A META-DATA INFORMED COMPARISON OF EXPERT JUDGMENT RELIABILITY

Ali Mosleh[1], Ellis S. Feldman[2]

[1] *University of Maryland, College Park, MD, mosleh@umd.edu*
[2] *University of Maryland, College Park, MD, efeld@umd.edu*

In situations where it is impracticable to calculate a quantity analytically, estimation via elicitation of subject matter experts, SMEs, may be used. Since such SME subjective judgments may have life and death, economic, or political impacts, their assessment in terms of reliability needs to be scrutinized. This assessment was examined via a set of research questions, applied to a set of meta-data containing elicited predictions and realized values. Two of the research questions are: 1) 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. An example of 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. If we disaggregate the meta-data into physical and probabilistic-related subsets, does the range of multiplicative factor bounding the estimate change? 2) 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?

The Figure of Merit (FOM) was the multiplicative error e/e', between prediction, e' and realized value, e. The meta-data was grouped into sets of related variables called themes. Weights were applied to each (e,e') "observation" in order to equalize the total weight applied to each theme, to each variable in a theme, and to each observation within a variable. Comparison of the empirical CDFs incorporating these weights, and corresponding to the two data types, showed that the type of quantity estimated does make a difference, with the probabilistic predictions showing greater multiplicative error. Additionally, quantile regression incorporating the weights was applied to fit a polynomial predicting e' given each e, to the meta-data. This facilitated exploration of the impact of level of e, on the extent of multiplicative error. For physical data, over values of e representing the 75th to the 5th percentile, the likelihood of a factor-of-two error ranged from approximately 30% to 50%, split approximately equally between over- and under-estimation. At the 90th percentile value of e, under- rather than overestimation was fifty percent more likely to occur (15% versus 10% probability of occurrence, respectively); the total likelihood of a factor-of-two error was 25% for this case. For probabilistic data, at the 5th percentile value of e (0.00001), e' was at least twice as likely to over-estimate rather than underestimate e by a factor of two (total probability of occurrence approximately 90%; there was nearly a 60% chance of a factor of ten error). The likelihood of occurrence decreases to approximately 80% (60%) at the 10th percentile (median) value of e, 0.001 (0.1); while the likelihoods of factor-of-two error are balanced for e=0.1, e' is eighty percent more likely to underestimate rather than overestimate e by a factor of five (16% versus 9%).

I. INTRODUCTION

Policy makers have used 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. The 1978 report to the U.S. Nuclear Regulatory Commission[1] 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 denoted as expert judgments. When these point estimates or probability distributions from multiple SMEs are combined, an aggregated expert judgment is formed.

A "set of experts can provide more information than a single expert"[2]. Such input needs to be aggregated or combined into single distribution to be used "as the basis of decision making"[3]. 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.

Although point estimates are less useful than interval estimates, the former continue to be observed in U.S. Government (USG) expert judgment elicitation studies[4, 5].

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"[6] or to develop probabilistic assessments given cost and feasibility constraints in obtaining hard data.

One of the fundamental issues in reliability analysis is "the uncertainty in the failure occurrences and consequences"[7] The author noted that the dominant contributors to risk are not necessarily the "design basis accidents" but rather looking at all feasible scenarios … with the probability of occurrence of such scenarios becoming an additional key aspect in order to rationally and quantitatively handle uncertainty" This key aspect, in part, motivated the research questions addressed in this paper in that expert judgment may be used to provide input into reliability analysis.

As part of ongoing research into the accuracy of expert judgment predictions, we examined observed values and predictions from an expert judgment meta-database. We investigated whether 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—mattered. Percentiles, standard deviations or ratios of physical quantities were considered to be physical. An example of 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. If we disaggregate the meta-data into physical and probabilistic-related subsets, does the range of multiplicative error around the estimate change? We attempted to determine, 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. We also explored whether multiplicative error varied with the level of the true value of the variable predicted.

II. DATA SOURCES

The data records used for this research include data from the Delft University of Technology (TUD.) and University of Maryland Center for Reliability and Risk Analysis (referenced hereafter as UMD). These data sources are collectively known as the Expert Judgement Extracts EJE. The EJE source consists of 606 TUD records and 1,182 UMD records for a total of 1,788 records. The data analysis operations applied to EJE to prepare the data for processing are documented in a work-in-progress UMD dissertation (A Meta-Data Informed Expert Judgment Aggregation and Calibration Technique).

The TUD data source[8] has been used in studies published in peer-reviewed literature[9, 10, 11, 12, 13]. 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. UMD data sources included Department of Mechanical Engineering graduate course work (A. Mosleh, personal communication, October, 2013) and two dissertations[14, 15].

TABLE 1: EJE: Summary-level Listing of Record, Prediction and Theme Count provides a summary-level listing of the records and predictions and theme count in EJE. 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. 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.

TABLE 1: EJE: Summary-level Listing of Record, Prediction and Theme Count

Records/Predictions/Themes↓EJDS Source→	TUD	UMD	EJE Total
Number of records in Physical Category	540	1,181	1,721
Number of records in Probabilistic Category	66	1	67
Subtotal Number of Records in Both Categories	*606*	*1,182*	*1,788*
Number of predictions in Physical Category	4,661	1,445	6,106
Number of predictions in Probabilistic Category	516	13	529
Subtotal Number of Predictions in Both Categories	*5,177*	*1,458*	*6,635*
Number of themes in Physical Category	27	16	43
Number of themes in Probabilistic Category	8	1	9
Subtotal Number of Themes in Both Categories	*35*	*17*	*52*

III. COMPARISON OF PREDICTION ACCURACY BETWEEN PHYSICAL AND PROBABILISTIC DATA

The measure of prediction accuracy used was the ratio r≡e/e' of realized value to predicted median value. This metric is scale invariant and accommodates the wide range of EJE data.

III.A. Comparison of Prediction Accuracy between Physical Data Subset and Probabilistic Dataset

The CDF of the ratios r≡e/e' for the subset of physical predictions for which e<1 (hereinafter, physical subset) was compared to the CDF of the r values for probabilistic predictions. Approximately 20% of the physical data records (by weight) have realized values e<1, and comprise the subset. To compute the CDF, each prediction in each data category was assigned a weight. Let $nthemes_{e<1}$ denote the number of themes in this category (19; each theme has weight ≈ 5%). Consider a single theme within this category. Let $nvar_{e<1}$ denote the number of variables in this theme having e values < 1. Consider a single such variable. Let nobs be the number of predictions associated with this variable. Then the inverse product $[nthemes_{e<1} \cdot nvar_{e<1} \cdot nobs]^{-1}$ is the weight associated with the prediction. This weighting scheme ensures that all e' associated with a given variable receive equal weight; all variables within a given theme receive equal weight; and all themes are weighted equally within the physical subset category.

An analogous process was used for the probabilistic category; note that since e<1 for all variables in this category, $nthemes_{e<1}$=nthemes and $nvar_{e<1}$=nvar. For each prediction e' in a given category, the ratio r was computed. The r values and their associated weights (the weights assigned to the e') were sorted by increasing value of r; weights were consolidated for r values which differed by less than 10^{-5}. An exception was made for values of r<<1 such as 6.5^{-6}. After consolidation, n_1=822 and n_2=261 unique (r, weight) pairs remained for the two data categories, respectively. Cumulating the weights yielded the CDFs for each category; the CDFs are shown in Figure 1: CDFs for Physical Subset and Probabilistic Data.

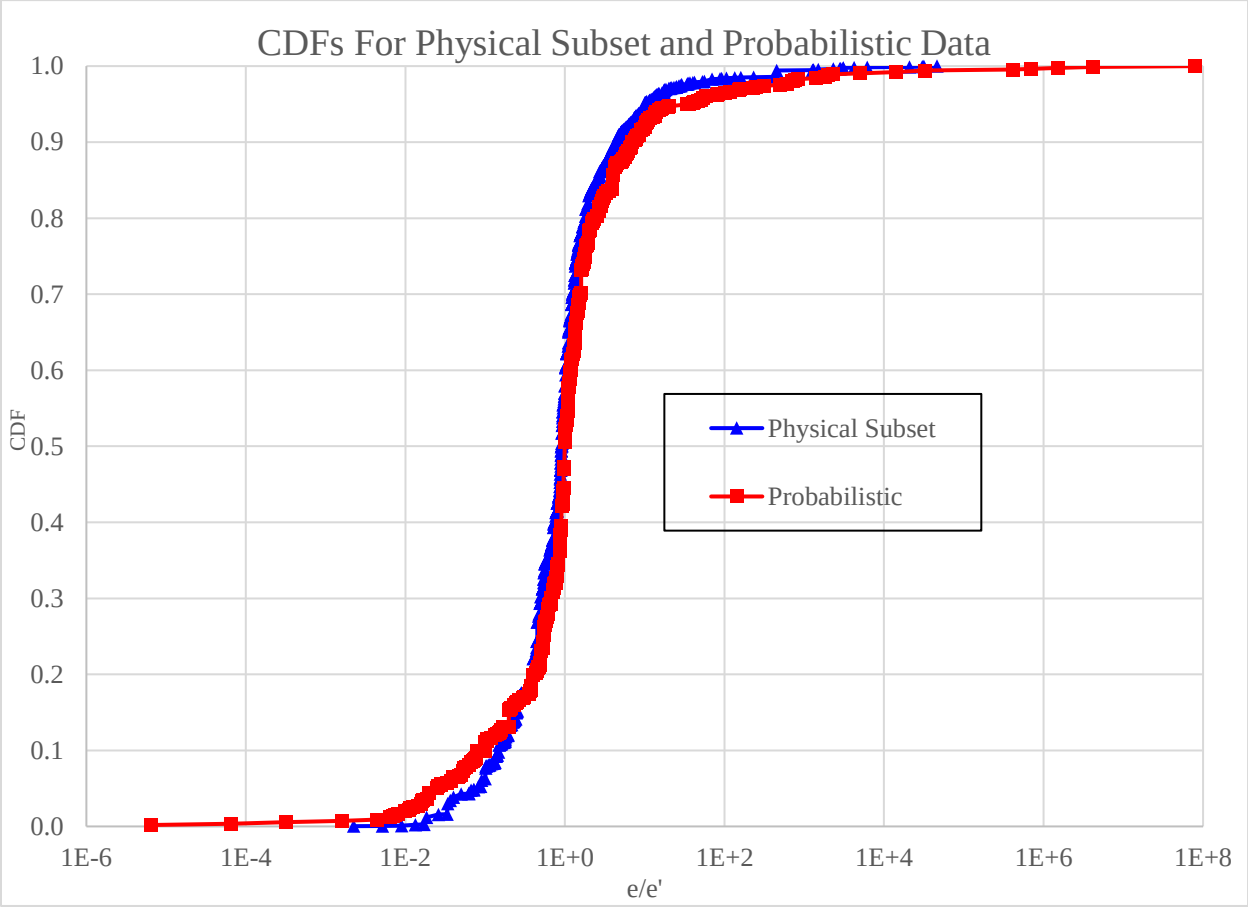

Figure 1: CDFs for Physical Subset and Probabilistic Data

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 means there is a probability of approximately six percent 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 2: Physical Data Subset 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. Figure 3: Physical Data Subset and Probabilistic Data – Underestimation Probability by Factor gives the analogous underestimation probabilities.

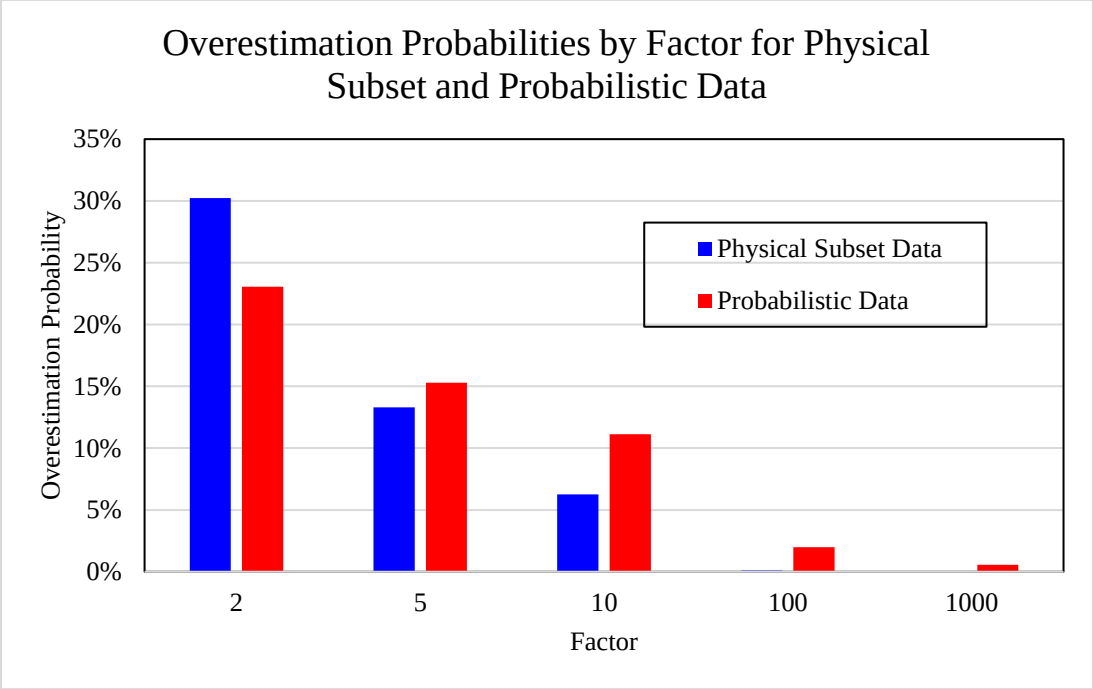

Figure 2: Physical Data Subset and Probabilistic Data – Overestimation Probability by Factor

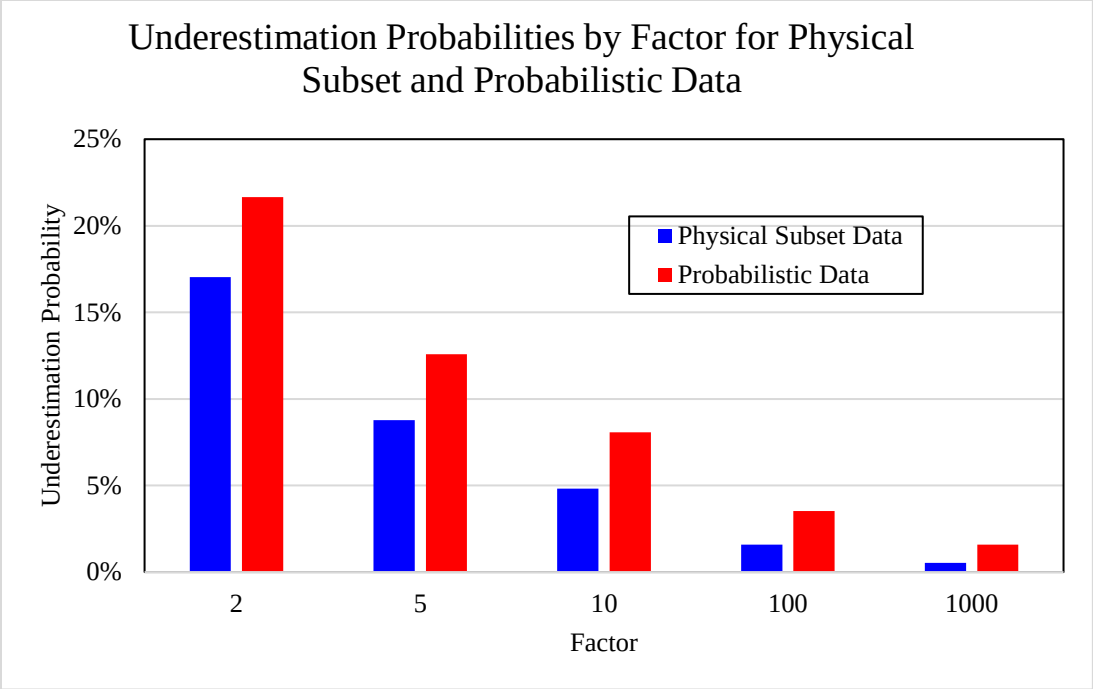

Figure 3: Physical Data Subset and Probabilistic Data – Underestimation Probability by Factor

The figures show 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 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 largest absolute difference, D between the two CDFs is approximately 0.113, occurring at r≈0.95. The numbers of pairs are sufficiently large to justify using the large-sample approximation for the K-S two-sided test. The critical value $D_{crit} = C_\alpha[(n_1+n_2)/(n_1 \cdot n_2)]^{-0.5}$, (per https://www.webdepot.umontreal.ca/Usagers/angers/MonDepotPublic/STT3500H10/Critical_KS.pdf). 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.)

It may also be noted if the K-S test were to be applied with all weights discarded, i.e. physical subset and probabilistic CDFs constructed giving all ratios e/e' equal weights, regardless of the theme or variable to which they belong, the difference D increases to 0.13 (at r=0.1), which would be significant at the $\alpha=0.01$ level of significance.

III.B. Comparison of Prediction Accuracy between Full Physical and Probabilistic Datasets

Having established a difference exists between 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 procedure 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 $n_1=3403$ 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 4: CDFs for Physical and Probabilistic Data.

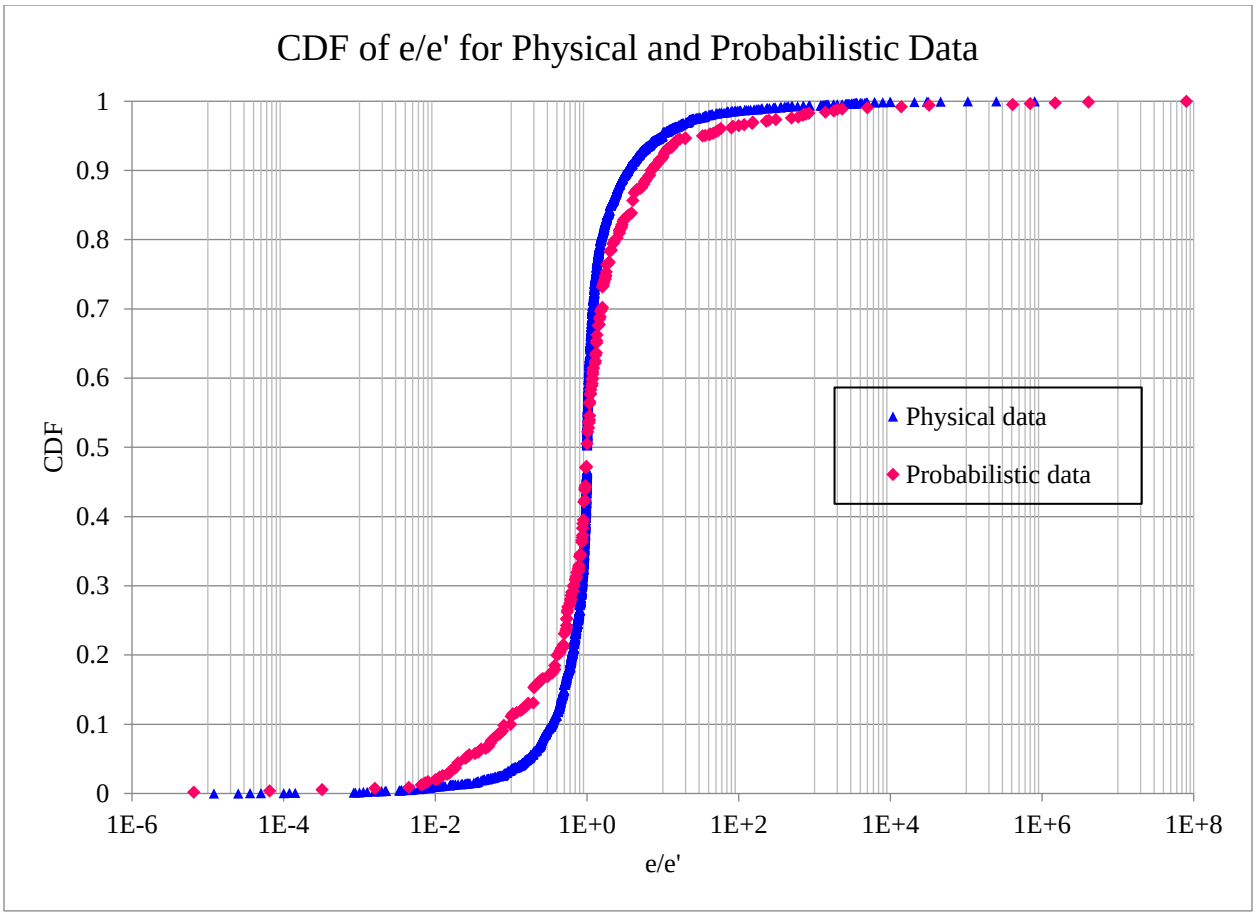

Figure 4: CDFs for Physical and Probabilistic Data

As before, given the CDFs, the probability that an elicited e' value over- or underestimates the realized value, e by a given factor is known. Figure 5: 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. Figure 6: Physical Data and Probabilistic Data – Underestimation Probability by Factor gives the probabilities of underestimating e by those same factors.

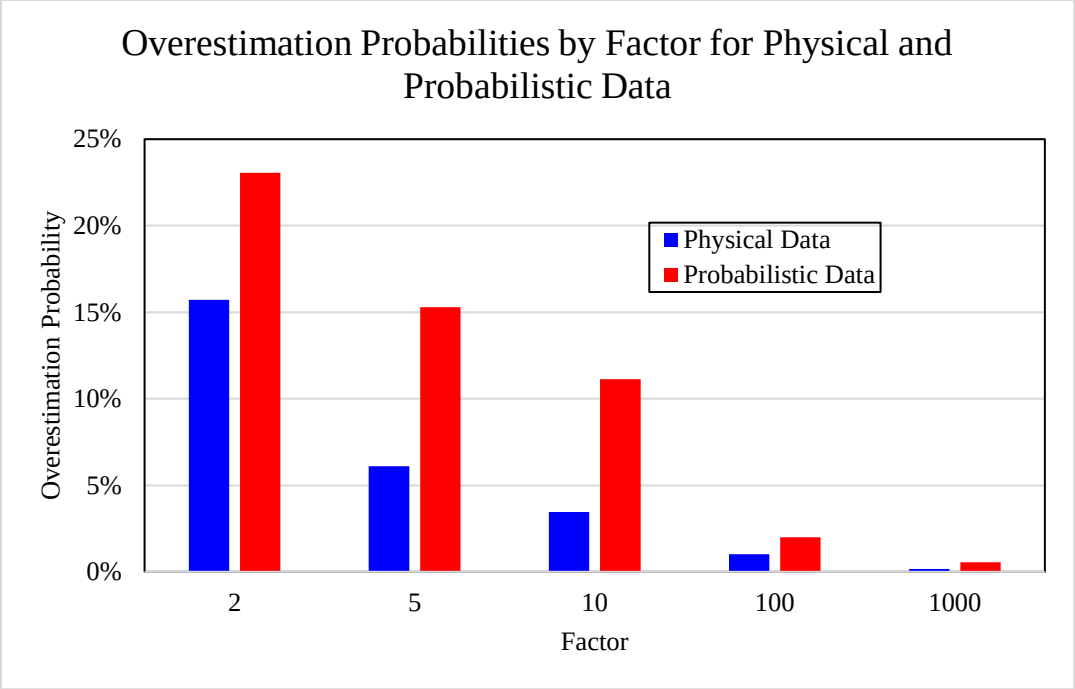

Figure 5: Physical Data and Probabilistic Data – Overestimation Probability by Factor

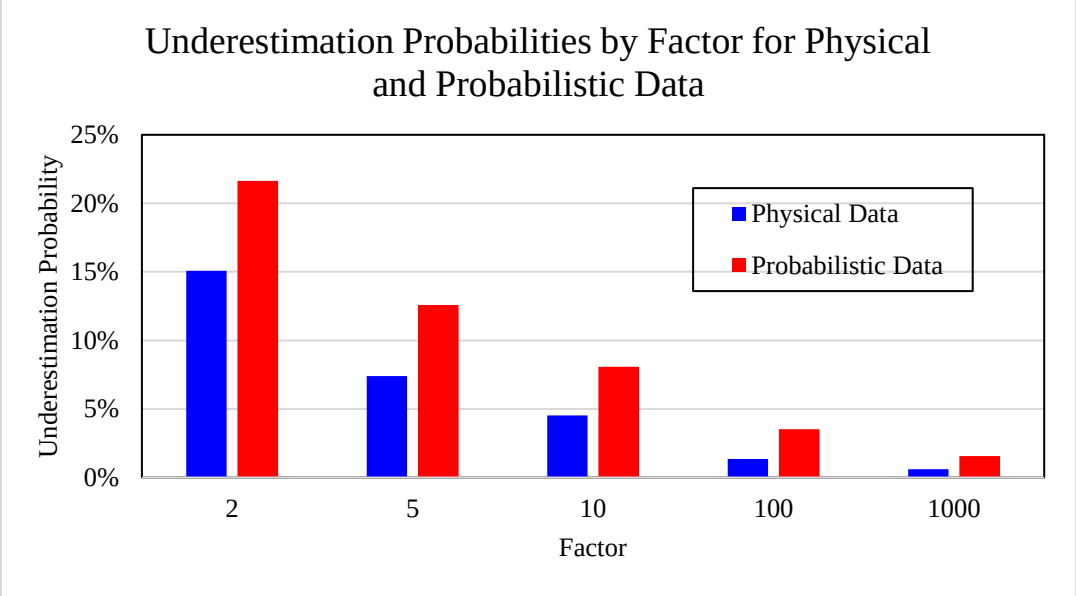

Figure 6: Physical Data and Probabilistic Data – Underestimation Probability by Factor

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 1000.

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.

III.C. Maximum Multiplicative Error, MME

We can consolidate over- and underestimation probabilities of e by using maximum multiplicative error, MME≡max(e'/e, e/e'), which is the Figure of Merit (FOM). For each (e,e') pair along with a corresponding weight (for a given data type), the MME was computed. 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 in below in Figure 7: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data.

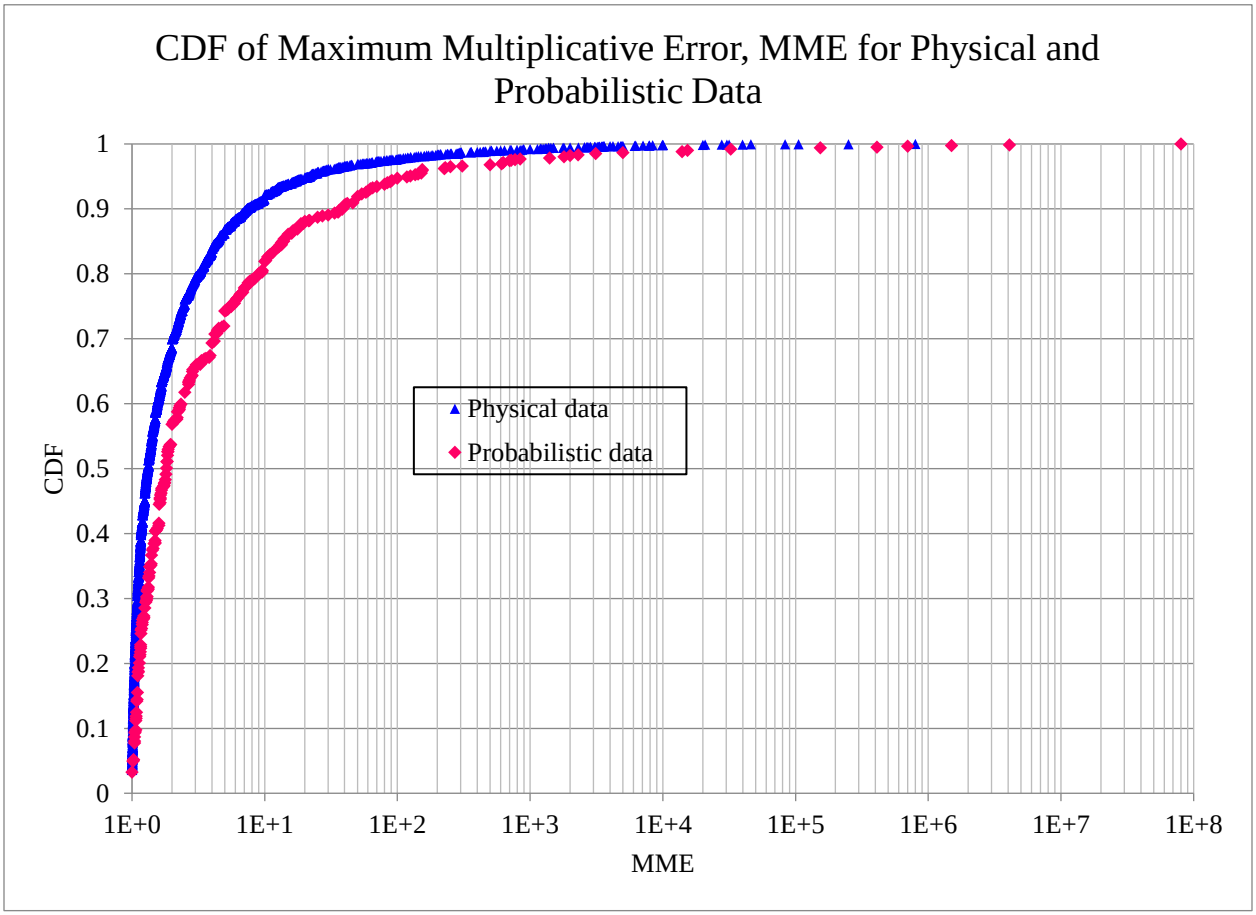

Figure 7: CDF of Maximum Multiplicative Error, MME for Physical and Probabilistic Data

The 80th and 90th percentiles, respectively 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. Given the CDF, the probabilities that MME equals or exceeds various factors were computed. Figure 8: Physical Data and Probabilistic Data – MME Exceedance Probability by Factor gives the probability that MME will equal or exceed a given factor for the two data categories, for the factors shown in Figure 8: Physical Data and Probabilistic Data – MME Exceedance Probability by Factor.

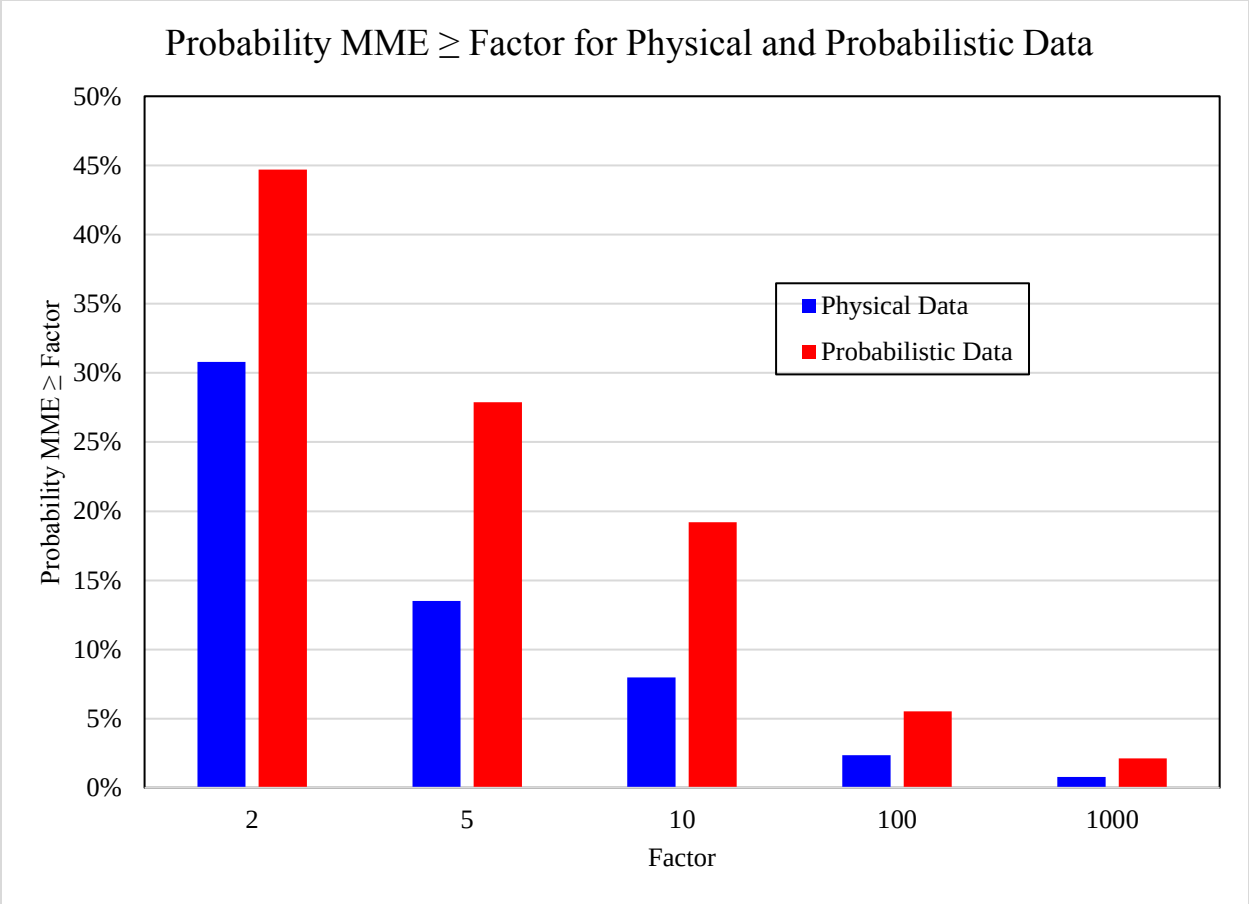

Figure 8: Physical Data and Probabilistic Data – MME Exceedance Probability by Factor

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.

III.D. Bounding Intervals

The 10^{th} and 90^{th} percentiles of the CDF of e/e' for physical data, and the 5^{th} and 95^{th} percentiles of the same CDF, can be used to obtain 80% and 90% bounding intervals for physical data e given e': [0.339e', 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', respectively: [0.08e', 7e'] and [0.02e', 36e'].

Since probabilistic data is bounded by one, adjustments are necessary in cases where the upper bound, 7e', exceeds unity. In this case, the lower bound associated with the two-sided 80% interval, 0.087e', may be used as a one-sided 90% lower bound. The one-sided 80% lower bound is obtained from the 20^{th} percentile of the CDF of r≡e/e' for probabilistic data; this 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 unity, the same lower bound associated with the two-sided 80% interval, 0.087e', may be used as a one-sided 90% lower bound, as before.

III.E. Influence of level of true value of variable predicted, on multiplicative error

As stated in the introduction, we also explored whether multiplicative error varied with the level of e, the true value of the variable predicted. To get a sense of the behavior of predictions e' with level of e, percentiles of ln(e/e') versus binned ln e were plotted for probabilistic data, using bins of width two with ranging from −18 to zero. (It was useful to log-transform the data first, given the large range of both e and e'.) See Figure 9: Percentiles of e'−e | e for Probabilistic Data.

The figure also shows the number of points in each bin: note that the leftmost two bins contain represent less than two percent of the data points.

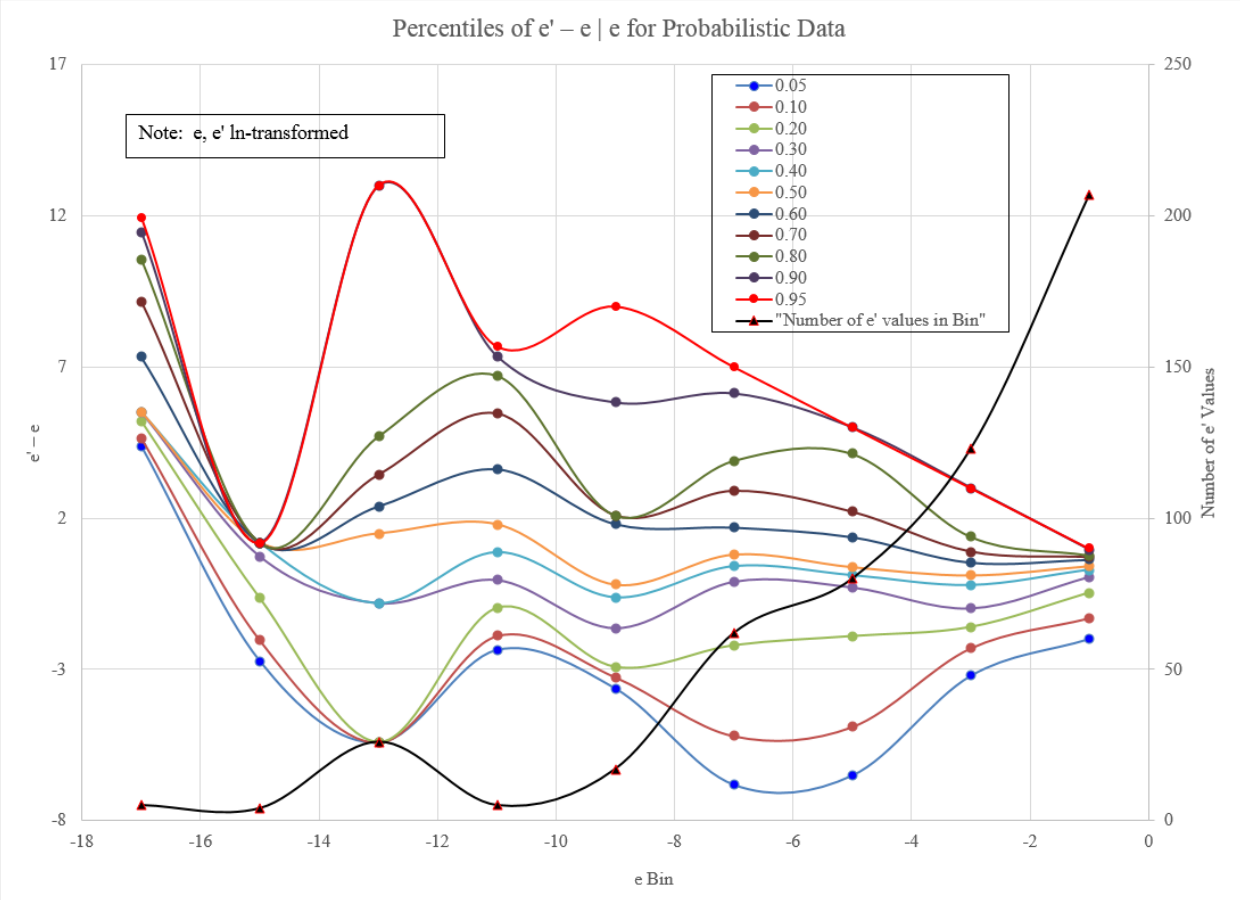

Figure 9: Percentiles of e'−e | e for Probabilistic Data

Figure 9 suggests that there is greater divergence between the prediction, e' and the true value, e at small values of e, and that the spread generally decreases as e approaches one. (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). Quantile regression was applied to fit a second-order polynomial predicting e' given each e, to the meta-data. The quadratic fit permitted curvature in the quantiles, without over-parameterizing the problem space. This facilitated exploration of the impact of level of e, on the extent of multiplicative error.

Quantile regression is based on minimizing the tilted absolute value function

Tilted_Abs(ρ,x)=x·(ρ–1_x), where 1_x=1 if x<0, 0 otherwise; and ρ is the quantile, e.g., 0.9.

The tilted absolute value function "*asymmetrically* [weights] absolute residuals residuals—simply giving differing weights to positive and negative residuals"[16] 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 record containing the (e, e') pair. The python function fmin was used to solve for the second-order polynomial coefficients minimizing

$$\sum \text{Tilted\_Abs}(\rho, e' - \text{polyval}(\text{coef}, e)) \cdot w_e \qquad (1)$$

over all pairs of log-transformed (e, e'), where polyval(coef,x) = $a_0+a_1x+a_2x^2$. The resulting quantile curves (5^{th}, 10^{th}, 20^{th}, 30^{th}, ..., 90^{th}, and 95^{th}) are plotted for probabilistic data below, in Figure 10: Quantiles of e' versus e for Probabilistic Data.

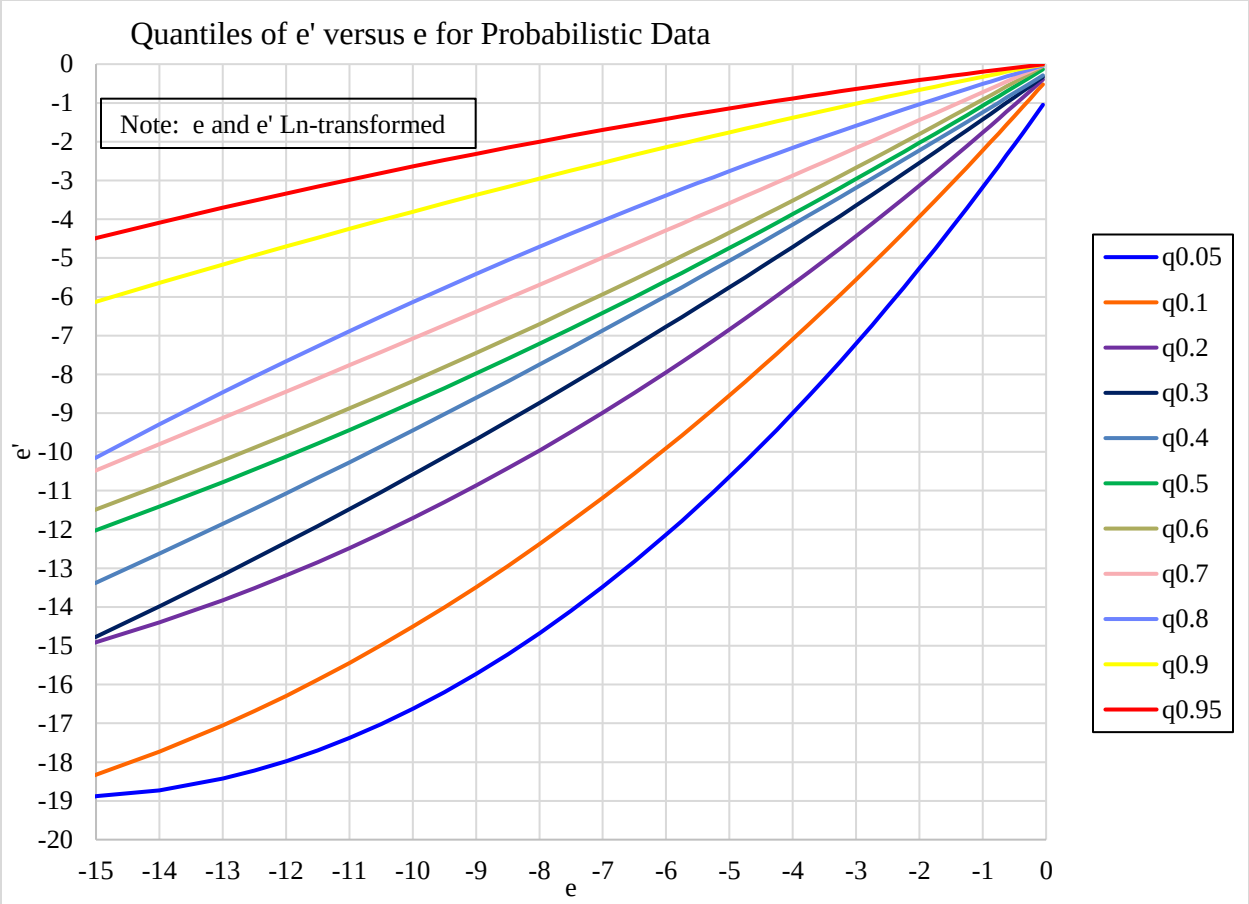

Figure 10: Quantiles of e' versus e for Probabilistic Data

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 **Error! Reference source not found.**. Quantiles of e' versus e for Physical Data. It will be observed that there is less curvature, i.e., variation in the distance between 5th and 95th quantiles, versus e, for physical data than for the probabilistic data above.

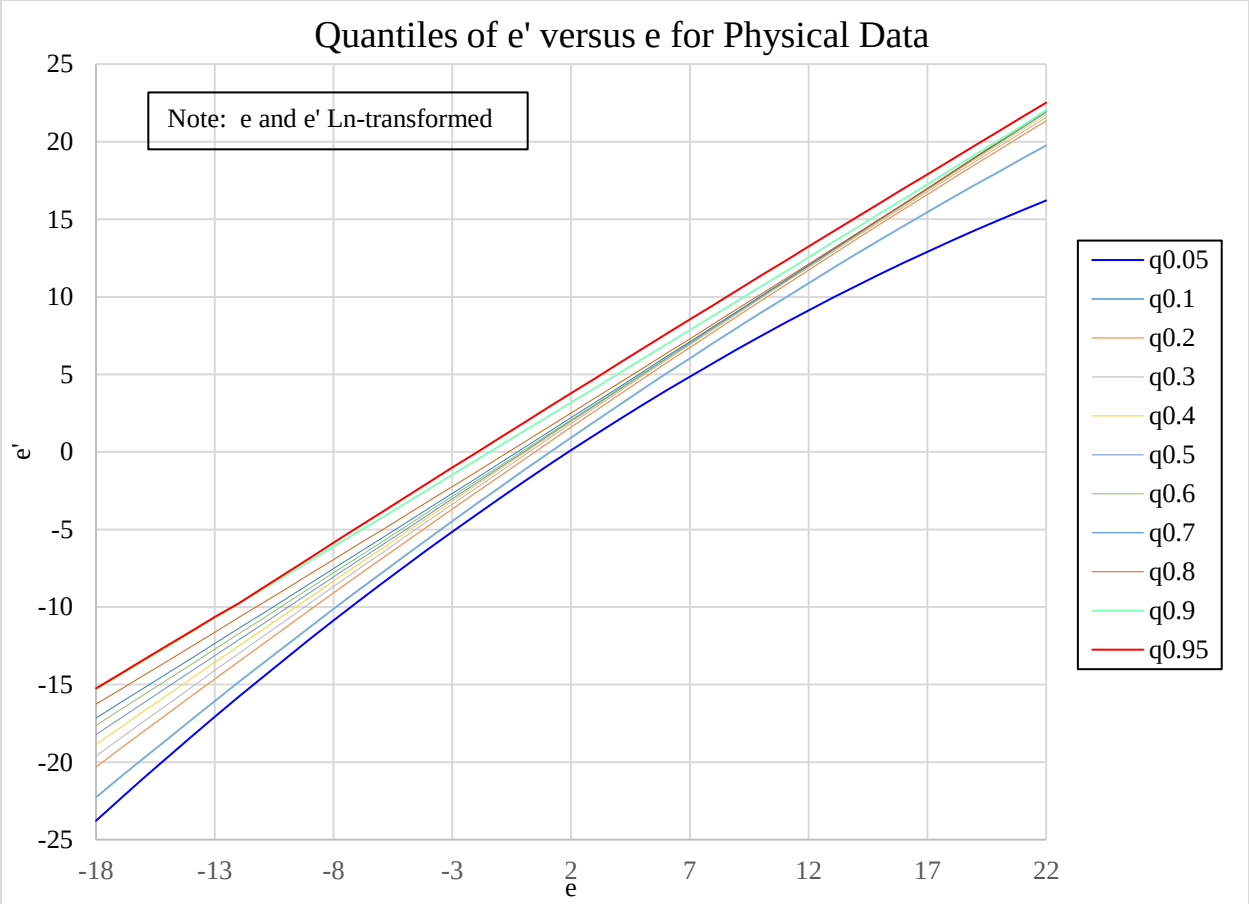

Figure 11: Quantiles of e' versus e for Physical Data

IV. OVER- AND UNDER-ESTIMATION PROBABILITIES VERSUS Ln(e)

Given the quantile curves, charts showing variation in over- and under-estimation probabilities with level of Ln(e) were created for probabilistic data. These probabilities are given in Figure 11: Probability that e' under- or overestimates e by various factors, s, for probabilistic data.

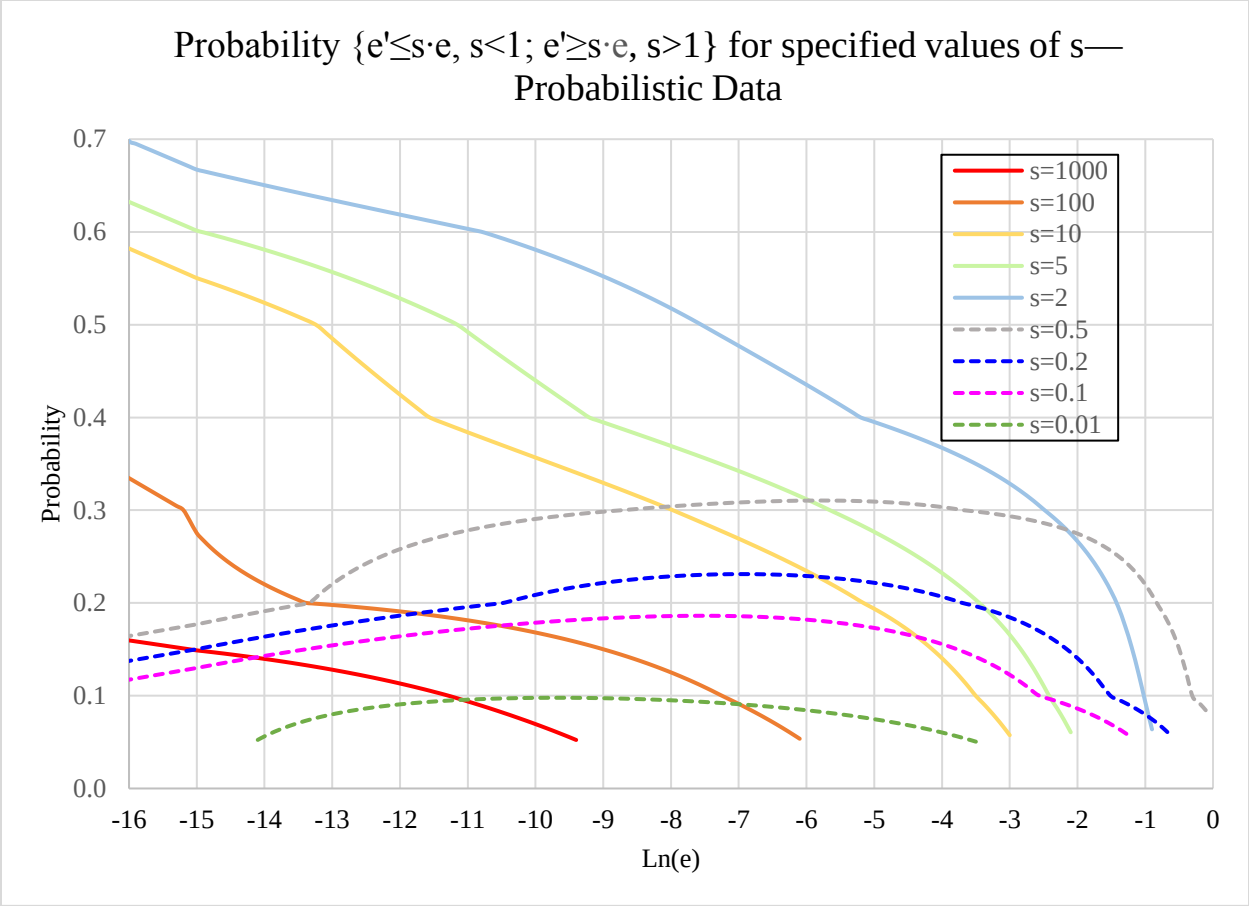

Figure 11: Probability that e' under- or overestimates e by various factors, s, for probabilistic data

An analogous process yielded the results in Figure 12: Probability that e' under- or overestimates e by various factors, for physical data.

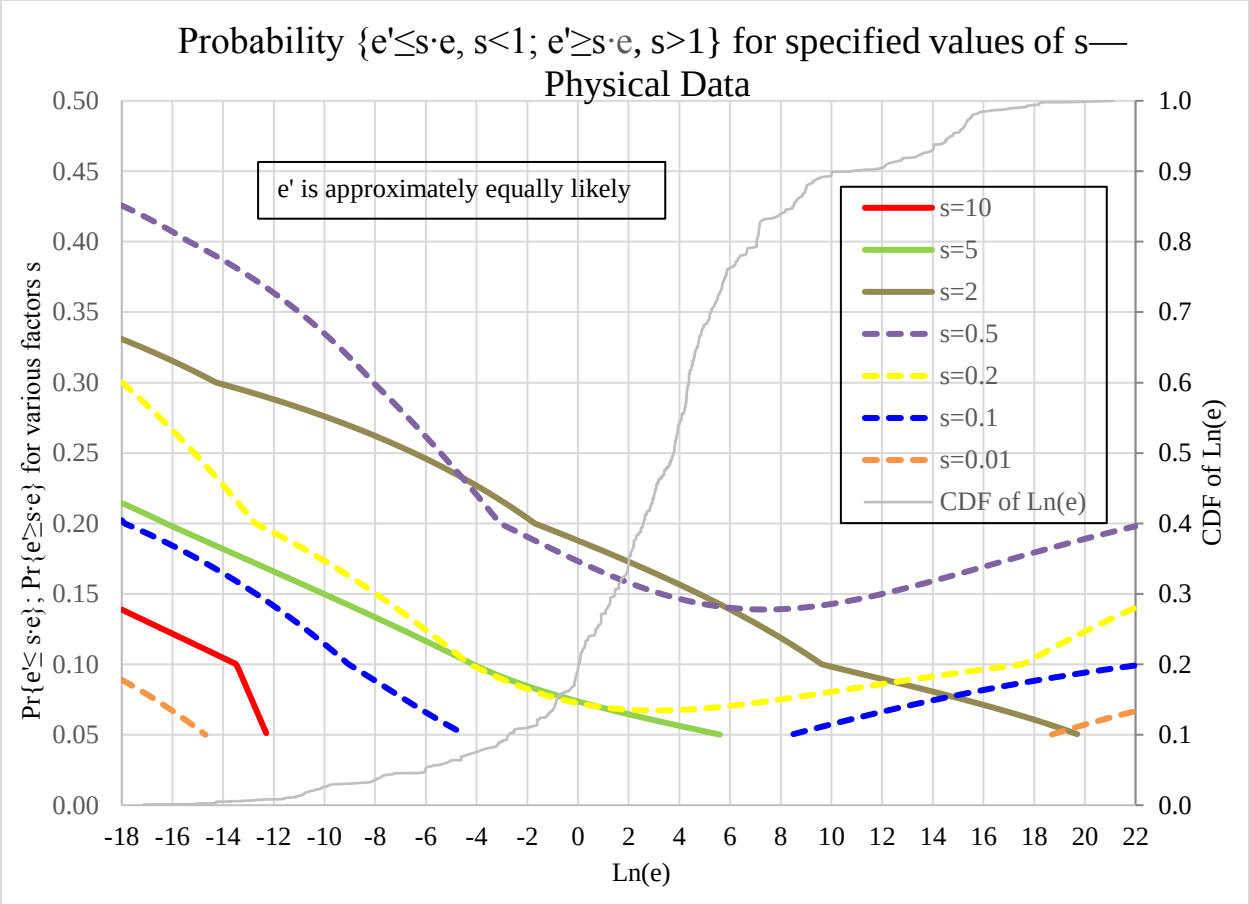

Figure 12: Probability that e' under- or overestimates e by various factors, for physical data

The figures show that for physical data, over values of e representing the 75th to the 5th percentile, the likelihood of a factor-of-two error ranged from approximately 30% to 50%, split approximately equally between over- and under-estimation. At the 90th percentile value of e, under- rather than overestimation was fifty percent more likely to occur (15% versus 10% probability of occurrence, respectively); the total likelihood of a factor-of-two error was 25% for this case. For probabilistic data, at the 5th percentile value of e (0.00001), e' was at least twice as likely to over-estimate rather than underestimate e by a factor of two (total probability of occurrence approximately 90%; there was nearly a 60% chance of a factor of ten error). The likelihood of occurrence decreases to approximately 80% (60%) at the 10th percentile (median) value of e, 0.001 (0.1); while the likelihoods of factor-of-two error are balanced for e=0.1, e' is eighty percent more likely to under- rather than overestimate e by a factor of five (16% versus 9%).

Note: The curves obtained by quantile regression are less accurate in the sparse tails of Ln(e). Additionally, the quadratic model is itself an approximation, and is intended to give a rough sense of the variation in multiplicative error with level of e, given the softness of the expert judgment data.

V. CONCLUSIONS

1. Comparison of the empirical CDFs of r≡e/e' show that the type of quantity estimated does make a difference: the probability that maximum multiplicative error, MME will exceed a given factor—whether the factor be 2,5,10,100, or 1000—is approximately twice as large for probabilistic data as for physical data.
2. Based on the CDFs, approximate 80% and 90% bounds for e given e' are [0.339e', 3.45e'] and [0.165e', 9.84e'], respectively, for physical data. Wider 80% and 90% bounds apply for probabilistic data: [0.08e', 7e'] and [0.02e', 36e'], respectively.
3. The level of e has a marked impact on overestimation error. The effect is greatest for probabilistic data. For this type of data, the probability of overestimating by a factor of s, where s = 2, 5, or 10, declines approximately linearly as Ln(e) increases, from above fifty percent at Ln(e)=−16, to less than five percent as·e approaches 0.6. For physical

data, the corresponding probabilities of overestimation are less than half of their counterparts for probabilistic data, at Ln(e)=−16; and they decline more gradually, not reaching five percent until Ln(e) is at least 200.

4. The level of e has a weaker impact on underestimation error. For both data types, the probability of underestimating by a factor of s, where s=2, 5, or 10, does not vary monotonically with Ln(e). Underestimation probabilities are generally confined to a region between 10 to 30 percent for probabilistic data. For physical data, with s=2, they can exceed 40 percent for Ln(e)<−16, decline gradually to a minimum in a broad region around Ln(e)=−2, then increase gradually to a level roughly half their starting level as Ln(e) approaches its upper limit of +22.

REFERENCES

1. 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).
2. CLEMEN, R. T., & WINKLER, R. L. (1999). Combining probability distributions from experts in risk analysis. Risk Analysis, 19(2), 187-203.
3. 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
4. US DEPARTMENT OF AGRICULTURE FOOD SAFETY AND INSPECTION SERVICE. (2007, DRAFT). Results of an additional expert elicitation on the relative risks of meat and poultry products. 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.
5. US 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.
6. 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.
7. ZIO, E. (2009). Reliability engineering: Old problems and new challenges. Reliability Engineering and System Safety, 94, 125-141
8. COOKE, R.M &. GOOSSENS L. L. H. J. (2008). TU Delft expert judgment data base. *Reliability Engineering and System Safety, 93,* 657–674. doi:10.1016/j.ress.2007.03.005.
9. 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.
10. SHI-WOEI, L. & BIER, V.M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety 93,*711–721
11. 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.
12. 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.
13. EGGSTAFF, J. W., MAZZUCHI, T. A. AND 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.
14. FORRESTER, Y. (2005). The quality of expert judgment: An interdisciplinary investigation (Doctoral Dissertation). Retrieved from http://hdl.handle.net/1903/3267.
15. 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.

16. KOENKER, R., & HALLOCK, K. F. (2001). Quantile regression. *Journal of Economic Perspectives, 15* (4), 143–156. http://digitalcommons.ilr.cornell.edu/hrpubs/19/.