PROBABILISTIC PROPERTY PREDICTION

D. Gary Harlow
Mechanical Engineering and Mechanics; Lehigh University, Bethlehem, PA 18015 USA

ABSTRACT

Because it is integral to modeling, experimentation, and manufacturing, uncertainty cannot be ignored. It, therefore, must be adequately characterized and managed especially for accurate design and life cycle estimation and prediction for engineering applications. Whenever new materials are developed or changes in design are suggested for critical engineering applications, usually only limited data and information are available. Thus, concerns about the magnitude and extent of the uncertainty are intensified, especially for high reliability applications. The purpose of this effort is to present a methodology so that uncertainty can be adequately taken into account. In this context, the uncertainty is assumed to encompass the individual uncertainties from all sources. The proposed methodology focuses on the integration of limited experimental data with science based modeling. This amalgamation is achieved through a combination of traditional Bayesian analyses and model synthesis with limited data to manage the inherent uncertainty. The methodology is specifically illustrated for the yield strength of a typical turbine disk alloy. One of the most important conclusions of the analysis is that the fusion of science based modeling with data greatly improves reliability estimation and prediction. As the accuracy of the modeling increases, the effect of uncertainty is reduced; however, even crude approximations for the model are more beneficial than using statistical data analysis alone. The methodology focuses on the estimation and prediction of the cumulative distribution function for the property of interest and statistical confidence bands for parameters within the distribution. Possibly the greatest benefit to the methodology is that it allows for a significant reduction in the number of data required for accurate estimation in contrast to purely statistical approaches.

1 INTRODUCTION

The effect of uncertainty is increasingly being considered in the design of complex engineered systems and components as well as the estimation and prediction of their life cycles. Uncertainty, which cannot be eliminated entirely from experimentation, manufacturing, and modeling, must be adequately characterized and managed. Typically limited data and information are available for new materials and new designs developed for insertion into engineering applications. The effects of uncertainty are intensified, especially when high reliability must be assured. A methodology is proposed to adequately manage uncertainty. Here, uncertainty is assumed to include contributions from any and all sources. Historically, uncertainty has been evaluated by accumulating as much data as possible given economic and time constraints; however, that data often do not resolve all of the crucial questions and concerns for the material behavior. Consequently, design and life cycle decisions are made from deficient information. Furthermore, certification of new materials for critical components is extremely labor intensive and time consuming. In order to address these concerns, extensive effort has been exerted in scientific modeling of the material behavior, which is essential for uncertainty management. The proposed methodology focuses on the science based modeling for yield strength (SBMYS) for a typical turbine disk alloy. The SBMYS developed previously and described in Harlow [1] will be synthesized with experimental data. The synergy of modeling and data is achieved by calibration of the SBMYS with small numbers of data in order to manage the uncertainty and to minimize the number of data required for accurate predictions.

An extensive set of yield strength data, obtained from production components, consisting of 129 samples tested at room temperature is used for the ensuing analyses and validation of the methodology. Figure 1 shows the data and corresponding simulations from the SBMYS. Note that

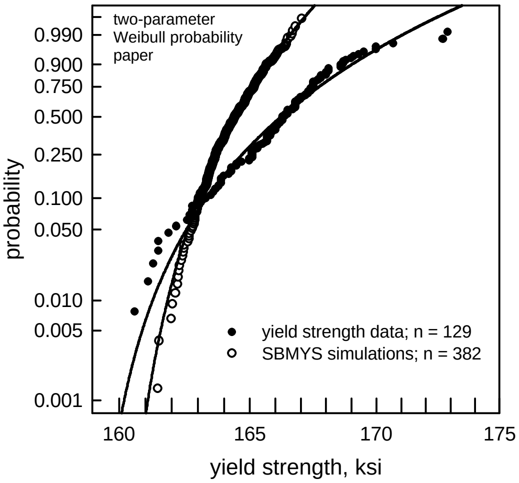

Figure 1: Production data and simulations from the SBMYS model, plotted on two-parameter Weibull probability paper.

the simulations are fairly good, but the variability is a bit too small. Consequently, there is some uncertainty between the model and the data that may be attributed to a variety of sources.

Brief comments on the Bayesian analysis used herein are followed by a presentation of the method used for data fusion to account for the uncertainty. The selection of an underlying sample cumulative distribution function (cdf) is vital for the analysis. The formulation of a convergence criterion, which is independent of large quantities of data, is followed by computations and observations derived from the synergistic approach incorporating limited data with the SBMYS.

2 BAYESIAN ANALYSIS

The basis for Bayesian analysis is that the cdf parameters are themselves random variables (rvs). Thus, the cdf, which statistically characterizes the data and adequately represents the physical phenomenon of interest, must be determined before any Bayesian analysis can be conducted. Although knowing the prior cdf for the parameters would be nice, there are reasonable approximations even when knowledge is scarce. Usually a non-informative cdf is used initially for the parameters, which may, in turn, depend on the form of the cdf.

The major difficulty is the numerical computation for the posterior cdf. The sampling-resampling method is used subsequently because it is a rather simple, but effective algorithm; see Rubin [2] and Smith and Gelfand [3]. One of the primary values of the Bayesian methodology is that it is easy to implement iteratively as additional information becomes available. After using a non-informative prior initially, subsequent computations use the previous posterior as the next prior. As the number of iterations increases, the validity of the estimation increases.

3 SELECTION OF THE CDFS

The cdf is usually chosen empirically to sufficiently characterize statistical data. In the ensuing analysis, however, the form of the cdf was selected in order to not only represent experimental data, but more importantly to characterize the behavior of the SBMYS. The most significant consequence from the results of the SBMYS that relates to the cdf is that there is a theoretical minimum for the yield strength. Hence, the lower tail of the cdf must be accurately estimated. The three-parameter Weibull cdf, which is an excellent choice for both the SBMYS and the data, is given by

$$F(x) = 1 - \exp\{-[(x-\gamma)/\beta]^{\alpha}\}, \quad x \geq \gamma \tag{1}$$

where α, β, and γ are the shape, scale, and location parameters, respectively. The efficacy of this selection is also seen on Fig. 1. Except for a few data, the fit is quite good for the yield strength data, and it is even better for the simulations. Thus, the Weibull cdf is an excellent choice.

The other cdf required for Bayesian analysis is the initial prior. Since it is assumed that the initial prior is unknown, a non-informative prior is chosen. It is also assumed that (α, β, γ) are statistically independent with a marginal uniform cdf over an appropriate interval. The ranges for the uniform cdfs are selected to be consistent with the data available for the initial computation. The most important range to specify is that for γ because it cannot exceed the smallest data.

4 SYNTHESIS OF MODEL SIMULATIONS WITH DATA

Since the production data are so extensive, it is assumed that they adequately represent the statistical behavior of the yield strength. Consequently, subsets of the production data of size *k*, $\{x_{i(j)} : 1 \leq j \leq k\}$, are used to calibrate the SBMYS simulations to account for discrepancies with the production data. The method of synthesis was chosen to match the average and standard deviation of the calibrated SBMYS simulations with those of the subset. Let $\{y_j : 1 \leq j \leq N_S\}$ be the SBMYS simulations, where N_s is the number of simulations. The linear transformation of y_j into z_j is

$$z_j = ay_j + b; \text{ where } a = s_x(k)/s_y; \text{ and } b = \overline{x}(k) - (s_x(k)/s_y)\overline{y}\ , \tag{2}$$

where $s_{x(k)}$ and s_y are the standard deviations and $\overline{x}(k)$ and $\overline{y}$ are the averages, respectively, for $\{x_{i(j)} : 1 \leq j \leq k\}$ and $\{y_j : 1 \leq j \leq N_S\}$.

5 CONVERGENCE CRITERION

Since the production data are so extensive, a small subset was randomly selected to represent data obtained through an experimental program, $\{x_{i(j)} : 1 \leq j \leq k\}$. All of the production data; however, are used for validation of the methodology. A number of numerical computations were performed fusing different random selections of the data with the SBMYS simulations. The required number is highly dependent on the convergence criterion used to establish the accuracy of the estimates obtained from the analysis. The criterion consists of two conditions that must be concurrently satisfied.

The first condition is similar to the Kolmogorov-Smirnov (K-S) goodness-of-fit test, see Leemis [4]. For the *k* data fused with SBMYS simulations, see eqn (2), the maximum deviation between the empirical cdf for the *k* data and the Bayesian estimate $\hat{F}_k(x)$ for the linearly transformed SBMYS simulations is desired. The criterion requires that the maximum deviation be at most $1/\sqrt{k}$ to force the convergence to be reasonably tight relative to the production data.

The second condition is imposed because for small *k* the calibration of the SBMYS simulations may not be as accurate as desired. An additional constraint guarantees uniform convergence of the Bayesian estimate $\hat{F}_k(x)$. This constraint restricts the maximum deviation between successive iterations of the Bayesian estimate:

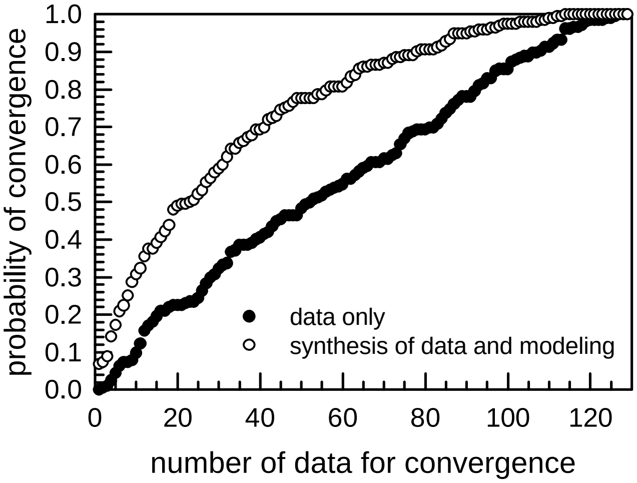

Figure 2: Probability of convergence as a function of the number *k* of randomly chosen data for data only and for linearly tuned SBMYS simulations.

$$\max\{|\hat{F}_{k+1}(x) - \hat{F}_k(x)|;\ x \geq 0\} < \varepsilon\,, \tag{3}$$

where ε is chosen to be sufficiently small; 0.05 for this example.

6 COMPUTATIONS AND ANALYSES

6.1 Data are analyzed solely with the Bayesian method.

To simulate distinct laboratory experiences, the data are randomly selected consecutively, without replacement. Only the data, without any model synthesis, are iteratively analyzed with Bayesian methods. When sufficient data have been selected to satisfy the convergence criterion, the ensuing estimation is compared to the best three-parameter Weibull cdf for the entire set of data for validation. The primary purpose of this exercise is to establish a baseline for the number of data needed to adequately model the yield strength when no scientific modeling is used.

Figure 2 shows the probability of convergence as a function of the number *k* of randomly selected data required. To have a reasonable estimate, 205 randomizations were considered. It is apparent that a rather large sample is required for convergence with high probability. Figure 3 is a graph of the fluctuations of the 2.5 percentile of $\hat{F}_k(x)$. This percentile was chosen because it is frequently used as input for design and certification standards. The large point on the right axis, taken from Fig. 1, is the best estimate using all production data. The dashed horizontal lines correspond to ±0.5 ksi deviation from the best estimate. Note that when *k* exceeds 92, convergence is within the ±0.5 ksi interval. A sample of 92 corresponds to a probability of convergence of 0.8 on Fig. 2. Thus, the number of data required for convergence with no additional input is quite large, in excess of 90.

6.2 Randomly selected data are synthesized with SBMYS simulations.

A subset of the production data of size *k*, randomly selected consecutively, without replacement, is fused with SBMYS simulations in accordance with eqn (2), and these calibrated simulations were

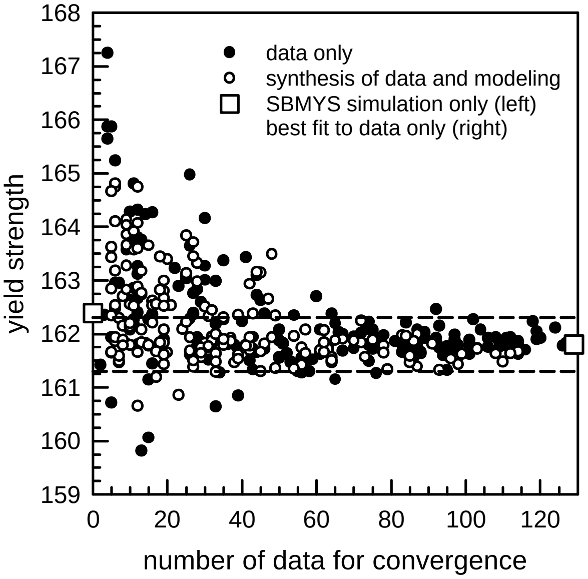

Figure 3. Fluctuations in the 2.5 percentile for the Bayesian estimates of the three-parameter Weibull cdf as a function of the number of randomly selected data.

subsequently analyzed by Bayesian methods. The iterative selection is terminated when the convergence criterion is satisfied. The number *k* of data required to satisfy the convergence criterion is substantially reduced because of the significant contribution of the synthesis with SBMYS.

Figure 2 also shows the probability of convergence as a function of *k* when data and modeling are synthesized. The difference between the approaches is striking. If a probability of 0.8 is considered, based on the convergence of the data alone, then *k* is reduced to 55. Likewise, Fig. 3 shows the fluctuations in the 2.5 percentile for the Bayesian estimates of the three-parameter Weibull cdf for the linearly calibrated SBMYS simulations as a function of the number *k* of data used in the synthesis. When *k* exceeds 49 the convergence is within ±0.5 ksi deviation from the best estimate for the 2.5 percentile. The ±0.5 ksi deviation was chosen because 1 ksi is approximately the width of the 50% confidence interval, determined from the Bayesian analysis, at the 2.5 percentile The large point on the left axis is the 2.5 percentile from the SBMYS simulations only. Although the SBMYS may not be as accurate as desired, it only deviates by about 0.6 ksi for this percentile, which is confirmation that the SBMSY is quite good.

In order to verify the validity of the convergence criterion for the methodology, the entire estimated cdf was statistically assessed with goodness-of-fit tests to the entire sample of data. For estimates satisfying the convergence criterion, the statistical tests confirmed that the cdfs were excellent representations of the entire production data. Thus, the convergence criterion defined above is quite reasonable, and its use in the proposed methodology is merited.

7 CONCLUSIONS

Strategically fusing limited experimental data with scientifically based probability modeling is essential for estimation and prediction of behavior outside the range of typical laboratory conditions. Nevertheless, uncertainty is a factor that must be considered. Herein, a methodology was proposed that accommodates uncertainty. Three key ingredients are necessary for success. First and foremost, an accurate scientifically based model is essential because modeling deficiency is a major source of uncertainty. Second, a good estimate for the cdf for the material property of interest is required for the Bayesian analyses. For yield strength, the three-parameter Weibull cdf is an excellent choice because the scientifically based model indicates such behavior and extensive production data confirms the choice. The overriding reason for this choice is the characterization of the lower tail behavior. Third, data to estimate model parameters, to calibrate the scientifically based probability model, and to validate the methodology are necessary. The calibration is needed to offset the effect of uncertainty. Of course, a well defined validation scheme, especially for new applications, is paramount.

One of the most significant results from this effort is that iteratively using standard Bayesian analysis, as new data are available, with scientifically based probability modeling improves estimates and confidence significantly. The analysis is crucially dependent on a convergence criterion. The proposed convergence criterion is independent of the existence of large amounts of data because it combines a statistical measure of accuracy with uniform functional convergence as data are available, regardless of sample size. The criterion was validated, and its strength lies in the accuracy of the scientifically based model. Possibly the most significant result for economic and time investments concerns the necessary sample size for testing. The combination of strategic testing with scientifically based modeling can drastically reduce the required experimental sample size. For statistical analysis using nothing other than the data itself, about 90 samples are required before the convergence criterion is satisfied. When the experimental data are fused with the scientifically based model, about 50 samples are needed to satisfy the same convergence criterion, which is a significant reduction in the quantity of experimental data. As the accuracy of SBMYS increases, the number of required data decreases. Again, modeling precision is paramount.

Thus, the proposed methodology has been shown to be extremely effective for modeling yield strength. No doubt, further refinements could be made; however, the approach seems to be quite sound. Although the application of the methodology to another situation for confirmation is warranted, the adoption of the procedure is recommended.

8 ACKNOWLEDGMENT

This effort was supported in part by the DARPA initiative, Accelerated Insertion of Materials (AIM) as administered under USAF contract F33615-00-2-5216.

9 REFERENCES

[1] Harlow, D.G., “The Effect of Randomness in Complex Models,” *Proceedings of the Ninth ISSAT International Conference on Reliability and Quality in Design*, Pham, H. and Yamada, S., eds, International Society of Science and Applied Technology, New Brunswick, NJ, 2003, pp. 284 – 288.

[2] Rubin, D.B., “Using the SIR Algorithm to Simulate Posterior Distributions,” *Bayesian Statistics*, Vol. 3, Bernado, J.M., et al., eds, Oxford University Press, 1988, pp. 395 – 402.

[3] Smith, A.F.M. and Gelfand, A.E., “Bayesian Statistics without Tears: A Sampling-Resampling Perspective,” *American Statistical Association*, Vol. 46, 1992, pp. 84 – 88.

[4] Leemis, L.M., *Reliability: Probabilistic Models and Statistical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1995.