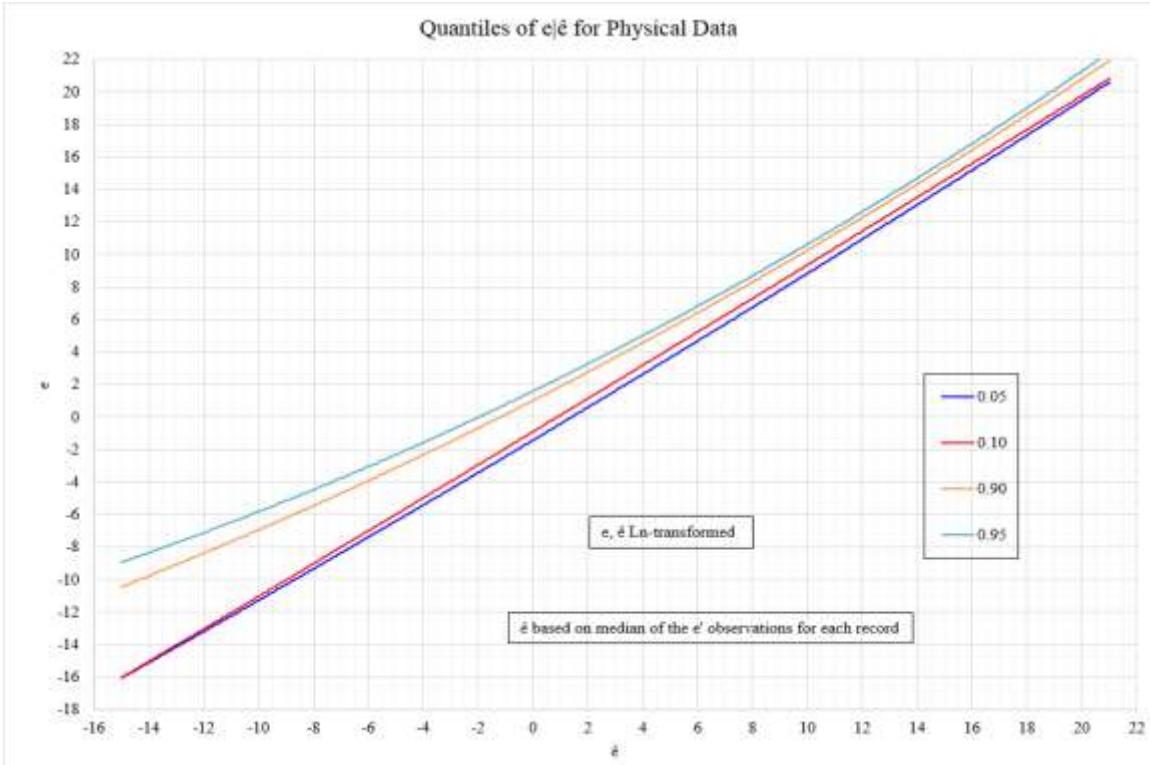


Figure 32: Quantiles of $e|\hat{e}$ for Physical Data



7.5 Summary

In conclusion, a simple parametric fit to physical data based on fitting a Cauchy distribution to the metadata can be used with some success to obtain 90% and 80% bounds around a point estimate \hat{e} obtained via the median. The approximation is as follows: for 90% bounds, take $(\hat{e}/4.1, \hat{e}\cdot 4.1)$; for 80% bounds, take $(\hat{e}/2, \hat{e}\cdot 2)$. For probabilistic data, a simple parametric fit based on the Cauchy may be applied to \hat{e} obtained via the median, however, it must be recognized that the resulting bounds will be at least twice as wide as those obtained by the more complex methods discussed in Chapter 4.

Chapter 8: Analysis of Test Data

8.1 Introduction

The Test Data (TD) set was collected solely during the course of this research, with the objective of comparing the performance of the aggregation methods per preceding chapters against new data separate from the EJE database. This TD set contains 26 probabilistic data records and 11 physical data records. There are five themes in each TD data category; two themes are common to each category. The relatively small size of the TD, 37 records in total encompassing 190 predictions, is noted as a limitation. This chapter includes a description of the TD set, the data processing results, and a comparison of aggregation results from the EJE and TD datasets.

8.2 Test Data Description

This section specifies the TD in terms of theme, source, and data category. A summary background description of the data is also provided.

i. Theme: Elections 2012. Probabilistic Data: one record; Physical Data: one record; The Washington Post (2012).

The Washington Post (2012) elicited predictions from Subject Matter Experts (SMEs) the day before the 2012 General Elections regarding different outcomes. The percent of the popular presidential vote for the incumbent was selected as the variable of interest for the Probabilistic Data Category. The number of Senate seats for the incumbent party was selected as the variable of interest for the Physical Data Category. The realizations were obtained from Federal Elections Commission (2013).

ii. Theme: Forecaster Reliability; Probabilistic Data; 21 records; Murphy and Winkler (1977).

Murphy and Winkler (1977) conducted a study regarding the reliability, accuracy, and skill of probability forecasts, where reliability was defined as “the degree of correspondence between forecast probabilities and the observed relative frequencies over a set of forecasts” (p. 2). Specifically, for precipitation, these forecasts are known as probability of precipitation (PoP) forecasts. PoP and realized values for two forecasters, A and B were read from Murphy & Winkler (1977), Figures 2 and 3 respectively. The variables are reported as Rainfall Forecast Probability Forecaster (A or B) Point (number unique to the forecaster).

iii. Theme: Unscheduled Outages; Probabilistic Data: two records; Physical Data: two records; FAA Internal Workplace Survey

In order to obtain additional expert judgment data to test the performance of the various aggregation approaches, four questions involving FAA equipment failures were composed in July, 2014 for a group of FAA colleagues familiar with FAA equipment and/or its use in air traffic control.

The first question was, “What will be the median duration of unscheduled full service outages (interrupt code FL) in August, 2014 attributable to software (cause code 86) in Technet?” The realized value was 0.45 hours. As it involved time, it was considered a physical quantity. Technet is an FAA logging system used by the technician workforce to track equipment outage cause, duration, impact and resolution. It was used as the source for realized values, e, after analysis on November 6, 2014 of counts and durations of pertinent outages.

The second question was, “What will be the median duration of unscheduled full service outages (interrupt code FL) in August, 2014 attributable to hardware (cause code 80) in Technet?” The realized value was 7.6 hours. As it involved time, it was considered a physical quantity.

The third question was “What will be split between the numbers of such SW and HW outages?” Answers were generally expressed as odds, e.g. “1 to 6” (SW:HW), or “200 to 1” (HW:SW); as such, they were considered probabilistic. They were transformed into probabilities, p that an outage was caused by software, via the formula $p=1/[ODDS+1]$, where ODDS denotes the HW:SW odds ratio, e.g. 6/1 or 200/1. The realized value was 0.001855, corresponding to a HW:SW split of 538:1.

The fourth question was “What is the fraction of all unscheduled outages (code 80 series) that will be attributed to personnel error (cause code 89, supplemental codes 5,6,7, or 8 [Personnel Error AF, AT, Other FAA, Non-FAA, respectively].)” This was probabilistic; the realized value was 1.24%.

iv. Theme: Antiretroviral Therapy Response; Probabilistic Data; one record; Zazzi et al. (2011).

A study was conducted to compare the EuResist expert system for antiretroviral therapy treatment with human expert estimates for the 8-week response to 25 treatment cases. The results are provided in Zazzi et al. (2011), Figure 1 Prediction of treatment outcome for the 25 patient cases by the 10 human experts and the EuResist expert system. According to Zazzi et al. (2011), “There were 15 treatment successes and 10 treatment failures”, (p. 211).

v. Theme: ASDE-X Effectiveness; Probabilistic Data; one record; FAA Internal
(data are FOIAble)

Data were available from a January, 2007 FAA expert judgment panel of four pilots and three controllers who were asked to predict the effectiveness of a safety system called ASDE-X, consisting of surface surveillance radar with aircraft data tags and conflict alerts, in preventing runway incursions. Against a pool of 38 serious (category A) incursions including five accidents, the panel was asked to state on a five-point Likert scale (1=Definitely not; 2=Probably no; 3=50/50; 4=Probably yes; 5=Almost definitely yes) whether the incursion would have been prevented if ASDE-X had been available. Each response was multiplied by 0.25, after subtracting one, to convert it into a probability between 0 and 1. For realized value, the FAA baselined ASDE-X effectiveness of 73% was used.

vi. Theme: Ongoing Software Development Effort; Physical Data; one record;
Jørgensen & Moløkken (2002).

Twenty SMEs were assigned to five different teams (four SMEs per team) and asked to estimate the number of hours for an identical software development effort for their assigned teams. The project was ongoing at the time of the elicitation; the SMEs were not informed of this fact. Three estimates were elicited from each SME; minimum, maximum, and most likely number of hours. The latter value was used to represent the predicted value. According to Jørgensen and Moløkken (2002), “The actual effort of the specified project turned out to be about 2,400 work-hours” (p. 426).

vii. Theme: Software Reliability Growth Models; Physical Data; one record;
Almering, van Genuchten, Cloudt, & Sonnemans, (2007).

Software Reliability Growth Models (SRGMs) are used to describe the process of finding and removing faults to improve the software reliability using a mathematical relationship. Figure 4 (Almering, et al., 2007, p. 86) shows five experts' fault estimations at four specific time periods (hours; t=0; t=260; t=570; t=950) during a project for developing software for high-end TV sets containing several million lines of code. The estimates for each of the five experts was read off the graph for the second time period at t=260 hours (20% of total time). This time period was selected since estimates were not adjusted, whereas they were adjusted for subsequent time periods after the experts reviewed the realized values.

viii. Theme: Software Hours Estimate Bid Proposal; Physical Data; six records; Faria, P. & Miranda, E. (2012).

The number of staff-days for six different software projects were elicited from 14 experts for five projects and thirteen experts for one project during the bid-phase. The realizations were presented in a table (Faria & Miranda, 2012, Table 3) and in graphs (Faria & Miranda, 2012, Figure 3). The experts' predictions were provided in graphs (Faria & Miranda, 2012, Figure 3).

8.3 Test Data Results

The MME scorecards for the TD set are shown in

Table 31: MME Scorecard for TD **Probabilistic Data** and

Table 32: MME Scorecard TD **Physical Data**. The construction of these scorecards is described in Chapter 4.

With respect to the Probabilistic Data, the Harmonic Mean had smallest average MME, at 2.27. This is contrary to the case for EJE data, where the harmonic mean had the worst performance in terms of average MME for both physical and probabilistic data. This result is an artifact of a single record from the Unscheduled outages theme, SeqID TDPROB26, involving the probability of a software versus hardware outage. There was a large discrepancy between the true, small probability of 0.001855 that an outage was caused by software rather than by hardware, and the orders of magnitude larger predictions, e' of all experts save one. The five predictions ranged from 0.004975 (representing a split of 1:200 SW to HW), to 0.33. Since the harmonic mean inverts e' values before averaging them, the expert with the small probability estimate received disproportionate weight, and the MME for this method of aggregation was approximately ten; MMEs for all other methods were approximately one hundred. Since this record was assigned a weight of 0.1 (it was one of two records in five probabilistic test data themes), all other methods, even if comparable against all other records, had weighted average MMEs larger than that associated with the harmonic mean, varying between 7.5 for Geometric Mean, to 13.6 for MLE.

When this highest MME record was removed from the average, the harmonic mean dropped from first place to the middle of the pack in terms of rank, and the median assumed first place at 1.11. Except for the Geometric Mean (reweighted average 2.0), all other aggregation methods had reweighted average MMEs less than approximately 1.3. The alpha stable and ROT MMEs were virtually the same for all MME computations for probabilistic as well as physical data. With respect to median MME for probabilistic data, the Median took first place, at 1.08; Geometric Mean was highest, at approximately 2.

Table 31: MME Scorecard for TD Probabilistic Data

MME	No. Records	Arithmetic Average	Geometric Mean	Harmonic Mean	Median	Alpha-Stable	Rule of Thumb	MLE	Bayesian
Average	26	11.35	7.52	2.27	12.85	9.39	9.4	13.6	11.69
Average Rank	26	5	2	1	7	3	4	8	6
Average*	25	1.22	2.04	1.15	1.11	1.13	1.13	1.25	1.32
Average* Rank	25	5	8	4	1	2	2	6	7
Median	26	1.11	2.01	1.14	1.08	1.12	1.12	1.31	1.35
Median Rank	26	2	8	5	1	3	3	6	7

* Largest MME eliminated

With respect to the Physical Data, as shown in Table 32, the average MME ranged from 8.5 (MLE) to 3.6 (Geometric Mean); seven of the eight values were less than 5. When the record with the largest MME is removed from each method, reweighted average MMEs fall between 2.7 (Arithmetic MEan) and 4.56 (MLE); the other six methods have reweighted average MMEs falling in a relatively narrow range (3.3, 3.6). The record from the Unscheduled outages theme involving median duration of

unscheduled full service outages attributable to software, contained the Maximum MMEs for the Arithmetic, MLE, and Bayesian methods. The realized value is 0.45 hours; the five predictions ranged from 0.25 to 20 hours. For the other aggregation methods, the record giving rise to the largest MME involved software hours bid. The realized value is 2,600 hours; the 14 observations ranged from 75 hours to 2,500 hours. With the exception of one prediction, all predictions were less than 700 hours. With respect to median MME, values ranged from 2.27 to 2.5; the Median had the lowest value; the Geometric Mean, the highest.

Table 32: MME Scorecard TD Physical Data

MME	No. Records	Arithmetic Average	Geometric Mean	Harmonic Mean	Median	Alpha-Stable	Rule of Thumb	MLE	Bayesian
Average	11	4.46	3.61	3.67	3.8	3.75	3.75	8.54	4.89
Average Rank	11	6	1	2	5	3	3	8	7
Average*	10	2.72	3.38	3.33	3.58	3.52	3.52	4.56	3.59
Average* Rank	10	1	3	2	6	4	4	8	7
Median	11	2.36	2.5	2.38	2.27	2.41	2.44	2.35	2.36
Median Rank	11	4	8	5	1	6	7	2	3

* Max MME eliminated

8.4 Comparison between EJE and TD Results

The following sections compare EJE and TD results for MME performance and bounds coverage for physical and probabilistic data. The EJE data are drawn from the TUD records.

8.4.1 Probabilistic Data

Once the TD record with highest MME was removed for each aggregation method, the Median took first place for both average and median MME, with values of 1.11 and 1.08, respectively. The corresponding values for the Geometric Mean were approximately twice these. For all other aggregation methods, values were within 20% of those associated with the Median method. The Median was also the top ranked method for probabilistic aggregation against the EJE data set.

8.4.2 Physical Data

For the EJE physical data set, the Median method had lowest average MME, both including and excluding the record with the maximum MME; while the Harmonic and Arithmetic Means were the worst and second worst methods, respectively, in both cases. The Median also had a median MME which was only one percent higher than the lowest MME. Thus, it was a strong performer against physical EJE data, with respect to accuracy. By contrast, for the TD data set, the Arithmetic and Harmonic Means have lowest and second lowest average MME when the highest MME is removed, at 2.72 and 3.33, respectively. The MLE has highest average MME, both including and excluding the highest MME record. At 3.8, the average MME for the Median is close to the lowest average MME value of 3.61 (Geometric Mean). When the highest MME record is removed from the average, the Median has a reweighted MME value of 3.58, part of a cluster of values for all other methods falling in the range (3.33, 3.59). For median MME, values range from 2.27 to 2.5; the Median has the lowest value. Thus, the Median still has relatively good performance against the TD data, consistent with its performance for physical data.

8.4.3 Bounds Coverage Against Probabilistic TD

Of the six aggregation methods for which RSS bounds coverage results were supplied against EJE data—Geometric, Median, Alpha-Stable, ROT, Bayesian, and Classical—bounds for the last method were not available for the test data. Per Table 33: Weighted Probabilistic TD Bounds Coverage, the remaining five methods have approximately 90% coverage of the weighted probabilistic test data, at both nominal 80% and 90% bounds.

Table 33: Weighted Probabilistic TD Bounds Coverage

Coverage	Alpha-Stable	Bayesian	Geometric	Median	ROT
90%	0.90	0.88	0.90	0.88	0.90
80%	0.90	0.88	0.90	0.88	0.90

Coverage against the first two of the five test data themes (vote share of the incumbent in U.S. presidential election 2012, and rainfall forecast probability) is even higher, at between ninety-five and one hundred percent for all methods. This is unsurprising, as the rainfall probability forecasts are considered relatively well-calibrated; and the vote share of the incumbent in U.S. presidential elections generally falls within a relatively narrow band. Similar results were obtained for the simple parametric fit bounds developed using the Cauchy distribution, as shown in Table 34: Probabilistic TD - Parametric Fit Bounds Developed Using Cauchy Distribution.

Table 34: Probabilistic TD - Parametric Fit Bounds Developed Using Cauchy Distribution

Coverage	Arithmetic	Geometric	Median
90%	0.90	0.90	0.90
80%	0.90	0.90	0.90

The scaling factors applied to aggregated \hat{e} values obtained under each of the three aggregation methods shown in the table—Arithmetic Mean, Geometric Mean, and Median—ranged from 3.1 to 6 on each side of \hat{e} for 80% bounds; and from 11 to 34 for 90% bounds. However, for the probabilistic test data set, with the exception of SeqIDs TDPLOB23 and TDPLOB26, no observation e' or aggregated estimate \hat{e} had a multiplicative excursion from e exceeding 3.05. For SeqID TDPLOB23, a single e' had an excursion exceeding a factor of six, however \hat{e} was within a factor of 2.2 of e , under each of the three aggregation methods. Since the one-sided bounds factors at 80% exceed 3.1 for each of the aggregation methods, they captured all of the e values associated with the probabilistic test data records—except for SeqID TDPLOB26, which involved the probability of a software-induced outage versus one caused by hardware. For this SeqID, e' observations diverged from e by factors as large as 180, and \hat{e} diverged from e by factors of 55 or greater. The one-sided 90% bounds intervals were not wide enough to capture the true value of e in this case. As this record had mass 0.1, the weighted coverage was 0.9 for both 80% and 90% bounding intervals.

Excursions between e' and e were generally smaller for the probabilistic test data than for the EJE data set, except for SeqID TDPLOB26 of the former, where three of five e' values exceeded 100 $\cdot e$. The corresponding weight associated with these three values was six percent; vice two percent of full probabilistic EJE data e' values exceeding 100 $\cdot e$. It is unsurprising that the parametric bounds at nominal 80% and 90% levels covered all test records except for SeqID TDPLOB26.

8.4.4 Bounds Coverage Against Physical TD

For physical test data, a somewhat reverse situation held, as shown in Table 35:

Weighted Physical TD Bounds Coverage the bounds developed using EJE physical data were not wide enough to cover the test data at nominal 80% and 90% levels.

Table 35: Weighted Physical TD Bounds Coverage

Coverage	Alpha-Stable	Bayesian	Geometric	Median	ROT
90%	0.20	0.73	0.40	0.50	0.30
80%	0.20	0.70	0.20	0.30	0.20

This was largely the result of a single theme, “Software Hours Bid Phase”, having twenty percent of the total mass, and involving relatively large excursions between e and e' values. Two thirds of its records had approximately forty percent or more of their observations an order of magnitude less than e; with generally only approximately ten percent of the e' values as large as e. The Bayesian method of aggregation captured the value of e for only a single one of the six records in this theme; the other methods did not capture any of the values of e in the theme.

With this theme removed, coverage is as shown in Table 36: Weighted Physical TD Bounds Coverage Subset "All themes except Software Hours Bid Phase"; mass 0.8.

Table 36: Weighted Physical TD Bounds Coverage Subset "All themes except Software Hours Bid Phase"; mass 0.8

Coverage	Alpha-Stable	Bayesian	Geometric	Median	ROT
90%	0.25	0.87	0.50	0.63	0.38
80%	0.25	0.87	0.25	0.38	0.25

The simple parametric fit bounds developed using the Cauchy distribution also failed to achieve 90% or 80% coverage at those nominal levels, as shown in Table 37: Physical TD - Parametric Fit Bounds Developed Using Cauchy Distribution.

Table 37: Physical TD - Parametric Fit Bounds Developed Using Cauchy Distribution

Coverage	Arithmetic	Geometric	Median
90%	0.77	0.73	0.73
80%	0.23	0.20	0.30

For 90% nominal bounds, the scaling factors ranged between 4.1 and 5.4 on each side of \hat{e} ; at 80% they are 2.2 for arithmetic mean, and 2.0 for the other aggregation methods. SeqID TDPHYS6, with mass 0.2, involving fault estimation at 260 hours, has an excursion of 2.5 between \hat{e} and e for the arithmetic mean; and excursions of 2.3 for the other two aggregation methods. Therefore, this sequence is not captured at the 80% bounding level. Similarly, for SeqID TDPHYS 2, with mass 0.1, involving durations of unscheduled full outages due to software, there were order of magnitude excursions between \hat{e} and e ; the e values were not captured at either the 80% or 90% levels.

8.5 Scorecards for TD and EJE

This section provides compares the performance of specific methods of aggregation for EJE and TD within each data category. Table 20: Physical Data Scorecard and Table 21: Probabilistic Data Scorecard in Chapter 4 presented summary physical and probabilistic scorecards, respectively, for six methods against EJE data. The six methods were: Geometric Mean, Median, Alpha-Stable, ROT, Bayesian, and Classical. This last method was eliminated from the TD comparison of methods as it is

not applicable to TD. Three additional methods, Arithmetic Mean, Harmonic Mean and MLE were not included in Table 20 and Table 21 since coverage and bounds width were not computed for them, and they are excluded from the summary scorecards for TD as well. The EJE data were re-ranked for the five remaining methods in order to enable this comparison. The ABSDist (Coverage–0.9) metric used in these two tables for EJE data, is computed differently for TD data. This was necessitated by the fact that most themes in the TD set contain two variables at most. Accordingly, ABSDist (Coverage–0.9) for TD is computed by taking a weighted sum of indicator variables 1_i over all records in each TD category, where the indicator variable 1_i equals one if the 90% bounds contain e , and zero otherwise; the weights are the record weights.

Table 38: EJE Physical Data Aggregation Methods for Comparison with TD Physical Data and Table 39:TD Physical Data Aggregation Methods for Comparison with EJE Physical Data provide the comparison data for EJEPHYS and TDPHYS respectively. Note that Table 38 is based on the same 410 records for consistency with Table 20 and the physical TD set includes 11 records. The Average* MME is consistently higher for EJE than TD; both rank the Alpha-Stable and the Bayesian in the second and last ranks respectively. The median MME is consistently lower for EJE compared with TD; both rank the Geometric Mean in last place. The ABSDist (Coverage–0.9) ranks are identical for EJE and TD, with the Bayesian in first place. However, the range of the coverage metric for TD is double that of EJE. Similarly, the ranks for the median one-sided multiplicative bounds width are identical: the Bayesian is widest, and the Alpha stable most narrow. However, the Alpha stable has the poorest coverage, while the Bayesian has the coverage closest to 90%.

Table 38: EJE Physical Data Aggregation Methods for Comparison with TD Physical Data

MME Physical EJE	Geometric Mean	Median	Alpha-Stable	Rule of Thumb	Bayesian
Average*	4.14	4.08	4.13	4.23	4.26
Average* Rank	3	1	2	4	5
Median	1.444	1.4	1.398	1.398	1.44
Median Rank	5	3	1	1	4
ABSDist (Coverage=0.9)	0.24	0.11	0.36	0.33	0.095
Rank	3	2	5	4	1
Median One-sided Multiplicative Bounds Width	2.46	3.43	1.62	1.74	5.24
Rank	3	4	1	2	5
Complexity	1	1	2	2	3
Σ of ranks	15	11	11	13	18
Placement	4	1	1	3	5

Table 39: TD Physical Data Aggregation Methods for Comparison with EJE Physical Data

MME Physical TD	Geometric Mean	Median	Alpha-Stable	Rule of Thumb	Bayesian
Average*	3.38	3.58	3.52	3.52	3.59
Average* Rank	1	4	2	2	5
Median	2.5	2.27	2.41	2.44	2.36
Median Rank	5	1	3	4	2
ABSDist(Coverage=0.9), where coverage was weighted	0.50	0.40	0.70	0.60	0.17
Rank	3	2	5	4	1
Median One-sided Multiplicative Bounds Width	1.99	3.22	1.52	1.60	5.08
Rank	3	4	1	2	5
Complexity	1	1	2	2	3
Σ of ranks	13	12	13	14	16
Placement	2	1	2	4	5

Table 40: EJE Probabilistic Data Aggregation Methods for Comparison with TD Probabilistic Data and

Table 41: TD Probabilistic Data Aggregation Methods for Comparison with EJE
 Probabilistic **Data** provide the comparison data for EJEPROB and TDPROB respectively. Note that Table 40 is based on 66 records for consistency with Table 21: Probabilistic Data Scorecard, and the probabilistic TD set includes 26 records.

The Average* for EJE is consistently higher than its TD counterparts; the Geometric Mean ranked last in each data set with the Bayesian and Median taking first place in EJE and TD respectively. Similarly, the median MME is consistently higher for EJE than its TD counterparts; with respect to this metric, the Median is best, and the Geometric Mean worst in each data set. The ABSDist (Coverage=0.9) for EJE is consistently higher than its TD counterparts. However, this is simply an artifact of the TD probabilistic coverage being essentially 90% for all methods.

The Geomean, Alpha-Stable, and ROT tied for first place for this metric in the TD set. The Median One-sided Multiplicative Bounds Width for EJE is consistently higher than its TD counterparts; however, with the exception of the Geometric Mean and Median (which swap ranks), the ranks are identical. The ROT, Alpha-Stable and Bayesian are ranked first, second, and fifth in each data set against this metric.

Table 40: EJE Probabilistic Data Aggregation Methods for Comparison with TD Probabilistic Data

MME Probabilistic EJE	Geometric Mean	Median	Alpha-Stable	Rule of Thumb	Bayesian
Average*	20.78	8.42	17.69	17.5	7.96
Average* Rank	5	2	4	3	1
Median	2.018	1.63	1.86	1.86	1.78
Median Rank	5	1	3	3	2
ABSDist (Coverage=0.9)	0.32	0.11	0.38	0.45	0.09
Rank	3	2	4	5	1
Median One-sided Multiplicative Bounds Width	3.14	4.97	2.13	1.96	9.28

Rank	3	4	2	1	5
Complexity	1	1	2	2	3
Σ of ranks	17	10	15	14	12
Placement	5	1	4	3	2

Table 41: TD Probabilistic Data Aggregation Methods for Comparison with EJE Probabilistic Data

MME Probabilistic TD	Geometric Mean	Median	Alpha-Stable	Rule of Thumb	Bayesian
Average*	2.04	1.11	1.13	1.13	1.32
Average* Rank	5	1	2	2	4
Median	2.01	1.08	1.12	1.12	1.35
Median Rank	5	1	2	2	4
ABSDist(Coverage=0.9), where coverage was weighted	0.00	0.02	0.00	0.00	0.02
Rank	1	4	1	1	4
Median One-sided Multiplicative Bounds Width	2.15	2.04	1.58	1.49	3.47
Rank	4	3	2	1	5
Complexity	1	1	2	2	3
Σ of ranks	16	10	9	8	20
Placement	4	3	2	1	5

8.6 Summary

With regards to overall placement for physical data type, the Median ranked first for both TDPHYS and EJEPHYS (with respect to EJEPHYS, it tied with Alpha Stable for first place). The Bayesian ranked last in each physical data set. With regard to overall placement for probabilistic data, the Median and Bayesian ranked first and second, respectively for EJEPHYS. For TDPROB, the ROT and Alpha-Stable ranked first and second. However, it can be argued that for TDPROB, the ranks relating to ABSDist coverage should not be considered, due to the comparatively insignificant differences in coverage between methods. In this case, the sum of ranks for the Median

and the Bayesian drop by three, causing the Median to assume first place.

Chapter 9: Conclusions and Recommendations

Expert judgment continues to be used as input to decisions having large economic or even life and death impacts. Therefore, it is important to assess its reliability. Despite challenges documented in estimating probabilities, extensive literature contrasting the assessment of non-probabilistic, i.e., physical, variables and probabilistic variables in the context of meta-data based expert judgment aggregation techniques, and the errors associated with the predictions developed from such variables, was not identified. This research gap suggested a distinction be drawn between probabilistic and physical variables in order to separate the differences between observed and predicted values, i.e., multiplicative excursions, generated for each variable type in the meta-database used in this research. The distinction was applied to nine aggregation models for point estimates and/or bounds as well as sensitivity to the number of experts, and to level of realized value, via a set of five research questions.

The first research question asked if the type of quantity estimated, "physical"—variables having units of mass, time, etc.—or "probabilistic"—variables representing likelihood of an event, or a frequency of occurrence—impacts the accuracy of elicited predictions. Additionally, given a point estimate elicited from an expert, what multiplicative factors should be applied to that estimate, in order to bound it in an interval, with a corresponding level of probability?

This research found that the type of quantity estimated indeed impacts prediction accuracy, where the latter is measured by the ratio of the elicited median, e' , to the realized value, e . Since e values for probabilistic values are bounded by 0 and 1, the probabilistic meta data set was first compared to that portion of the physical meta data

whose e values were also bounded by 0 and 1; hereinafter physical data subset. It was shown that there is indeed a difference between these two data sets. For example, there was approximately twice the likelihood that an individual prediction e' overestimates the realized value of a probabilistic variable by a factor of ten, compared to a prediction for a physical variable. This difference persisted when the probabilistic data type was compared to the entire physical data.

Arguments were provided for using the ratio of realized value to predicted value (for TUD data, the predicted median value was used) as the measure of accuracy, as opposed to the difference between the two. The cumulative distribution function (CDF) of these ratios for probabilistic data was compared to the corresponding CDF for the subset of physical data for which $e < 1$. Both CDFs were constructed by applying weights to predictions e' so as to provide for equal weighting of all predictions associated with a given variable, all variables within a given theme, and all themes within a given subset of data.

The CDFs showed that probabilistic data is far more likely than physical subset data to be overestimated by large factors. For example, while rare, factor-of-one hundred overestimations occur approximately twenty times more often for probabilistic data than for physical subset data (2% vs. 0.1%). Similarly, for probabilistic data, there is roughly twice the likelihood of overestimating the realized value by a factor of ten or more. On the other hand, more modest overestimations, e.g. factor of two are actually somewhat more likely (30% vs. 23%) to occur for physical subset data.

Underestimations by factors of 2, 5, and 10 were at least 25% more likely to occur for probabilistic data; and at least twice as likely to occur at factors of 100 and 1000. It

was noted that although the CDFs were computed analytically, if the Kolmogorov-Smirnov two-sided non-parametric test for equality of distributions were applied, the null hypothesis H_0 that there is no difference between the distributions of the ratios e/e' for the two data categories would be rejected at $\alpha=0.05$. The null hypothesis would also have been rejected if the test had been applied to unweighted data, i.e. giving all ratios equal weights. Additionally, regarding the chance of overestimating the realized value by a factor of ten, the large sample test for equality of binomial proportion applied to the unweighted ratios would reject equality between probabilistic and physical subset data, at a level of significance $\alpha<<0.01$.

The disparity between the two data types persisted when the entire physical data set was compared to the probabilistic data set. Factor-of-ten overestimation remained statistically significantly more likely to occur for probabilistic data. Overestimation errors by factors of 2, 5, and 10 were approximately half as likely to occur for the full physical data set as for the physical data subset; corresponding underestimation probabilities were within 15% of the latter.

Over- and underestimation errors were consolidated into a derived metric, maximum multiplicative error, MME. This was defined as the maximum of $(r, 1/r)$, where r equals the ratio e/e' . The CDF of MME was computed, along with its complement. The latter gives the probability that MME exceeds a given value. For probabilistic data, there is a ten percent chance that a single prediction, e' will either overestimate or underestimate the realized value by a factor of forty or more; the analogous value for physical data is 7.3. In general, exceedance probabilities at factors of 2, 5, 10, 100 and 1000 for probabilistic data were roughly twice their physical data

counterparts. The exceedance probability was nearly 50% at a factor of 2, for probabilistic data; 30% for physical data. The respective probabilities declined to approximately 5% and 2%, at a factor of 100.

Given a single point estimate, e' , percentiles of the previously-developed CDFs for the appropriate data type were used to obtain intervals around e' , with associated levels of probability. In particular, the 10th and 90th percentiles of the CDF of e/e' for physical data, and the 5th and 95th percentiles of the same CDF, were used to obtain 80% and 90% bounding intervals for physical data e given e' : [0.340e', 3.45e'] and [0.165e', 9.84e'], respectively. An analogous procedure gave the corresponding 80% and 90% bounding intervals for probabilistic data e given e' : [0.098e', 7e'] and [0.025e', 36e'], respectively. The 90% intervals for probabilistic data are approximately five times wider on each side of e' multiplicatively, compared to physical data; the 80% intervals are approximately 2.5 times wider.

Future work could consider comparative accuracy when meta-data is further disaggregated by type. In particular, with respect to probabilistic data, it was noted that Hilbert (2011) proposed that there are two methods, (i) likelihood and probability and (ii) frequency for computing probability estimates. Analogously, this distinction may be considered non-frequentist and frequentist, as proposed by Kendall (1949), wherein the former expresses “a degree of rational belief” and the latter “defines probability in terms of frequencies of occurrence of events, or by relative proportions in 'populations' or 'collectives'" (p.101). One area for future research would be to compare predictions for Frequentist versus Non-Frequentist categories per Kendall's (1949) definition in terms of MME.

The second research question asked, given a set of point estimates elicited from experts, how should they be combined to yield an aggregate estimate and associated bounds? Six methods were considered for aggregating individual expert point estimates, e' into a single aggregated estimate of the median, \hat{e} , in conjunction with calibrated bounds around \hat{e} : The methods are:

- Geometric Mean; bounds computed via a Gaussian mixture approach using two normal distributions fitted to deweighted sets of (e, e') observations
- Median; bounds computed via quantile regression using a quadratic model in \ln -space to reflect the observed decrease in spread between quantiles for binned probabilistic data, as e increases
- Alpha stable distribution; parameters obtained via optimization in python to fit a power-law transformation of deweighted observations; bounds follow from the fact that sums of alpha stable are themselves alpha stable, with the same decay parameter
- A Rule of Thumb (ROT) method based on the fact that the alpha stable reduces to a normal distribution when its decay parameter approaches two; since the decay parameter for each data type was, in fact, either two or close to it, the power law transformation reduced to $|\ln(e/e')|^{0.5} \cdot \text{sign}(\ln(e/e')) \sim N(0, \sigma^2)$, where $\sigma=1.15$ and 0.83 for probabilistic and physical data, respectively; bounds were obtained accordingly
- A Bayesian method, incorporating as a likelihood function densities associated with quantile curves generated via quantile regression for e' given e ; and using a log-uniform model for e given location and spread, with location and spread distribution obtained via application of MathWave EasyFit to deweighted realized values and ratios of TUD 95th to 5th values, respectively.

- The Classical method of Cooke, for which 5th, 50th, and 95th aggregated estimates were provided after data cleanup.

The methods were ranked according to five criteria: accuracy, calibration, informativeness; sensitivity to outliers, and complexity. The method with the best performance against a given criterion received a rank of 1, ties excepted. Ranks were summed to determine the overall best methods against each data type. Accuracy criteria consisted of average and median MME, on a record-weighted basis. Because a single extremely large MME value could dominate the average (this occurred for the Classical method), the largest MME for each method was dropped before computing a reweighted average. Ranks with respect to accuracy were reported for three additional aggregation methods: Arithmetic Mean, Harmonic Mean, and Maximum Likelihood Estimate (MLE). (The MLE used the likelihood function incorporated in the Bayesian aggregation method.) These three generally placed among the three or four least inaccurate methods, against both physical and probabilistic data. (An exception was the Arithmetic Mean against probabilistic data, where it came in third place in weighted average MME, but 80% higher than the top-ranked method.) They were not further considered in terms of developing calibrated bounds.

As an example of the advantage in terms of increased accuracy conferred by aggregation, the chances of over- or underestimating e by various factors (2,5,10,100, and 1000) were computed before and after aggregation into \hat{e} via the Median. Aggregating reduced the combined chances by between 1.2 and 5.6, at each factor.

As previously stated, intervals or bounds around \hat{e} are required to give a sense of the uncertainty associated with estimate. The second type of ranking criterion for

comparing aggregation methods reflects calibration of the bounds. For nominal 90% bounds computed using a given aggregation techniques, the coverage of the bounds against the variables in each theme was computed. The absolute distance, ABSDist, between the coverage and 0.9 was averaged over all the themes associated with a given data set, and the resulting metric, called ABSDist(Coverage–0.9) is reported and ranked for each aggregation technique.

Since for the same level of calibration, narrower bounds are more informative than wider ones, the third criterion used was the one-sided multiplicative bounds width. This was given by the square root of the ratio of the 95th to 5th bounds points. Certain methods had a few extremely large widths, e.g., approximately 28,000 and 12,000 for the Classical. These could both reflect uncertainty around e , and dominate a weighted average of widths. For this reason, the median width (on a weighted basis) was used instead as the ranking criterion.

The fourth criterion, was sensitivity to outliers, where outliers were defined by comparing $e'_i - \bar{e}$ to the Median Absolute Deviation of these quantities, and \bar{e} is the median of the observations e'_i .

The final criterion was complexity, ranging from a value of 1 for least complex, i.e., methods provided as-is in office automation software (such as Arithmetic Mean), to 3 for most complex, where aggregation technique requires code leveraged from other techniques and/or additional development or stand-alone development beyond intermediate complexity level of 2. Bayesian and Classical methods were considered most complex.

Results were as follows: with respect to accuracy, the Alpha stable and the Median had the lowest sum of ranks for physical and probabilistic data sets, respectively. (With respect to average MME only, the Median and Bayesian had best performance, respectively.) The Classical method had worst performance on average MME only, but placed second and third best against physical and probabilistic data, respectively, with respect to median MME. The Geometric Mean had the highest rank for median MME.

With respect to calibration, for both data sets, the Bayesian and Classical had smallest and second smallest ABSDist(Coverage=0.9), respectively. (The Median was either tied with Classical or placed third against this criterion.) However, the Bayesian had approximately twice the median one-sided multiplicative bounds width of the Classical. The Alpha stable and closely related ROT had the narrowest bounds (half the median width associated with the Classical), but at the expense of poor coverage: roughly 50%. The Geometric Mean yielded widths greater than the Alpha stable but less than the Classical, with coverage that was still poor: approximately 70%. The ABSDist(Coverage=0.9) rankings for Alpha stable, ROT and Geometric Mean reflect these facts: they were 6, 5, and 4, respectively.

The Classical method was most sensitive to outliers in each data set; the Bayesian was least sensitive to probabilistic data outliers, but placed second worst for physical data outliers. The Geometric Mean and Alpha-Stable were among the top three ranked methods against this criterion.

After summing ranks, the Bayesian placed best for probabilistic data despite being tied with the Classical method for complexity, and the Alpha-Stable placed best for physical data. The Median and the Alpha -Stable were no worse than second or third

place overall. The Classical method placed last or next to last for both data sets. Note, however, that no single method ranked best against all criteria. Therefore, if a different weighting scheme had been applied to the criteria, as opposed to a sum of the ranks, a different result could have been obtained.

The third research question asked if the level of the quantity estimated mattered. How does the range of multiplicative bounding factor change between estimates of infrequent events, and estimates of frequent events? The quantile curves for e' given e , developed in connection with Bayesian aggregation, facilitated exploration of the impact of the magnitude of e on over- and underestimation chances. The magnitude of e was found to have a significant impact on these chances. For probabilistic data at $e=-15$, there is a spread of 15 units between the 5th and 95th quantiles of e' given e (both e and e' Ln-transformed). Approximately two thirds of the spread is above e ; one third below, at this point. This is consistent with the known tendency of elicitees to overestimate small values of e for probabilistic data. The spread narrows to approximately one unit at $e=-0.05$, with the bulk of the spread necessarily below e at this point. Narrower spreads are observed for physical data, but over a wider range of e values: nine units at $e=-18$ (with two thirds of the spread below the value of e); decreasing to four units, split approximately evenly above and below e at $e=2$; then increasing to six units at $e=22$, almost all of which is below e at this point. As stated in Chapter 5, the varying distances between quantiles impact the chances of over- or underestimating e by given factors.

Quantile regression was also employed to obtain analogous quantile curves for aggregated (via the Median) \hat{e} given e . For probabilistic data, the spreads between the 5th and 95th quantiles were smaller for \hat{e} at $e=-15$: about 9 units. Similarly, for physical

data, there was less curvature and smaller variation in the distance between 5th and 95th quantiles for aggregated \hat{e} .

The magnitude of e has a notable impact on multiplicative error for probabilistic data. For example, e' is approximately twice as likely to over- rather than underestimate e by factors of 2, 5, and 10 at $\ln(e) = -9$; and one and one half times as likely to over- rather than underestimate e by these same factors at $\ln(e) = -7$. The probabilities of over- or underestimating by factors of two, five, and ten become approximately equal at $\ln(e) = -2.2$, -3.3 , and -4.3 , respectively. Since e is bounded by one, underestimation becomes more likely than overestimation as e continues to increase; at $\ln(e) = -1$, underestimation by a factor of two is more than twice as likely to occur as overestimation.

For physical data, e' is approximately equally likely to over- as to underestimate e by factors of 2 and 5, over a broad region of $\ln(e)$ extending from approximately -4 to $+5$ (over half of the data points). For smaller e values, underestimation becomes more likely than overestimation: opposite the trend observed for probabilistic data. Underestimation probabilities also increase as $\ln(e)$ increases beyond 5; a factor-of-two underestimation is twice as likely to occur as similar overestimation at $\ln(e) = 14$.

Aggregation reduces the over-and underestimation chances at various levels of e . For example, for probabilistic data at $\ln(e) = -5$, aggregation cuts the likelihood of factor-of-ten overestimation in half (20% to 10%). Similarly, the likelihood of factor-of-ten underestimation is reduced 50% at this point. As before, underestimation becomes more likely than overestimation, as e approaches one.

An example of the improvement associated with aggregation of physical data is that at $\text{Ln}(e) = -4$, the chances of factor-of-five over- or underestimation before aggregation are approximately equal, at ten percent; whereas, after aggregation, these are approximately six and eight percent, respectively. In conclusion, the same overall trends in behavior of over- and underestimation error with magnitude of e are observed, but the effects are damped by aggregation.

Chapter 5 included a preliminary investigation of whether certain methods worked best in terms of having the greatest accuracy, over certain regions of values of e , as opposed to applying one method group-wide. Descriptive statistics (average, median and maximum) were reported for MMEs arising from each of the aggregation methods, categorized into bins according to $\text{Ln}(e)$. Six bins and 23 bins were used for probabilistic and physical data, respectively. The aim was to see if specific behavior of the descriptive statistics over subgroups would provide information which could otherwise be lost in a group-wide estimate.

For each statistic, and each aggregation method, the Spearman rank correlation coefficient, ρ of the statistic with the bin midpoint was computed. These coefficients were negative for physical data, and statistically significant ($p < 0.05$) for the median and average (except for Arithmetic Mean and Harmonic Mean with respect to the average.) Inferencing was limited for the probabilistic data, since it only had six bins. All coefficients were negative, except for average and maximum associated with the Harmonic Mean. This trend is consistent with poorer performance for small values of e .

In order to gain visual perspective on possible patterns, series of identical ranks within a given method of aggregation across three or more bins for the physical data,

were highlighted. Several of these streaks run for six bins, which appeared to be statistically significant ($p < 0.05$). However, the only streak involving a first place rank (1) occurred for both the average and maximum statistics for Bayesian aggregation over the bins ranging from [2.5,3] to [4.3333 to 4.6666], i.e., six bins. The corresponding values of e (not in Ln-space) run between approximately 12 and 106. Future research may wish explore this result in more detail.

In summary, as stated in that chapter, there is a general tendency for the representative MMEs for any given method of aggregation to decrease with increasing bin limits, consistent with the general trend of decreasing accuracy with decreasing numbers. The median statistic appears to be less subject to variation across the bins compared with the average and maximum statistics. No one method dominates the probabilistic bins, with respect to average MME. As discussed, for physical bins, with the exception of a streak of six bins in which the Bayesian had best performance, there is no consistent pattern to best performance of any particular aggregation method over bins.

The next research question, explored in Chapter 6, asked how the quality of the aggregated estimate varied with the number of experts used. Does a larger number of experts result in a tighter credible interval for the same level of probability. Two aggregation methods were used which were readily computable, Geometric Mean and Median. For a given number of experts, s ranging from 1 to 12, and for each record having $n \geq s$ observations e' , all possible $\binom{n}{s}$ distinct subsets of e' were drawn, without regard to order, and \hat{e} and associated MME computed for the subset. These MMEs were averaged for each record, after which the sum product using the (normalized) record weights was applied to obtain the global average MME at the given number of experts, s .

For both probabilistic and physical data, there is an order of magnitude reduction in average MME as s is increased from 1 to 2. For probabilistic data, a further order of magnitude reduction occurs between $s=2$ and $s=4$. A final order of magnitude improvement occurs by $s=7$, after which improvement essentially stops.

For physical data, a more moderate factor of two reduction is observed between $s=2$ and $s=4$; and a final factor of two reduction, to 6.7 is observed between $s=4$ and $s=5$. Further increases in s are associated with fluctuations in average MME.

The order of magnitude reductions in average MME associated with increasing the number of experts, s from 1 to 2 relied on the entire meta-data set for data of a given type. However, as noted in Chapter 6, had attention been confined to a particular subset of the data, gains may have been much more modest. As explained in Chapter 6, per Shanteau (2015), certain domains involving human behavior, may be dynamic and have poor expert judgment performance. The EJE INFOSEC (information security) theme seems to fit this description. By contrast, a physical system with stable stimuli was conducive to better performance. The pm25 theme, involving particulate emissions seemed to fit this description. For the latter theme, the ratio of maximum to minimum e' against each of its variables averaged less than 1.14. The average MME when only a single prediction from this theme was chosen, was 1.12. Increasing the number of experts to six decreased the average MME over this theme by one percent.

By contrast, for the newer, information security-related theme involving human behavior, the ratio of maximum to minimum e' against each of its probabilistic variables averaged 30,000. The corresponding average MME is 684 for a single expert; declining to 6.5 when the number of experts is increased to six. This means that increasing the

number of experts does not always have a big payoff; the applicable domain is significant.

The fifth and last research question, addressed in Chapter 7, asked if there are simple parametric distributions which perform reasonably well in yielding bounds around the estimate? The motivation for asking the question is such distributions, if they exist, might comprise an easy way to obtain confidence bounds, compared to more complex techniques involving Bayesian aggregation or quantile regression. Deweighted samples of $\ln(e/\hat{e})$ obtained using several simple aggregation methods, were run through MathWave EasyFit. The methods included Geometric Mean, Arithmetic Mean and Median.

The Cauchy distribution was the simple distribution providing the best fit for all three methods, for both physical and probabilistic data. Given the Cauchy's location and scale parameters, m and s , respectively, the quantile corresponding to the p th percentile point is given by $\hat{e} \cdot \exp(q)$, where $q = m + s \cdot \tan(\pi \cdot (p - 0.5))$. For aggregation via the median, $m=0$ and $s=0.22336$. Accordingly, 90% bounds are $(0.244\hat{e}, 4.097\hat{e})$. This can be approximated as $(\hat{e}/4.1, \hat{e} \cdot 4.1)$. The one-sided bounds width of 4.1 is approximately 20% larger than that shown in Table 12 for bounds around \hat{e}_{Median} obtained using quantile regression. ABSDist(Cov=0.9) was 0.106, close to its counterpart of 0.10 shown in Table 12. The 80% bounds are $(.503\hat{e}, 1.989\hat{e})$, which can be approximated as $(\hat{e}/2, \hat{e} \cdot 2)$. Coverage was 76%. Bounds factors for the Geometric Mean were less than two percent wider than their counterparts for the Median. For the Arithmetic Mean, however, the one-sided bounds width was significantly wider, at 5.4. Since the median is one of the top performing aggregation methods, the Cauchy-based simple parametric bounds of

$(\hat{e}/4.1, \hat{e}\cdot 4.1)$ and $(\hat{e}/2, \hat{e}\cdot 2)$ seems to represent a viable technique for obtaining 90% and 80% intervals around a point estimate \hat{e} obtained via the median.

For probabilistic data, the analogous Cauchy-based 90% bounds around \hat{e} obtained via the Median were $(0.089\hat{e}; 12.539\hat{e})$; the corresponding one-sided multiplicative bounds width is 11.9, which is more than twice the corresponding width of approximately 5.0 shown in Table 15 in Chapter 4 for 90% bounds around the median obtained via quantile regression. Coverage was 87%. Cauchy-based bounds around geometric and arithmetic means had significantly greater widths.

It appears viable to apply the Cauchy-based simple bounds technique around aggregated \hat{e}_{Median} point estimates. For probabilistic data, the median bounds width will be more than twice the value of its counterpart associated with quantile regression. Additionally, equal width bounds are always applied, irrespective of the aggregate point estimate \hat{e} and implied magnitude of e . This ignores the fact that for probabilistic data, there tends to be a wider spread around small \hat{e} values than around relatively large \hat{e} values approaching one.

In the scorecard table for aggregation methods against physical data, Table 20, the Alpha stable is ranked first. It had the tightest bounds width. However, it also had the worst coverage, as measure by the ABSDist (Coverage-0.9) criterion. By contrast, the Bayesian had the best coverage against this criterion, but the widest bounds. In the scorecard table, ranks were simply summed, and the method receiving the lowest sum was declared the best. However, if calibration had been given more weight than bounds width, the ranking of aggregation methods could have been different. In their analysis of alternative combinations of combined forecasts produced by European Central Bank

Survey of Professional Forecasters over a ten-year period, Genre, Kenney, Meyler, and Timmerman (2010) concluded “our results would argue in favour of reporting a suite of alternative combinations which forecast users could draw on taking into account the historical track record of individual combination methods and the prevailing economic context” (p. 35).

This research showed that no single aggregation method performed best against both EJE and test data. The harmonic mean, which had the worst performance in terms of average MME against both EJE probabilistic and physical data, had the best or second best performance against the corresponding test data sets.

The Alpha-Stable, ROT, and Bayesian aggregation techniques assumed the e' are independent of each other. However, the distributions of e' in some records can be bimodal, as shown in Appendix B: Shapiro-Wilk Analysis. For future research, the aggregation techniques used to obtain bounds could be refined to reflect this behavior, possibly through the incorporation of a copula function applied to the marginal densities per Sklar's theorem.

Finally, there is an advantage to having domain knowledge. As shown in Chapter 6 the applicable domain is significant, both in terms of numbers of accuracy as a function of number of experts, and bounds widths. Increasing the number of experts from one to six may buy a one percent reduction in MME for a relatively stable domain such as pm25; or it may increase accuracy by two orders of magnitude, for a dynamic, relatively poorly understood domain involving human behavior, such as information security (INFOSEC). Similarly, the Bayesian aggregation model used in this research assumed a prior for location based on all the realized values of a given data type in the meta-data

base; this represented a span of 40 units in Ln-space for physical data (-18 to +22). However, for one of physical data themes, involving predictions of stock prices, all values ranged between 1210 and 1340, and all realized values between 1163 and 1319. This represents an excursion of at most 0.15 in Ln-space. Future research could explore aggregation methods based on subsets of meta-data disaggregated by domain.

Appendix A: EJE and TD Referencing

The following listing provides the EJE Theme and corresponding TUD Case(s) where applicable, and the SeqID range by data category for the records in that theme. The data categories are further broken out by EJE and TD. The Reference column identifies the source of each theme's data.

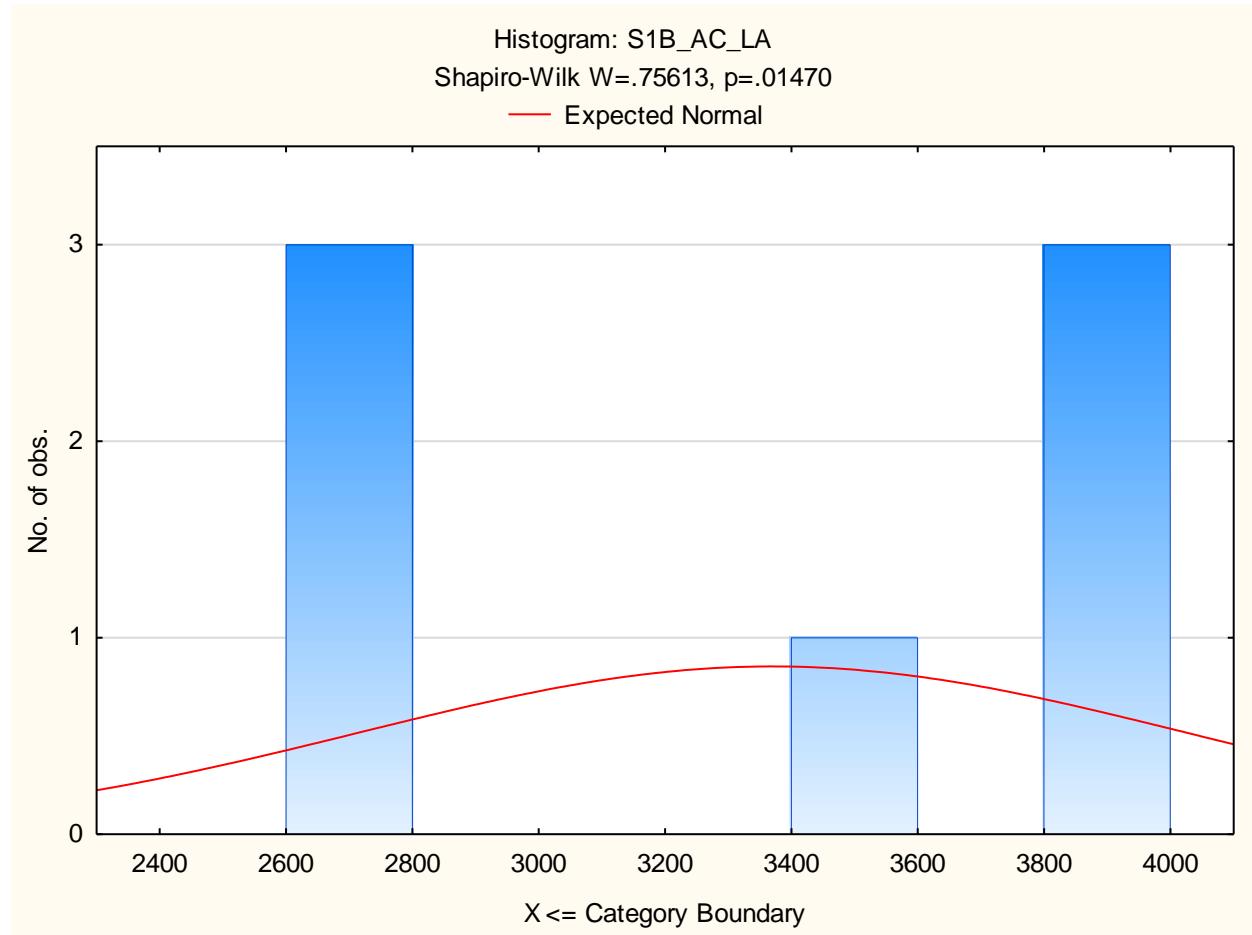
EJE Theme	TUD Case(s)	EJEPHYS	EJEPROB	TUDPHYS	TUDPROB	Reference
A_SEED	A_SEED	1-8				EXCALIBUR
ACNEXPTS	ACNEXPTS	9-18				EXCALIBUR
Agricultural Prices		19-70				McIntosh & Bessler (1988).
Airline Passengers		71-130				BaFail, A. O (2004).
AOTDAILY	AOTDAILY	131-168				EXCALIBUR
AOTRISK1	AOTRISK1	169-179				EXCALIBUR
Aviation	ATCEP Error 5 experts 31 jan 08; FCEP Error 5 experts 31 jan 08; pilots	180-194	1-12			EXCALIBUR
Benzene Concentration		195-197				Walker et al. (2003)
BSWAAL	BSWAAL	198-205				EXCALIBUR
CALCE Experts		206-210				UMD coursework (Prof. Mosleh)
Crop Yield		211-934				Dumler (2003).
dams	dams	935-938	13-19			EXCALIBUR
DEPOS1	DEPOS1	939-962				EXCALIBUR
DISPER1 (amalgamated with TNODISP1)	DISPER1; TNODISP1	963-1034				EXCALIBUR
Electricity Peak Profile		1035-1118				Cather & Thompson. (2005).
EXP_DISP	EXP_DISP	1119-1141				EXCALIBUR
EXP_WD	EXP_WD	1142-1160				EXCALIBUR
GL-invasive-species	GL-invasive-species	1161-1173				EXCALIBUR
Greece_NL_CARMA	CARMAExpStudy; CARMA-Greece-Assessments	1174-1189	20-23			EXCALIBUR
GROND5	GROND5	1190-1199				EXCALIBUR
Gross Product		1200-1231				Tennessee Valley Authority (2003)
Hotel Occupancy		1232-1260				Schwartz & Cohen (2004).
INFOSEC	INFOSEC	1261-1264	24-27			EXCALIBUR

EJE Theme	TUD Case(s)	EJEPHYS	EJEPROB	TUDPHYS	TUDPROB	Reference
Labor Force		1265-1280				Stekler & Thomas (2005).
Ladders	Ladders	1281-1284	28-33			EXCALIBUR
Land Use Planning		1285-1294				Rodier. (2005).
New Product Shares		1295-1338				Ehrman & Shugan (1995)
NH3EXPTS	NH3EXPTS	1339-1348				EXCALIBUR
Nickel Species		1349-1352				Ramachandran et al. (2003).
ONINX	ONINX	1353-1399				EXCALIBUR
OpRiskBank	OpRiskBank	1400-1415				EXCALIBUR
PBEARLYH	PBEARLYH	1416-1430				EXCALIBUR
PBINTDOS	PBINTDOS	1431-1485				EXCALIBUR
PhD Surveys 2005		1486-1491				Forrester (2005)
pm25	pm25	1492-1497	34-39			EXCALIBUR
RETURNafter	RETURNafter	1498-1528				EXCALIBUR
S_SEED	S_SEED	1529-1559				EXCALIBUR
SE Asia Population Cohorts		1560-1607				Kahn (2003).
SO3EXPTS	SO3EXPTS	1608-1617				EXCALIBUR
Space Flight Risk	ESTEC1; ESTEC-2; ESTEC-3	1618-1643	40-63			EXCALIBUR
UMD Campus		1644-1655	64			UMD coursework (Prof. Mosleh)
US Population		1656-1707				Campbell (2002).
Volcanoes	MONT1; Volcrisk	1708-1721	65-67			EXCALIBUR
	Elections 2012			1	1	The Washington Post (2012)
	Unscheduled outages			2-3	23; 26	FAA Survey (FOIA-ble)
	Ongoing software development effort (hours)			4		Jørgensen & Moløkken (2002).
	Software growth reliability model			5		Almering, van Genuchten, Cloudt, & Sonnemans, (2007).
	Software Hours Bid Phase			6-11		Faria & Miranda (2012).
	Forecaster Reliability				2-22	Murphy & Winkler (1977).
	Antiretroviral therapy				24	Zazzi et al. (2011).
	ASDE-X				25	FAA (FOIA-ble)

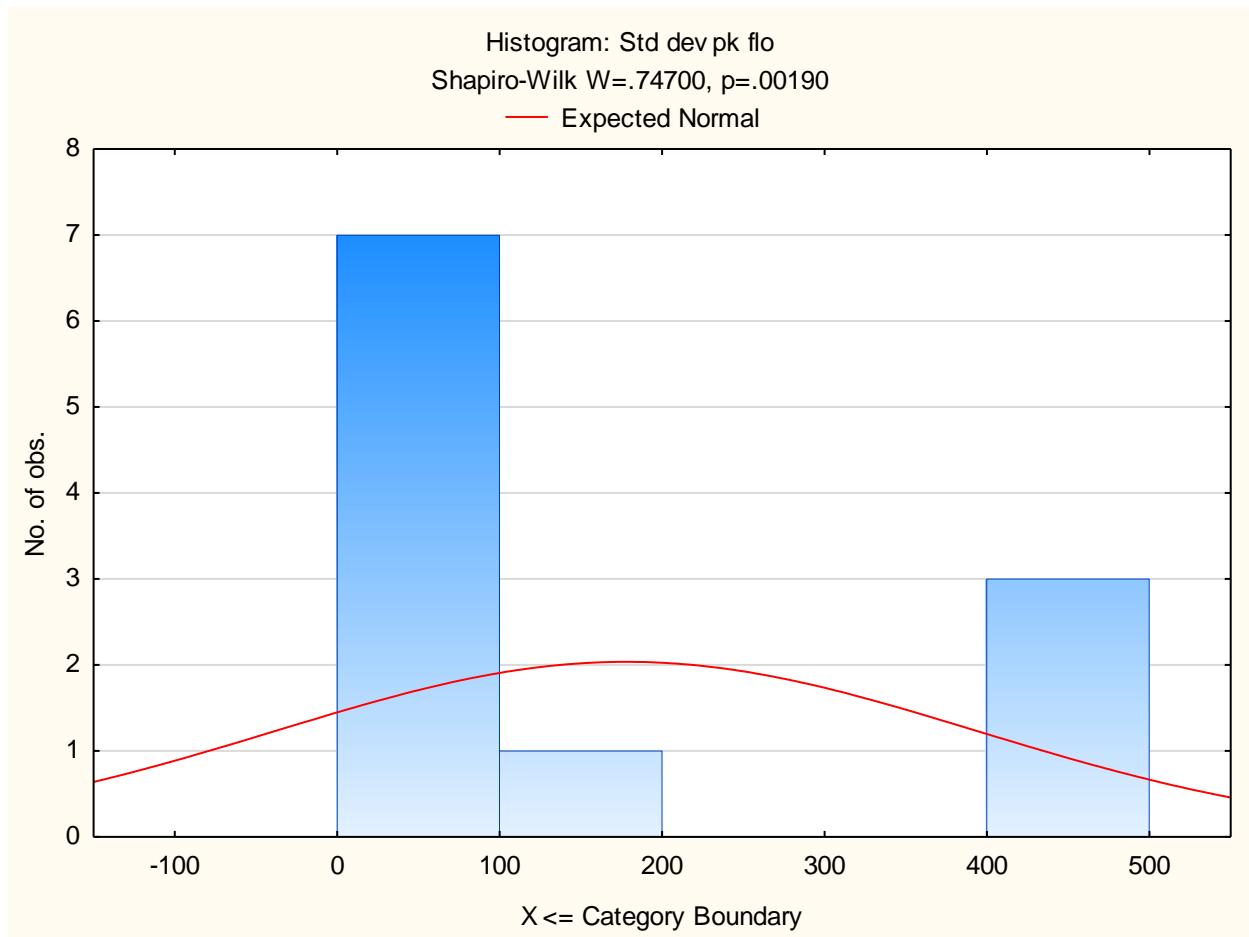
Appendix B: Shapiro-Wilk Analysis

Physical Data

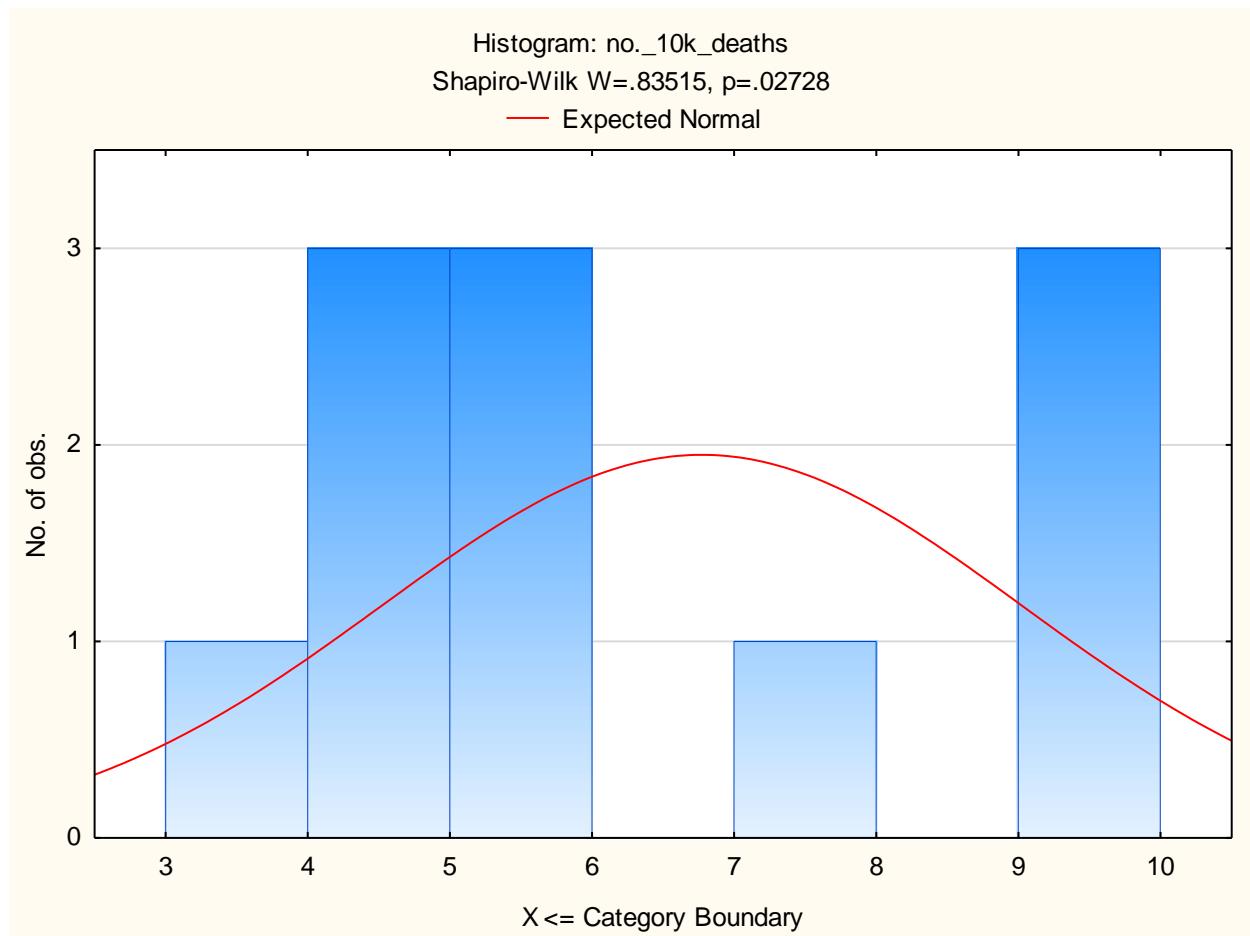
SeqID EJEPHYS2; n=7; Theme=A_SEED



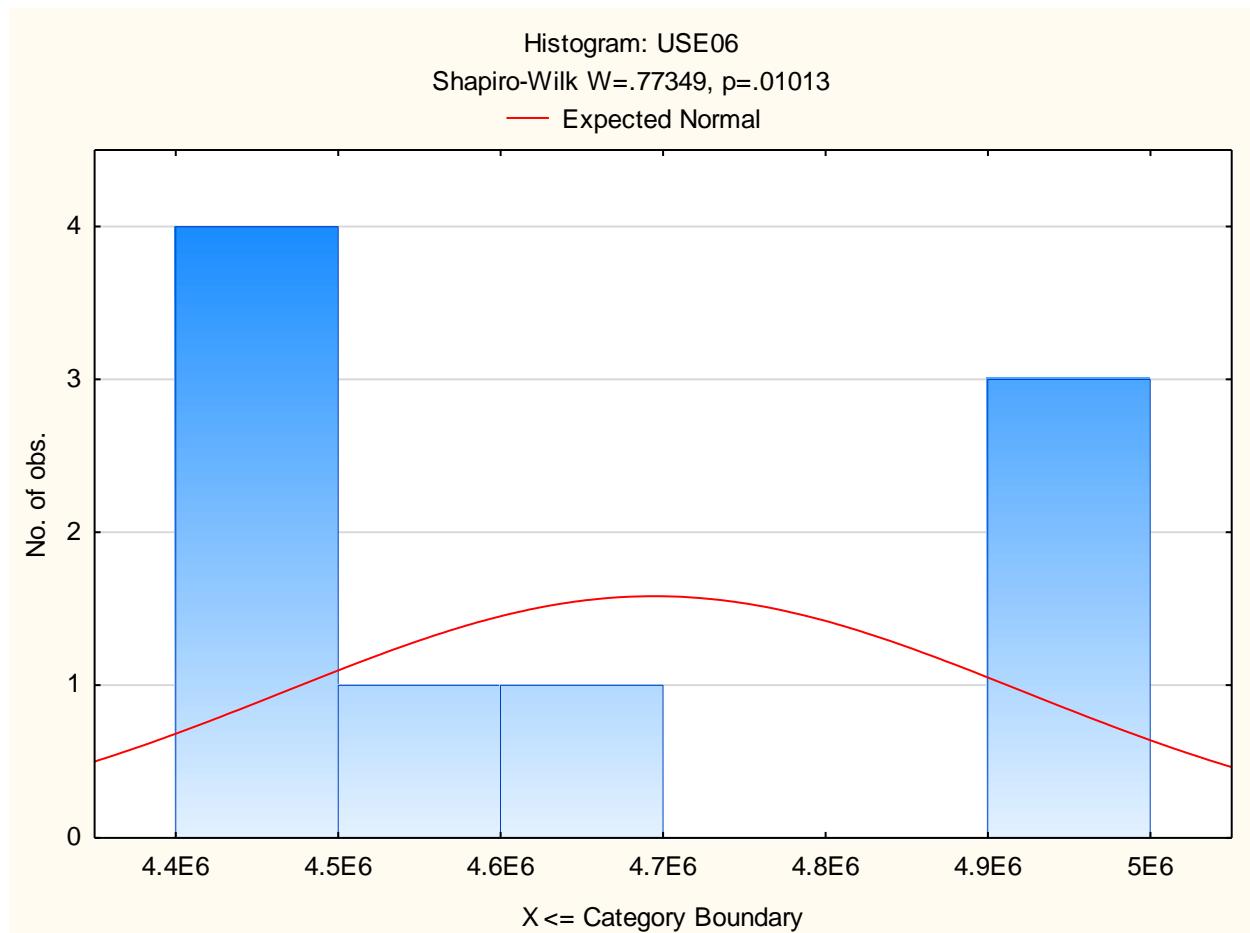
SeqID EJEPHY963; n = 11; Theme: DISPER1 (amalgamated with TNODISP1)



SeqID EJEPHY1716; n=11; Theme: Volcanoes

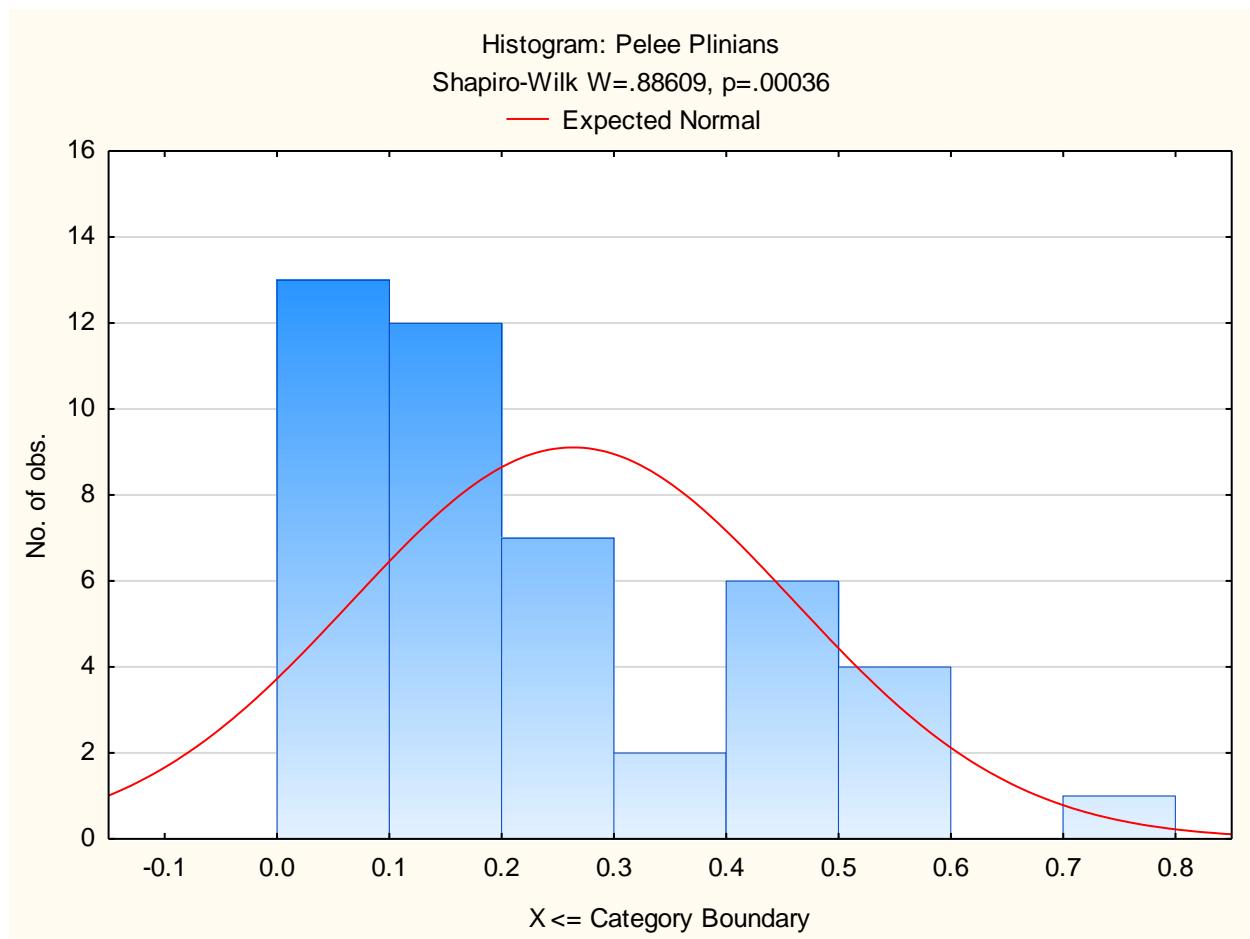


SeqID EJEPHY1168; n=9; Theme: GL-invasive-species

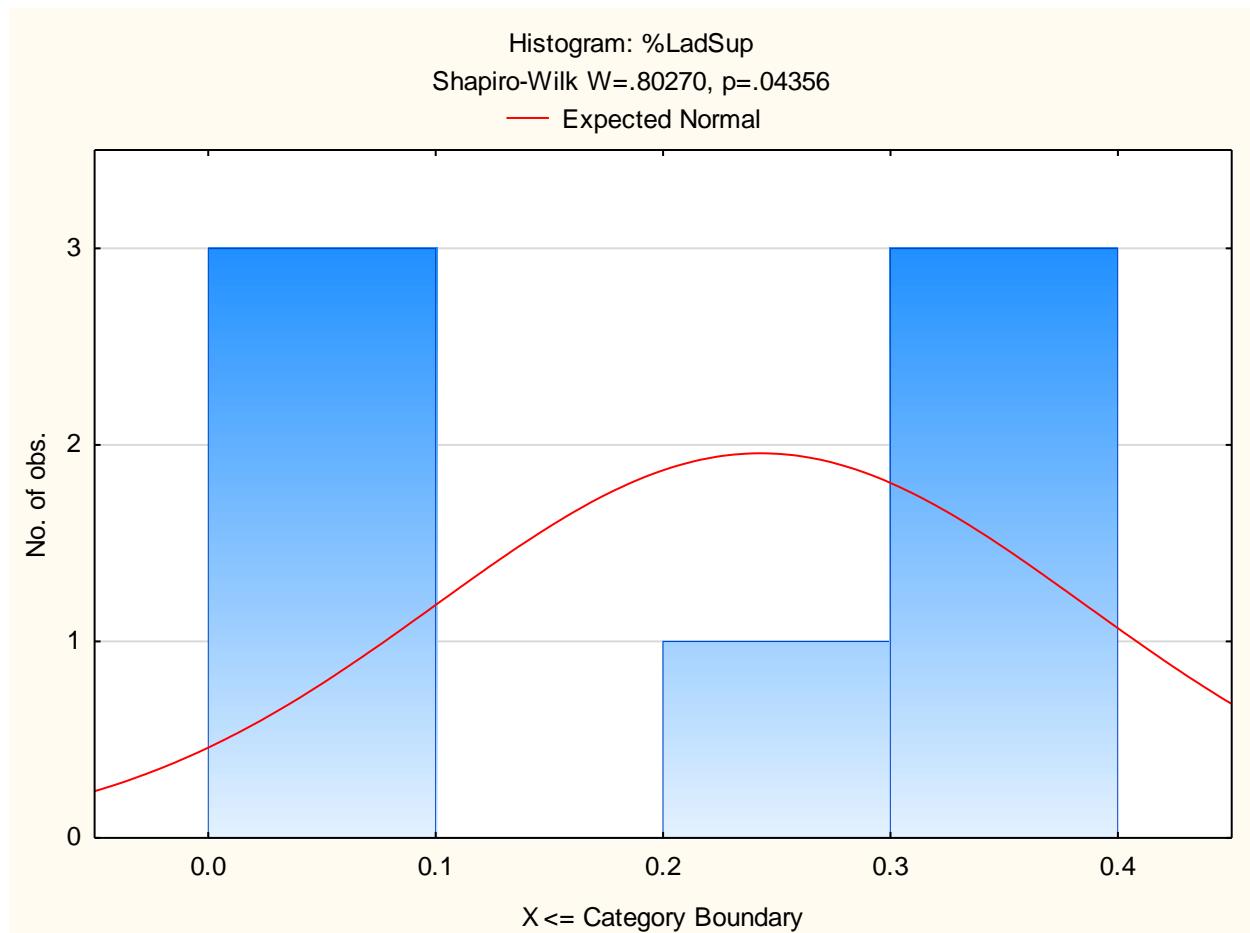


Probabilistic

SeqID EJEPROB67; n=45; Theme: Volcanoes



SeqID EJEPROB1; n=7; Theme: Ladders



Appendix C: Sample of MathWave EasyFit Outputs

The following three tables were generated by MathWave EasyFit when used to fit a location parameter for physical data from a deweighted sample of 123,765 records.

Table B-1: Descriptive Statistics

Statistic	Value	Percentile	Value
Sample Size	123765	Min	-17.14
Range	38.275	5%	-6.008
Mean	3.5894	10%	-2.659
Variance	31.477	25% (Q1)	0.754
Std. Deviation	5.6105	50% (Median)	3.781
Coef. of Variation	1.5631	75% (Q3)	5.824
Std. Error	0.01595	90%	10.786
Skewness	-0.0457	95%	14.743
Excess Kurtosis	1.2345	Max	21.135

Table B-2: Goodness of Fit Summary

#	Distribution	Kolmogorov		Anderson		Chi-Squared	
		Smirnov		Darling			
		Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.10234	21	1921.9	16	32432	11
2	Burr (4P)	0.07082	8	874.6	5	24586	5
3	Cauchy	0.05055	2	635.9	2	28666	10
4	Chi-Squared (2P)	0.15653	28	4397.9	25	62964	24
5	Dagum (4P)	0.06797	4	925.79	6	25927	9
6	Erlang (3P)	0.0959	14	1925.1	17	35632	18
7	Error	0.07023	6	765.22	4	19251	2
8	Error Function	0.32156	34	26617	33	130190	27
9	Exponential (2P)	0.41443	36	31663	34	336290	30
10	Fatigue Life (3P)	0.10037	17	1904.9	13	32801	13
11	Frechet (3P)	0.21803	31	11971	29	1.6524E+26	36
12	Gamma (3P)	0.09471	13	1983.1	19	35129	16
13	Gen. Extreme Value	0.09275	12	1960.2	18	63057	25
14	Gen. Gamma (4P)	0.10084	18	1913.6	14	33610	15
15	Gen. Logistic	0.0738	9	964.31	8	24933	6
16	Gen. Pareto	0.12738	25	40576	35	N/A	
17	Gumbel Max	0.11397	22	3944	24	2.9742E+16	35

#	Distribution	Kolmogorov		Anderson		Chi-Squared	
		Smirnov		Darling			
		Statistic	Rank	Statistic	Rank	Statistic	Rank
18	Gumbel Min	0.1475	27	5673	26	69831000	34
19	Hypersecant	0.06855	5	702.92	3	20548	3
20	Inv. Gaussian (3P)	0.10192	20	1904.4	12	32916	14
21	Johnson SU	0.08697	11	1203.5	10	25502	7
22	Kumaraswamy	0.11865	24	2630	22	35913	19
23	Laplace	0.04068	1	316.45	1	16427	1
24	Levy (2P)	0.52353	37	42952	37	589690	32
25	Log-Logistic (3P)	0.07072	7	939.25	7	25705	8
26	Logistic	0.08289	10	1098.8	9	24137	4
27	Lognormal (3P)	0.09682	16	1878.8	11	35148	17
28	Normal	0.10157	19	1915.2	15	32514	12
29	Pearson 5 (3P)	0.09616	15	2004.7	20	38752	21
30	Pearson 6 (4P)	0.23085	32	14467	30	138530	28
31	Pert	0.15955	29	5878.3	27	57307	22
32	Phased Bi-Exponential	0.53097	38	47572	38	545190	31
33	Power Function	0.32281	35	19213	32	160640	29
34	Rayleigh (2P)	0.2985	33	15806	31	115690	26
35	Student's t	0.57657	39	128320	39	1236500	33
36	Triangular	0.18701	30	6832.8	28	57409	23
37	Uniform	0.14281	26	41194	36	N/A	
38	Wakeby	0.05343	3	3084.3	23	N/A	
39	Weibull (3P)	0.11785	23	2616.6	21	35932	20
40	Burr	No fit (data min < 0)					
41	Chi-Squared	No fit (data min < 0)					
42	Dagum	No fit (data min < 0)					
43	Erlang	No fit (data min < 0)					
44	Exponential	No fit (data min < 0)					
45	Fatigue Life	No fit (data min < 0)					
46	Frechet	No fit (data min < 0)					
47	Gamma	No fit (data min < 0)					
48	Gen. Gamma	No fit (data min < 0)					
49	Inv. Gaussian	No fit (data min < 0)					
50	Johnson SB	No fit					
51	Levy	No fit (data min < 0)					
52	Log-Gamma	No fit					
53	Log-Logistic	No fit (data min < 0)					
54	Log-Pearson 3	No fit					

#	Distribution	Kolmogorov		Anderson		Chi-Squared	
		Smirnov		Darling			
		Statistic	Rank	Statistic	Rank	Statistic	Rank
55	Lognormal	No fit (data min < 0)					
56	Nakagami	No fit					
57	Pareto	No fit					
58	Pareto 2	No fit					
59	Pearson 5	No fit (data min < 0)					
60	Pearson 6	No fit (data min < 0)					
61	Phased Bi-Weibull	No fit					
62	Rayleigh	No fit (data min < 0)					
63	Reciprocal	No fit					
64	Rice	No fit					
65	Weibull	No fit (data min < 0)					

Table B-3: Fitting Results

#	Distribution	Parameters
1	Beta	$\alpha_1=4194.6 \quad \alpha_2=3085.2$
		$a=-554.72 \quad b=414.25$
2	Burr (4P)	$k=0.84193 \quad \alpha=3.2240E+6$
		$\beta=8.9589E+6 \quad \gamma=-8.9589E+6$
3	Cauchy	$\sigma=2.4738 \quad \mu=3.5003$
4	Chi-Squared (2P)	$v=24 \quad \gamma=-20.52$
5	Dagum (4P)	$k=0.83378 \quad \alpha=35.183$
		$\beta=97.898 \quad \gamma=-93.67$
6	Erlang (3P)	$m=228 \quad \beta=0.37455 \quad \gamma=-81.809$
7	Error	$k=1.3296 \quad \sigma=5.6105 \quad \mu=3.5894$
8	Error Function	$h=0.12603$
9	Exponential (2P)	$\lambda=0.04824 \quad \gamma=-17.14$
10	Fatigue Life (3P)	$\alpha=0.00704 \quad \beta=796.64 \quad \gamma=-793.08$
11	Frechet (3P)	$\alpha=2.5753 \quad \beta=18.053 \quad \gamma=-18.552$
12	Gamma (3P)	$\alpha=207.46 \quad \beta=0.39559 \quad \gamma=-78.524$
13	Gen. Extreme Value	$k=-0.25788 \quad \sigma=5.2292 \quad \mu=1.6592$
14	Gen. Gamma (4P)	$k=1.7884 \quad \alpha=106.21$
		$\beta=7.6507 \quad \gamma=-100.16$
15	Gen. Logistic	$k=0.01451 \quad \sigma=3.0022 \quad \mu=3.5177$
16	Gen. Pareto	$k=-0.94277 \quad \sigma=17.17 \quad \mu=-5.2484$
17	Gumbel Max	$\sigma=4.3745 \quad \mu=1.0643$

#	Distribution	Parameters
18	Gumbel Min	$\sigma=4.3745 \mu=6.1144$
19	Hypersecant	$\sigma=5.6105 \mu=3.5894$
20	Inv. Gaussian (3P)	$\lambda=6.5286E+6 \mu=590.18 \gamma=-586.59$
21	Johnson SU	$\gamma=0.05642 \delta=2.1485$
		$\lambda=10.77 \xi=3.9046$
22	Kumaraswamy	$\alpha_1=4.4604 \alpha_2=702.76$
		$a=-19.671 b=90.805$
23	Laplace	$\lambda=0.25207 \mu=3.5894$
24	Levy (2P)	$\sigma=18.191 \gamma=-17.567$
25	Log-Logistic (3P)	$\alpha=157.26 \beta=464.47 \gamma=-460.95$
26	Logistic	$\sigma=3.0932 \mu=3.5894$
27	Lognormal (3P)	$\sigma=0.03589 \mu=5.0514 \gamma=-152.74$
28	Normal	$\sigma=5.6105 \mu=3.5894$
29	Pearson 5 (3P)	$\alpha=342.02 \beta=36137.0 \gamma=-102.34$
30	Pearson 6 (4P)	$\alpha_1=3.0474 \alpha_2=2.4928E+8$
		$\beta=1.9295E+9 \gamma=-17.248$
31	Pert	$m=4.2567 a=-17.366 b=21.369$
32	Phased Bi-Exponential	$\lambda_1=0.01782 \gamma_1=-18$
		$\lambda_2=0.47383 \gamma_2=20.351$
33	Power Function	$\alpha=1.5093 a=-17.183 b=21.135$
34	Rayleigh (2P)	$\sigma=15.205 \gamma=-17.168$
35	Student's t	$v=2$
36	Triangular	$m=4.094 a=-17.203 b=21.275$
37	Uniform	$a=-6.1283 b=13.307$
38	Wakeby	$\alpha=180.79 \beta=13.235 \gamma=4.6985$
		$\delta=-0.05423 \xi=-13.568$
39	Weibull (3P)	$\alpha=4.4805 \beta=25.457 \gamma=-19.733$
40	Burr	No fit (data min < 0)
41	Chi-Squared	No fit (data min < 0)
42	Dagum	No fit (data min < 0)
43	Erlang	No fit (data min < 0)
44	Exponential	No fit (data min < 0)
45	Fatigue Life	No fit (data min < 0)
46	Frechet	No fit (data min < 0)
47	Gamma	No fit (data min < 0)
48	Gen. Gamma	No fit (data min < 0)
49	Inv. Gaussian	No fit (data min < 0)
50	Johnson SB	No fit

#	Distribution	Parameters
51	Levy	No fit (data min < 0)
52	Log-Gamma	No fit
53	Log-Logistic	No fit (data min < 0)
54	Log-Pearson 3	No fit
55	Lognormal	No fit (data min < 0)
56	Nakagami	No fit
57	Pareto	No fit
58	Pareto 2	No fit
59	Pearson 5	No fit (data min < 0)
60	Pearson 6	No fit (data min < 0)
61	Phased Bi-Weibull	No fit
62	Rayleigh	No fit (data min < 0)
63	Reciprocal	No fit
64	Rice	No fit
65	Weibull	No fit (data min < 0)

Bibliography

- Abrahamson, N. A., Coppersmith, K. J., Koller, M., Roth, P., Sprecher, C., Toro, G. R., & Youngs, R. (2004) *Probabilistic seismic hazard analysis for Swiss nuclear power plant sites*. (PMT-SB-0001). Retrieved on November 30, 2014 from <http://www.swissnuclear.ch/upload/cms/user/PEGASOSProjectReportVolume1-new.pdf>.
- Aitken, C., Roberts, P., & Jackson, G. (2010). *Fundamentals of probability and statistical evidence in criminal proceedings: guidance for judges, lawyers, forensic scientists and expert witnesses*. Manual. Royal Statistical Society, London.
- Retrieved on March 27, 2015 from <http://www.rss.org.uk/site/cms/contentviewarticle.asp?article=1132>.
- Almering, V., van Genuchten, M., Cloudt, G., & Sonnemans, P. J. M. (2007). Using software reliability growth models in practice. *IEEE Software*, 24 (6) 82-88. doi:10.1109/MS.2007.182.
- Ashton, A. H. & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, 31, 1499-1508.
- Aspinall, W. (2008). Briefing. *Expert judgment elicitation using the classical model and EXCALIBUR* (Briefing). Retrieved from <http://dutiosc.twi.tudelft.nl/~risk/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf>.
- Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, 463, 294-295. doi:10.1038/463294a.

- Aspinall, W. & Cooke, R.M. (1998) Expert judgement and the Montserrat Volcano eruption. *Proceedings of the 4th international conference on Probabilistic Safety Assessment and Management PSAM4*. Vol 3, pp. 13-18.
- BaFail, A. O (2004). Applying data mining techniques to forecast number of airline passengers in Saudi Arabia (domestic and international travels). Retrieved from <https://trid.trb.org/view.aspx?id=748691>.
- Barberis, N. (2013). The Psychology of tail events: Progress and challenges. *American Economic Review: Papers & Proceedings*, 103(3), 611–616. doi.doi.org/10.1257/aer.103.3.611.
- Batanero, C. & Díaz, C. (2012). Training school teachers to teach probability: reflections and challenges. *Chilean Journal of Statistics*, 1, 3-13.
- Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. *School Science and Mathematics*, 108(8), 355–361.
- Bjerstedt, T.W. (1996, January). *Principles and guidelines for formal use of expert judgment by Yucca Mountain site characterization project*. Presentation presented at the Nuclear Technical Waste Review Board of the US Department of Energy, Office of Civilian Radioactive Waste Management, La Vegas, NV. Retrieved on November 26, 2014 from <http://www.nwtrb.gov/meetings/1996/jan/bjerstedt.pdf>.
- Board of Governors of the Federal Reserve System. (2013). *Capital planning at large bank holding companies: Supervisory expectations and range of current practice*. Retrieved from <http://www.federalreserve.gov/bankinforeg/stress-tests/ccar/August-2013-Estimation-Methodologies-for-Losses-Revenues-and-Expenses.htm#subsection-1934-08640AF5>.

- Bolger, F. & Rowe, W. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, 35(1), 5-11. doi: 10.1111/risa.12272
- Boone, I., Van der Stede, Y., Bollaerts, K., Messens, W., Vose, D., Daube, G., Aerts, M., & Mintiens, K. (2009). Expert judgement in a risk assessment model for *Salmonella* spp. in pork: The performance of different weighting schemes. *Preventive Veterinary Medicine*, 92, 224–234. doi:10.1016/j.prevetmed.2009.08.020.
- Borak, S., Härdle, W., & Weron, R. (2005). Stable distributions. In P. Cizek; W. Härdle, & R. Weron. (Eds), *Statistical tools for finance and insurance* (pp. 21-44). doi:10.1007/b139025.
- Brown, B. & Helmer, O. (1964). Improving the reliability of estimates obtained from a consensus of experts (RAND Corporation Report P2986). Retrieved from <http://www.rand.org/content/dam/rand/pubs/papers/2008/P2986.pdf>.
- Camerer, C., & Johnson, E. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In A. Ericsson and J. Smith Ed(s), *Toward a general theory of expertise: Prospects and limits*. New York: Cambridge University Press.
- California Department of Transportation (2007). Division of Design Cost Estimating. Retrieved from http://www.dot.ca.gov/hq/oppd/pdpm/chap_pdf/chapt20.pdf.
- 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). Retrieved from <https://www.census.gov/population/www/documentation/twps0057/twps0057.pdf>

Carmines and Zeller (1979). *Reliability and validity assessment*. Sage University Papers.

Retrieved from http://www.uky.edu/~clthyn2/PS671/carmines_zeller_671.pdf

Cathers, C. A. & Thompson, G. D. (2005). Forecasting short-term electricity load profiles

(Cardon Research Papers). Retrieved from

<https://ag.arizona.edu/arec/sites/cals.arizona.edu.arec/files/publications/2006-07cathersthompson.pdf>

Chernick, M. R. (2011). *The essentials of biostatistics for physicians, nurses, and*

clinicians. USA, NJ. Wiley. Retrieved from

<http://sgh.org.sa/Portals/0/Articles/The%20Essentials%20of%20Biostatistics%20or%20Physicians,%20Nurses,%20and%20Clinicians.pdf> John Wiley & Sons, Inc.

Published 2011.

Clemen, R.T., & Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research*, 33(2), 427-442.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5, 559-583.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2), 187-203.

Clemen, R. T. (2008) Comment on Cooke's Classical Method. *Reliability Engineering & System Safety*, 93(5), 760-765. doi:10.1016/j.ress.2007.03.005.

Cooke, R.M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

Cooke R. M. & Slijkhuis, K. A. (2003). Expert judgment in the uncertainty analysis of dike ring failure frequency. In: Blischke, W. R., & Prabhakar, M. (Eds). *Case*

- studies in reliability and maintenance* (pp. 331–352). doi: 10.1002/0471393002.ch15
- Cooke, R. M., & Probst, K. N. (2006). Highlights of the expert judgment policy symposium and technical workshop” Retrieved on October 19, 2014 from <http://www.rff.org/Documents/Conference-Summary.pdf>.
- Cooke, R. M., & Goossens L. L. H. J. (2008). TU Delft expert judgment data base. *Reliability Engineering and System Safety*, 93(5), 657–674. doi:10.1016/j.ress.2007.03.005.
- Cooke, R. M., & Kelly, G. N. (2010). *Climate change uncertainty quantification: Lessons learned from the Joint EU-USNRC project on uncertainty analysis of probabilistic accident consequence codes*. Retrieved on October 19, 2014 from <http://www.rff.org/RFF/Documents/RFF-DP-10-29.pdf> RFF DP 10-29 May 2010.
- Critical Values for the Two-sample Kolmogorov-Smirnov test (2-sided) (n.d.). Retrieved from https://www.webdepot.umontreal.ca/Usagers/angers/MonDepotPublic/STT3500H10/Critical_KS.pdf.
- Curtright, A., Morgan, G. M., & Keith, D. W. (2008). Expert assessments of future photovoltaic technologies. *Environmental Science & Technology*, 42(24), 9031-9038. doi: 10.1021/es8014088.
- Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi Method to the use of experts. *Management Science*, 9(3), 458-467. Retrieved from <http://socsci2.ucsd.edu/~aronatas/project/academic/delphi%20method%20of%20convergence.pdf>.

- Davidson, V. J., Ravel, A., Nguyen, T. N., Fazil, A., & Ruzante, J. M. (2011). Food-specific attribution of selected gastrointestinal illnesses: estimates from a Canadian expert elicitation survey. *Foodborne Pathogens and Disease*, 8(9), 983-995. doi:10.1089/fpd.2010.0786.
- Devilee, J. L. A., & Knol, A. B. (2011). *Software to support expert elicitation: An exploratory study of existing software packages* (RIVM Letter report 630003001/2011) The National Institute for Public Health and the Environment (RIVM). Retrieved on November 30, 2014 from http://www.rivm.nl/en/Documents_and_publications/Scientific/Reports/2012/mei/Software_to_support_expert_elicitation_An_exploratory_study_of_existing_software_packages.
- Dumler, T. J. (2003). Rainfall and Farm Income. *Proceedings of Risk and Profit Conference, 2003, Manhattan, Kansas, August 14-15, 2003 Kansas State University Agricultural Experiment Station and Cooperative Extension Service*. Retrieved from <http://www.agecon.ksu.edu/home/Research&Extension/risk%20and%20profit/papers.htm>
- Edwards, W. (1975). Cognitive processes and the assessment of subjective probability distributions: Comment. *Journal of the American Statistical Association*, 70(350), 290-293.
- Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2012). A Performance-based statistical expert judgment model to assess technical performance and risk. *INCOSE*

International Symposium, 22: 2101–2112. doi: 10.1002/j.2334-5837.2012.tb01460.x.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke's classical model. *Reliability Engineering and System Safety*, 121(2014), 72–82.
<http://dx.doi.org/10.1016/j.ress.2013.07.015>.

Ehrman, C. M. & Shugan, S. M. (1995). The forecaster's dilemma. *Marketing Science*, 14(2) 123-127. Retrieved from <http://dx.doi.org/10.1287/mksc.14.2.123>

European Food Safety Authority (2014). *Guidance on expert knowledge elicitation in food and feed safety risk assessment*. EFSA Journal 2014;12(6):3734. Retrieved on November 30, 2014 from

<http://www.efsa.europa.eu/en/efsajournal/doc/3734.pdf>.

European Food Safety Authority (n.d.). Scientific panel renewal. Retrieved on December 7, 2014 from

http://www.efsa.europa.eu/en/corporate/doc/factsheetpanelsrenewal_en.pdf.

Fama, E. F. (1965). The behavior of stock market prices. *Journal of Business*, 38(1), 34-105.

Faria, P. & Miranda, E. (2012, Assissi 17-19 Oct.) Expert Judgment in Software Estimation During the Bid Phase of a Project -- An Exploratory Survey Software Measurement and the 2012 Seventh International Conference on Software Process and *Product Measurement (IWSM-MENSURA)*, 2012 Joint Conference of the 22nd International Workshop, pp. 126-131. doi: 2012 10.1109/IWSM-MENSURA.2012.27.

Federal Elections Commission. (2013). *Federal Elections 2012. Election Results for the U.S. President, the U.S. Senate and the U.S. House of Representatives.*

<http://www.fec.gov/pubrec/fe2012/federalelections2012.pdf>

Forrester, Y. (2005). *The quality of expert judgment: An interdisciplinary investigation* (Doctoral Dissertation). Retrieved from <http://hdl.handle.net/1903/3267>.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, 19 (1), 44-63.

Gayer, T. (2015, March). *Energy efficiency, risk and uncertainty, and behavioral public choice* (Institute for Humane Studies/Mercatus Keynote address presented at the Brookings Institute). Retrieved on March 29, 2015 from
<http://www.brookings.edu/research/speeches/2015/03/06-energy-efficiency-public-choice-gayer>.

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1), 114-148, 1986.

Genre, V., Kenney, G., Meyler, A., & Timmerman, A. (2010). *Combining the forecasts in the ECB survey of professional forecasters. Can anything beat the simple average?* European Central Bank Working Paper Series Number 1277.
<https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1277.pdf?d26def670a20ade78c896d61dc2b5ff7>

Glaser, M., Langer, T. & Weber, M. (2012). True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *Journal of Behavioral Decision Making*, 26(5), 405-417.

Gould, J. E. (2002). *Concise handbook of experimental methods for the behavioral and biological sciences*. Boca Raton, FL: CRC Press.

Hammit J. K., & Zhang, T. (2013). Combining Experts' Judgments: Comparison of Algorithmic Methods using Synthetic Data. *Risk Analysis*, 33(1), 109–120. doi: 10.1111/j.1539-6924.2012.01833.x

Helmer O., & Rescher N. (1959). On the epistemology of the inexact science (RAND Report R-353). Retrieved from
<http://www.rand.org/content/dam/rand/pubs/reports/2006/R353.pdf>.

Hendry, D.F., & Clements, M. P. (2002). Pooling of forecasts. *Econometrics Journal*, 5, 1–26. Retrieved from
<http://www.nuff.ox.ac.uk/Users/Hendry/Papers/DFHMPCFrcncEctJ.pdf>.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin* 138(2), 211-237. doi: 10.1037/a0025940.<http://dx.doi.org/10.1037/a0025940>.

Hoek, G., Boogaard, H., Knol, A., de Hartog, J. K., Slottje, P., Ayres, J.G., Borm, B., Brunekreef, B., Donaldson , D., Forastiere, F., Holgate, S., Kreyling, W.K., Nemery, B., Pekkanen, J., Stone, V., Wichmann, H. E., & van der Sluijs, J. (2010). Concentration response functions for ultrafine particles and all-cause mortality and hospital admissions: Results of a European expert panel elicitation. *Environmental Science & Technology*, 44(1), 476-482. doi: 10.1021/es9021393.

Hogarth, R. M. (1975a). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, 70(350), 271-289.

Hogarth, R. M. (1975b). Cognitive processes and the assessment of subjective probability distributions: Rejoinder. *Journal of the American Statistical Association*, 70(350), 294.

Hora, S.C. (2004). Probability judgments for continuous quantities: linear combinations and calibration. *Management Science*, 50, 597-604.

Hora, S. & Jensen, M. (2002). Expert judgement elicitation. The Swedish Radiation Protection Authority (SSI Report: 2002:19). Retrieved on October 19, 2014 from <http://www.stralsakerhetsmyndigheten.se/Global/Publikationer/Rapport/Stralskydd/2002/ssi-rapp-2002-19.pdf>.

Hora, S. C. & Jensen, M. (2005). Expert panel elicitation of seismicity following glaciation in Sweden. Swedish Radiation Protection Authority (SSI Report 2005-20). Retrieved on November 30, 2014 from <http://www.stralsakerhetsmyndigheten.se/Publikationer/Rapport/Stralskydd/2005/200520>

Huer, R. J. (1981). Strategic deception and counterdeception: A cognitive process approach. *International Studies Quarterly*, 25(2), 294-327.

Israelski, E. W. (2010). Basic Human Attributes. In Gardner-Bonneau (Eds). *Handbook of human factors in medical device design*. CRC Press Taylor & Francis Group. Retrieved from

https://books.google.com/books?id=jAemLm2zu_oC&pg=PA53&lpg=PA53&dq=overestimation+of+low+probabilities&source=bl&ots=NP2i1aci3Z&sig=LEb5myu_o3UVUNukIiVvSAZLqX8&hl=en&sa=X&ei=I1QYVZauH4KhgwT87oCY

Aw&ved=0CCUQ6AEwAzgK#v=onepage&q=overestimation%20of%20low%20 probabilities&f=false

Johnston, R. (2003) Reducing analytic error: Integrating methodologists into teams of substantive experts. *Studies in Intelligence*, 47 (1). Retrieved on October 19, 2014 from <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol47no1/article06.html>.

Jørgensen, M. & Moløkken, K. (2002). Combination of software development effort prediction intervals: why, when and how? *SEKE '02 Proceedings of the 14th international conference on Software engineering and knowledge engineering*, pp. 425-428. doi:10.1145/568760.568833

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Keith, D.W. (1996). When is it appropriate to combine expert judgments? An editorial essay. *Climatic Change*, 33, 139-143.

Kendall, G. M. (1949). On the reconciliation of theories of probability. *Biometrika*, 36, (1/2), 101-116. <http://www.jstor.org/stable/2332534>.

Khan, H.T.A. (2003) A comparative analysis of the accuracy of the United Nations' population projections for six Southeast Asian countries (IIASA Interim Report IR-03-015). Retrieved from <http://pure.iiasa.ac.at/7066/>.

Knol, A. B., de Hartog, J. J., Boogaard, H., Slottje, P., van der Sluijs, J. P., Lebret, E. Cassee, F. R., Wardekker, J. A., Ayres, J. G., Borm, J. P., Brunekreef, B., Donaldson, K., Forastiere, F., Holgate, S.T., Kreyling, W.G., Nemery, B., Pekkanen, J. V., Stone, V., Wichmann, H. E., & Hoek, G. (2009). Expert

- elicitation on ultrafine particles: likelihood of health effects and causal pathways.
Particle and Fibre Toxicology, 6(19). doi:10.1186/1743-8977-6-19.
- Knol, A.B., Slottje, P., van der Sluijs, J. P., & Lebret, E. (2010). The use of expert elicitation in environmental health impact assessment: a seven step procedure.
Environmental Health, 9(19). doi:10.1186/1476-069X-9-19.
- Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15 (4), 143–156. <http://digitalcommons.ilr.cornell.edu/hrpubs/19/>.
- Lee, J. C., & McCormick, N. J. (2012) Risk and safety analysis of nuclear systems. Wiley & Sons, Hoboken: NJ. doi: 10.1002/9781118043462.ch1
- Lewis, H. L., Budnitz, H. J., Couts, C., von Hippel, F., Lowenstein, W. B., & F. Zachariasen, F. (1978). Risk Assessment Review Group Report to the U. S. Nuclear Regulatory Commission (NUREG/CR-0400).
- Library of Congress, Congressional Research Service (2010), *The Nunn-McCurdy Act: Background, analysis, and issues for Congress*, (Report R41293). Retrieved from <http://www.fas.org/sgp/crs/misc/R41293.pdf>
- Lichtendahl, K. C., Grushka-Cockayne, Y., & Winkler, R. L., (2013) Is it better to average probabilities or quantiles? *Management Science*, 59(7) 1594–1611.
<http://dx.doi.org/10.1287/mnsc.1120.1667> ©2013 INFORMS
- Lin, S-W. (2011). Jackknife evaluation of uncertainty judgments aggregated by the Kullback–Leibler distance. *Applied Mathematics and Computation*, 218, 469–479. doi:10.1016/j.amc.2011.05.087.
- Lin, S-W., & Bier V. M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety*, 93(5), 711–721. doi:10.1016/j.ress.2007.03.005.

Lin, S-W, & Cheng C-H. (2008) Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities? *IEEE International Conference on Industrial Engineering and Engineering Management* (pp.425 – 429). doi: 10.1109/IEEM.2008.4737904.

McIntosh, C. S. & Bessler, D. A. (1988). Forecasting agricultural prices using a Bayesian composite approach. *Southern Journal of Agricultural Economics. December*, 73-80. Retrieved from <http://ageconsearch.umn.edu/bitstream/29269/1/20020073.pdf>

Mahmood, A., Chitre, M., & Armand, M. A. (2012). Improving PSK performance in snapping shrimp noise with rotated constellations. *Proceedings of the Seventh ACM International Conference on Underwater Networks and Systems*. Article Number 12. doi: 10.1145/2398936.2398952.

Martz, H. F., Bryson, M. C., & Waller, R. A. (1985). Eliciting and aggregating subjective judgements – some experimental results. Los Alamos National Laboratory LU-UR—84-3193.

Meyer, M. A. & Booker, J. M. (2001). Eliciting and analyzing expert judgment: a practical guide. *American Statistical Association and the Society for Industrial and Applied Mathematics*, Alexandria, Virginia, USA. Retrieved on November 30, 2014 from http://books.google.com/books?hl=en&lr=&id=ZLt_y-patXYC&oi=fnd&pg=PR2&dq=Meyer+MA+and+Booker+JM,+2001.+Eliciting+and+analyzing+expert+judgment&ots=clvta_OkAQ&sig=cOagmRrq8gQyO1yLu_ybNdXiEqjg#v=onepage&q=payment&f=false.

Morris, P.A. (1977). Combining expert judgments: A Bayesian approach. *Management Science* 20, 679-693. doi: <http://dx.doi.org/10.1287/mnsc.23.7.679>

- Mosleh, A., & Apostolakis, G. (1986). The assessment of probability distributions from expert opinions with an application to seismic fragility curves. *Risk Analysis*, 6(4), 447–461. doi: 10.1111/j.1539-6924.1986.tb00957.x
- Mumpower, J. L., & Stewart, T. R. (1996). Expert judgement and expert disagreement. *Thinking and Reasoning*, 2(2/3), 191-211.
- Murphy, A. H., & Winkler, R. L (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? Retrieved from <http://nwas.org/digest/papers/1977/Vol02No2/1977v002no02-MurphyWinkler.pdf>
- Murphy, A. H. (1993). What is a good forecast? An essay of the nature of goodness in weather forecasting. *Weather Forecasting*, 8, 281–293. doi: [http://dx.doi.org/10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](http://dx.doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2)
- National Research Council (1994). Science and Judgment in Risk Assessment. Retrieved from <http://www.nap.edu/openbook.php?isbn=030904894X>.
- Nicholls, N., (1999). Cognitive illusions, heuristics, and climate prediction. *Bulletin of the American Meteorological Society*, 80 (7), 1385-1398.
- Nolan, J. (2009). Stable distributions models for heavy tailed data. Retrieved from <http://academic2.american.edu/~jpnolan/stable/chap1.pdf> on March 29,2014.
- Nunes, A. & Kirlik, A. (2005). An Empirical Study of calibration in air traffic control expert judgment. Proceedings of the human factors and ergonomics society. 49th Annual Meeting (422-426). Retrieved on March 27, 2015 from http://www.aviation.illinois.edu/avimain/papers/research/pub_pdfs/hfes/nunkir.pdf.

O'Hagan, A. Buck, C.E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J.,
Oakley, J. E. & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Wiley.

Ouchi, F. (2004). A Literature review on the use of expert opinion in probabilistic risk analysis (World Bank Policy Research Working Paper 320). Retrieved from
<http://econ.worldbank.org>.

Ramachandran G., Banerjee S., & Vincent J. H. (2003). Expert judgment and occupational hygiene: Application to aerosol speciation in the nickel primary production industry. *The Annals of Occupational Hygiene*, 47(6) 461–475.
Retrieved from doi: 10.1093/annhyg/meg066.

Ranjan, R., & Gneiting, T. (2009). Combining probability forecasts. *Journal of the Royal Statistical Society B*, 72, Part 1, 71–91.

Rimmer, R. H., & Nolan, J. (2005). Stable distributions in Mathematica. *The Mathematica Journal*, 9(4), 776-789.

Rodier, C. J. (2005). Verifying the accuracy of land use model used in transportation and air quality planning: A case study in Sacramento, California region (MTI Report 05-02). Retrieved from

<http://transweb.sjsu.edu/MTIportal/research/publications/summary/0502.html>

Roman, H. A., Hammitt, J. K., Walsh, T. L., & Stieb, D.M. (2012). Expert elicitation of the value per statistical life in an air pollution context. *Risk Analysis*, 32(12), 2133-2151. doi: 10.1111/j.1539-6924.2012.01826.x.

Rowe, G. & Wright, G. (2001). Expert Opinions in Forecasting: The Role of the Delphi Technique. In J. Scott Armstrong (Ed.) *Principles of forecasting: A handbook for*

- researchers and practitioners* http://link.springer.com/chapter/10.1007%2F978-0-306-47630-3_7.
- Sackman, H. (1975). Summary evaluation of Delphi. *Policy Analysis*, 1(4), 693–718.
Retrieved from <http://www.jstor.org/stable/42784280>
- Salvendy, G. (ed.) (2012) *Handbook of human factors and ergonomics*, 4th Edition Wiley & Sons, Hoboken: NJ.
- Schwartz, Z. & Cohen, E. (2004). Hotel Revenue management forecasting - Evidence of expert-judgment bias. Retrieved from
<http://cqx.sagepub.com/content/45/1/85.full.pdf>
- Seaver, D. A. (1976). Assessment of group preferences and group uncertainty for decision making. (Defense Technical Information Center Technical Report. Jul 75-Sep 76). Retrieved from <http://www.dtic.mil/get-tr-doc/pdf?AD=ADA033246>.
- Shanteau, J. (2015). Why task domains (still) matter for understanding expertise. *Journal of Applied Research in Memory and Cognition*, 4 (2015) 169–175. Retrieved from <http://dx.doi.org/10.1016/j.jarmac.2015.07.003>.
- Shirazi, C. H. (2009). *Data-informed calibration and aggregation of expert judgment in a Bayesian framework* (Doctoral Dissertation). Retrieved from
<http://hdl.handle.net/1903/9883>.
- Slottje, P., Sluijs, J. B., & Knol, A. B. (2008). Expert Elicitation: Methodological suggestions for its use in environmental health impact assessments (RIVM letter report 630004001/2008.) Retrieved on October 1, 2013 from
http://www.nusap.net/downloads/reports/Expert_Elicitation.pdf.
- Slovic, P. (1987). Perception of risk. *Science*, 236, 280–285.

Snedecor, G. W. & Cochran, W.G. (1978). *Statistical methods*. Iowa State University Press; 6th edition (1978).

Sniezek, J. A., Schrah, G. A., & Dalal, S. R. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making, 17*, 173–190. doi: 10.1002/bdm.468

Stekler, H. O. & Thomas, R. (2005). Evaluating BLS labor force, employment and occupation projection for 2000. Retrieved from
<http://www.bls.gov/opub/mlr/2005/07/art5full.pdf>

Sweidan, M., Williamson, M., Reeve, J. F., Harvey, K., O'Neill, J. A., Schattner, P., & Snowdon, T. (2010). Identification of features of electronic prescribing systems to support quality and safety in primary care using a modified Delphi process. *BioMed Central Medical Informatics and Decision Making, 10*(21). doi:10.1186/1472-6947-10-21.

Tennessee Valley Authority (2003). Methodology and results from socioeconomic modeling. Final environmental assessment – Appendix B. Retrieved from
http://www.tva.gov/environment/reports/rates/appendix_b.pdf.

The Washington Post (2012). Crystal Ball Contest (November 3, 2012).
<http://www.washingtonpost.com/wp-srv/special/opinions/outlook-crystal-ball-contest/>;

Tian, Y., Huffman, G.J., Adler, R. F. Tang, L., Sapiano, M., Maggioni, V., & Wu, H. (2013). Modeling errors in daily precipitation measurements: Additive or multiplicative? *Geophysical Research Letters, 40*, 2060–2065. doi:10.1002/grl.50320, 2013

Tversky, A. & Kahneman D. (1974). Judgment under Uncertainty: Heuristics and Biases.

Science, 185(4157), 1124-1131. <http://www.jstor.org/stable/1738360>.

Tversky, A. & Kahneman D. (1982). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

United Kingdom Department for Environment, Food and Rural Affairs. (2005). Social Cost of carbon: A closer look at uncertainty (Final project report). Retrieved on November 30, 2014 from

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/243814/sei-scc-report.pdf

U.S. Department of Agriculture (1999). Price forecasting applications economic research service determination for corn and wheat. (Technical Bulletin No 1878). Retrieved from <http://www.ers.usda.gov/media/1761730/tb1878h.pdf>

U.S. Department of Agriculture, Food Safety and Inspection Service. (2007). Results of an additional expert elicitation on the relative risks of meat and poultry products (Draft Report Contract No. 53-3A94-03-12, Task Order 27). Retrieved on November 29, 2014 from http://www.fsis.usda.gov/wps/wcm/connect/2d081ff1-cdc3-4975-94d4-948930b6e141/RBI_Elicitation_Report.pdf?MOD=AJPERES.

U.S. Department of Agriculture, Food Safety and Inspection Service. (2012). Expert elicitation on the market shares for raw meat and poultry products containing added solutions and mechanically tenderized raw meat and poultry products. Retrieved on November 29, 2014 from

http://www.fsis.usda.gov/wps/wcm/connect/3a97f0b5-b523-4225-8387-c56a1eeee189/Market_Shares_MTB_0212.pdf?MOD=AJPERES

U.S. Department of Agriculture (n.d.) Explanatory Notes for Stochastic Budget Outlay Estimates.

https://www.fsa.usda.gov/Internet/FSA_File/pb09_stochastic_explanation.pdf

U.S. Department of Commerce National Institute of Standards and Technology. (2010). Engineering Statistics Handbook, Retrieved from
<http://www.itl.nist.gov/div898/handbook/>.

U.S. Department of Commerce National Institute of Standards and Technology. (2012). Latent print examination and human factors: Improving the practice through a systems approach: the report of the expert working group on human factors in latent print analysis. Retrieved on March 27, 2015 from
<http://www.nist.gov/oles/upload/latent.pdf>.

U.S. Environmental Protection Agency. (2001). RAGS Volume 3 Part A ~Process For Conducting Probabilistic Risk Assessment Appendix G. Retrieved on March 20, 2015 from <http://www.epa.gov/oswer/riskassessment/rags3adt/pdf/appendixg.pdf>.

U.S. Environmental Protection Agency (2011). Expert elicitation task force white paper. Retrieved on November 29, 2014 from <http://www.epa.gov/stpc/pdfs/ee-white-paper-final.pdf>.

U.S. Environmental Protection Agency (n.d.). Expert elicitation of the deep uncertainty surrounding the market and non-market damages of climate change. Retrieved on December 7, 2014 from

[http://yosemite.epa.gov/EE/epa/eed.nsf/41565cd88a5ab1a3852575a6006ab35e/7a57118f091e961985257d3c000162b7.](http://yosemite.epa.gov/EE/epa/eed.nsf/41565cd88a5ab1a3852575a6006ab35e/7a57118f091e961985257d3c000162b7)

United States General Accountability Office (1997) *Air traffic control. Improved cost information needed to make billion dollar investment decisions* (GAO/AIMD-97-2). <http://www.gao.gov/assets/160/155713.pdf>

United States General Accountability Office (2005). *Further refinements needed to assess risks and prioritize protective measures at ports and other critical infrastructure* (GAO-06-91). <http://www.gao.gov/assets/160/157672.pdf>

United States General Accountability Office (2008a). *FAA has increased efforts to curb runway incursions* (GAO-08-1169T). <http://www.gao.gov/assets/130/121346.pdf>

United States General Accountability Office (2008b). *2010 Census, Census Bureau should take action to improve the credibility and accuracy of its cost estimate for the decennial census* (GAO-08-554). <http://www.gao.gov/assets/280/276782.pdf>.

United States General Accountability Office. (2009). *GAO Cost Estimating and Assessment Guide. Best practices for developing and managing capital program costs* (GAO-09-3SP).

United States General Accountability Office (2012). *Cost estimating practices have improved, and continued evaluation will determine effectiveness* (GAO -15-210R). <http://www.gao.gov/assets/670/667509.pdf> GAO-15-210R Missile Defense

United States General Accountability Office (2014). *NRC needs to improve its cost estimates by incorporating more best practices* (GAO-15-98). Retrieved on March 20, 2015 from <http://www.gao.gov/assets/670/667501.pdf>

Usher, W. & Strachan, N. (2013). An expert elicitation of climate, energy and economic uncertainties. *Energy Policy*, 61(2013), 811–821.

<http://dx.doi.org/10.1016/j.enpol.2013.06.110>.

Van der Fels-Klerx, H. J., Cooke, R. M., Maarten, M. N., Goossens, L. H., & Havelaar, H. A. (2005). A structured expert judgment study for a model of campylobacter transmission during broiler-chicken processing. *Risk Analysis*, 25(1), 10-124. doi: 10.1111/j.0272-4332.2005.00571.x

Walker, K. D., Evans, J. S., & Macintosh, D. (2001). Use of expert judgment in exposure assessment Part I. Characterization of personal exposure to benzene. *Journal of Exposure Analysis and Environmental Epidemiology*, 11(4), 308 – 322. PMID:11571610.

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) 1-16. Retrieved from
<http://dx.doi.org/10.1038/sj.jea.7500253>.

Weinstein, B. D. (1993). What is an expert? *Theoretical Medicine*, 14(1), 57-73. PMID: 8506540.

Winkler, R. L. (1968). The consensus of subjective probability distributions. *Management Science*, 15(2), B61-B75. Retrieved on October 1, 2013 from
<http://www.jstor.org/stable/2628853>.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, 66(336), 675-685.

<http://www.jstor.org/stable/2284212>.

Winkler, R. L. (1975). Cognitive processes and the assessment of subjective probability distributions: Comment. *Journal of the American Statistical Association*, 70(350), 290-291.

Winkler, R. L., & Clemen, R. T. (2004) Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, 1(3), 167–176. doi 10.1287/deca.1030.0008.

Xiaohu, G., Guangxi, Z., & Yaoting, Z. (2004). On the testing for Alpha-Stable distributions of network traffic. *Computer Communications*, 27(5), 447-457.

Retrieved from: <http://dx.doi.org/10.1016/j.comcom.2003.10.004>

Zazzi, M, Kaiser, R, Sonnerborg, A, Struck, D, Altmann, A, Prosperi, M, Rosen-Zvi, M, Petroczi, A, Peres, Y, Schulter, E, Boucher, C A, Brun-Vezinet, F, Harrigan, P R, Morris, L, Obermeier, M, Perno, C F, Phanuphak, P, Pillay, D, Shafer, R W, Vandamme, A M, vanLaethem, K, Wensing, A M J, Lengauer, T & Incardona, F. (2011). Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study). *HIV Medicine*, 12(4) 211-218. doi: 10.1111/j.1468-1293.2010.00871.x

Zientek, L. R., Carter, T. A., Taylor, J. M., & Capraro, R. M. (2011). Preparing prospective teachers: An examination of attitudes toward statistics. *The Journal of Mathematical Sciences and Mathematics Education*, 5, 25-38.