Improved Test Planning and Analysis Through the Use of Advanced Statistical Methods

Lawrence L. Green[1]
NASA Langley Research Center, Hampton, VA 23681
Katherine A. Maxwell[2]
University of Minnesota, Minneapolis, MN 55455
Dr. David E. Glass[3] and Dr. Wallace L. Vaughn[4]
NASA Langley Research Center, Hampton, VA 23681
Weston Barger[5]
University of Washington, Seattle WA 98195
Mylan Cook[6]
Brigham Young University, Provo, UT 84602

The goal of this work is, through computational simulations, to provide statistically-based evidence to convince the testing community that a distributed testing approach is superior to a clustered testing approach for most situations. For clustered testing, numerous, repeated test points are acquired at a limited number of test conditions. For distributed testing, only one or a few test points are requested at many different conditions. The statistical techniques of Analysis of Variance (ANOVA), Design of Experiments (DOE) and Response Surface Methods (RSM) are applied to enable distributed test planning, data analysis and test augmentation. The D-Optimal class of DOE is used to plan an optimally efficient single- and multi-factor test. The resulting simulated test data are analyzed via ANOVA and a parametric model is constructed using RSM. Finally, ANOVA can be used to plan a second round of testing to augment the existing data set with new data points. The use of these techniques is demonstrated through several illustrative examples. To date, many thousands of comparisons have been performed and the results strongly support the conclusion that the distributed testing approach outperforms the clustered testing approach.

Nomenclature

ANOVA	=	analysis of variance
COV	=	coefficient of variation
DOE	=	design of experiments
DX9	=	Design Expert software, version 9
e_{RMS}	=	root mean square error
e_{COV}	=	error in the estimated variability
OFAT	=	one factor at a time
RSM	=	response surface methods
SME	=	Subject Matter Expert
XML	=	Extensible Markup Language, a subset of the Standard Generalized Markup Language

[1] Distinguished Research Associate, Vehicle Analysis Branch, MS451, Senior Member AIAA.

[2] Graduate Student, Mathematics Department, Morrill Hall 100 Church St. S.E.

[3] Senior Researcher, Structural Mechanics & Concepts Branch, MS190, Senior Member AIAA.

[4] Senior Researcher, Advanced Materials & Processing Branch, MS188AMS190.

[5] Graduate Student, 225 Schmitz Hall Campus Box 355850.

[6] Undergraduate Student, A-41 ASB.

I. Introduction

THE current processes used for obtaining experimental data for materials follow the guidelines established in MIL-HDBK-17-1F[1] and its predecessor documents. The history of this document dates back as far as 1943[2]. Version 1F of the handbook was released in 2002[1]. This approach, identified herein as clustered testing, results in multiple, repeated tests points selected at a limited number of test conditions. For testing involving numerous independent variables, a rectangular grid of testing conditions, typically known as the one-factor-at-a-time (OFAT)[3] method of testing, is also routinely employed. One now outdated foundation of the OFAT testing strategy is that complicated statistical analyses could not be routinely performed on old computers. Three distinct disadvantages of the OFAT testing approach include: (1) OFAT requires more runs for the same precision in effect estimation, (2) OFAT cannot estimate multi-factor interactions, and (3) OFAT can miss optimal settings of factors[3].

Considerable progress has been made in computational capabilities since the time[4-5] when the OFAT method was the default standard testing method. Peak computer speeds and the number of cores that can be utilized have increased by orders of magnitude in that time frame. Yet, the process to obtain and analyze experimental data for composite materials has remained largely unchanged for years.

This paper proposes a different approach to composite materials testing, data analysis and data augmentation. The proposed process (distributed testing) is based upon the statistical technique known as Analysis of Variance (ANOVA)[6-13]; ANOVA enables the use of Design of Experiments (DOE)[14] and Response Surface Methods (RSM)[15]. These classical statistical methods were too computationally intensive for use many years ago, but now can be routinely performed on basic desktop or laptop computers. A key feature of these methods is that many or all of the factor values are changed simultaneously, resulting in greater precision in effect estiamtion[3]. This paper utilizes the ANOVA methods of a commercially available software package from Stat-Ease[16] called Design-Expert (version 9, DX9)[17], but similar capabilities can be performed through the programming languages of R[18], Java[19], Python[20] or Matlab[21]. NASA does not endorse the use of any of these products, but they have been investigated and found to have many capabilities useful to this effort.

In general usage, DOE or experimental design is used for planning any information-gathering exercises where variation is present, regardless of whether the variation is under the full control of the experimenter or not[14]. However, in statistics, these terms are usually used for controlled experiments. Experimenters are often interested in modeling and analyzing the effect of some process or intervention (the "treatment") on specific objects (or people, plants, animals, or things). As such, design of experiments is thus a technique that has very broad application across all the natural and social sciences and engineering[14]. While the authors of this paper are not advocating the use of the DX9 software, this package contains many useful and necessary capabilities that can be easily accessed and employed by non-experts in statistics. Among these capabilities is a class of DOE known as optimal designs, in which the minimum number of points is used to define a specific RSM functional form[16-17]. In the DX9 software, the functional form is a single or multi-factor polynomial to which additional mathematical transformations (logarithmic, exponential, square root, etc.) can be applied. Optimal designs allow the functional form model parameters to be estimated without bias and with minimum-variance. In comparison to an optimal design, a non-optimal design requires a large number of experimental runs to estimate the parameters to the same order of precision[14, 16-17, 22]. The optimality of a design depends on the statistical model and is assessed with respect to a statistical criterion, which is related to the variance-matrix of the estimator[22]. In practical terms, optimal experiments can [significantly] reduce the cost of experimentation.

In this research, the class of D-Optimal designs is used. These space-filling designs seek to maximize the determinant of the Fischer information matrix of the design[22]. For these designs, the variance of the parameter estimations for a specified model is minimized when the determinant of Fischer information matrix is maximized[23]. Essentially, D-Optimal designs seek to optimally cover a given domain with the fewest number of data points required to create a given RSM model, while maximizing the predictive capability of the resulting RSM model. The number of points required depends on both the number of factors (independent variables) and the order of the polynomial model that is to be created[11]. Computations are used to repeatedly simulate a wide variety of possible testing campaigns. For example, as shown in Figure 1, two sample data sets (7 points each, black delta and red gradient symbols) are shown to represent the relative strength (in percent) of concrete[24]. The value of the response (dependent variable) is shown on the y-axis as a function of the factor temperature in degrees Celsius (independent variable), plotted along on the x-axis. The symbols are the actual measured relative minimum and maximum measured strength of two different types of concrete. The solid blue line is the mean value curve and the dashed blues lines are the 95% confidence interval about the mean value curve, as computed by the DX9 software. Without know the type of concrete that may be used for a given application, considerable variability in the relative strength (on the order of 40%) is possible throughout the domain. At a temperature of 600°C, the measured data have spread from about 40% to 90% relative strength.

In this work, the results of different testing strategies (ranging from fully clustered to fully distributed) are statistically analyzed and evaluated over a large number of sample sets, so as to produce statistical evidence in support of distributed testing. By using computational simulations of testing, results that would take years to obtain through actual testing can be obtained in minutes or hours. The proposed process uses a more distributed, rather than clustered, testing strategy. With the distributed testing approach, many fewer test points are required to achieve equivalent accuracies of the mean value and variability[3, 22]. In addition, RSM creates smoother, more globally representative models of many material properties, which in turn creates smoother A- and B-Basis confidence bounds, typically associated with the strength of materials[1].

To date, many hundreds of thousands of examples (perhaps approaching a total of millions of examples) have been analyzed using this computational simulation of testing process. In each simulation, comparisons were made between the distributed testing approach and the clustered testing approach. Without going into details here, numerous phases of the simulated testing process must account for statistical variability and uncertainty. In each example case, an initial truth model was arbitrarily defined to represent some material property behavior of interest; the truth model contained some inherent variability. Then, many simulated test sample draws from the truth model population were produced, using some choice of simulated distributed or clustered testing approaches. The sampled data were then fit via RSM to produce analytic surfaces representing the simulated "measured data". Finally, the RSM models were quantitatively compared to the initial truth model in order to determine the validity of each testing approach by several metrics. As each of these process steps, described in more detail in Section II of this paper, can be subject to variability and uncertainty, several replications for each phase in the simulation process were obtained. Enforced replication enables statistically sound conclusions to be drawn. The conclusions should be independent of any particular process step or implementation approach.

Over the course of more than one year, each author has independently performed similar analyses using their own methods and tools, and all have reached similar conclusions. Simulations were initially run using Microsoft Excel and the DX9 software. Later, an XML interface to the DX9 software was introduced to automate the interaction process and to allow for thousands of cases to be analyzed overnight on a single office computer. These exercises have since been repeated in the languages R, Matlab, Java, and Python; numerous scripts have been developed that can be provided to potential users. Additional automation steps and parameterization of the problem setup were also later employed. Again, NASA does not endorse the use of any of these products.

Figures 2, 3 and 4 (typical of the initial results) demonstrate a proof of concept for the distributed versus clustered testing approach. Without going into all the details at this point, these results hint at the strength of the statistically-based evidence[25, 26] that has been obtained to date for the clustered and distributed testing approaches More rigorous examinations of the clustered versus distributed testing approaches are presented in Section III of this paper. Figure 2 shows the performance of the distributed testing approach in better replicating the underlying truth model, compared with that of the clustered testing approach, for both a one-factor example, Fig. 2(a), and a five-factor example , Fig. 2(b). Each used 1000 single-factor truth models of various polynomial orders. The distributed testing approach resulted in a better approximation to the underlying truth model about 93% of the time for the one-factor truth models and about 98% of the time for the five-factor truth models. Using different color bands for distributed analyses with increasing numbers of data samples considered, Figure 3 shows the 95% confidence bands, typical of distributed testing (with the number of distributed data samples ranging in number from 9 to 75), around a truth model curve derived by applying the RSM process to the clustered data points shown with red symbols. The width of the confidence bands quickly decreases with an increasing number of distributed points considered as the bands converge to the underlying truth model. Figure 4 illustrates a comparison of the B-Basis allowables curves[1] that can be derived from the clustered and distributed testing approaches. In Fig. 4(a), the red symbols are simulated clustered data. The analytic truth model from which the data samples were drawn is also shown (solid blue line). Computed B-Basis values for each cluster are shown as the black symbols, with both piecewise linear segments connecting the B-Basis values (solid black lines) and a least squares fit through the B-Basis values (black dashed line) included in the figure. Within the materials testing community, the idea of applying RSM methods to generate B-Basis curves, rather than discrete values is quite novel, even controversial. In Figure 4(b), this idea is carried one step farther, wherein a set of fully distributed data samples (red symbols) replaces the clustered data in Fig 4(a). RSM techniques are then applied to this distributed data to obtain a smooth least squares curve that reflects an approximate truth model, and the B-Basis curve is then derived by considering the variability of all the available data around the smooth least squares curve. The resulting B-Basis offset in Fig. 4(b) is less than either B-Basis offset in Fig 4(a) and the resulting curve is smoother and more reflective of a natural behavior than the piecewise linear B-Basis boundary in Fig 4(a).

The remainder of this paper describes in detail the process description, analysis methods employed, and sample results for the process. Finally, a few conclusions are drawn from the work.

II. Analysis Methods

In this work, the entire testing process (test planning, data acquisition and data analysis) is simulated through computations. First, an initial “truth” model is constructed as a single-factor or multi-factor polynomial curve or surface to represent some physical behavior. The truth model description includes some degree of expected or imposed variability in the response. Upper and lower values for each of the factors are set, which establishes a domain in which the truth model will be analyzed. Then, a test strategy is selected. The test strategy determines how the data points will be sampled within the available single- or multi-factor domain; the determination of how the data points will be arranged in the domain may employ the DOE technique. For a single factor test scenario, (e.g., the only independent variable is temperature) the test strategy determines how the independent variable values will be chosen. The number of test points to be obtained is also prescribed. Then, factor values are selected in accordance with the selected test strategy. Response values are then computed from the truth model description, including the imposed variability; these response values are the simulated test data values for the given factor values. Having obtained a set of simulated test data at discrete locations within the domain, data pairs (factor, response) are then fit as a polynomial function of the factor values using RSM techniques. This yields an analytic, predictive model of the simulated test that can be interpolated anywhere within the domain. Having obtained an analytic model of the simulated test data, comparisons of that test model can be made with the underlying truth model to assess and quantify the goodness of fit with the truth model. In order to account for the variability and uncertainty associated with any particular truth model and any particular test model, this process is repeated over and over within a nested Monte-Carlo-based computational simulation of experimental testing. Given a sufficiently large number of simulated test samples, statistical evidence emerges to support the choice of the distributed testing approach. Each of these elements will be discussed in more detail subsequently.

A. Truth Model Creation

The process step of truth model creation creates the conditions under which the various test strategies are compared. The truth model represents the uncertain physical behavior being tested. In this work, the truth model consists of a single factor mean value polynomial curve of order less than or equal to six, surrounded by a band of plausible data variability or uncertainty. Polynomials can be used to capture many physical behaviors, and can approximate other mathematical behaviors, such as exponentials, over a limited domain; additional transformations to the polynomial form are also available. The first step in generating the truth models is making a choice of the polynomial order; this determines the number of terms in the polynomial. The mean value curve polynomial coefficients for those terms are then randomly chosen from a uniform distribution of the interval [-70, 70]. For example, a third-order truth model (cubic) polynomial response (y) in one factor (x) may be given by the equation:

$$y = 52.37 - 19.13x + 5.18x^2 - 20.44x^3 \qquad (1)$$

The truth model mean value curve given by Eq. 1 is deterministic; it has no inherent uncertainty or variability. In this work, the data variability or uncertainty about the mean value is represented separately from the deterministic portion. For simplicity, the data variability or uncertainty is represented by a normal distribution, though other choices are possible and should be investigated. A normal distribution requires two mathematical parameters, a mean value and a standard deviation. The data variability or uncertainty is typically expressed as a coefficient of variation (COV), a nondimensional measure of the data variability. The COV is equal to the standard deviation divided by the mean value. This ratio is typically multiplied by 100 so the COV can be expressed as a percent value. Experimentalists sometimes claim to have knowledge about much variation might be expected for a given data set. A typical assumed value for the COV might be 10% to 20% (reflective of some early design margins of safety) but in this study the COV varies from 1% to 100%.

Given a polynomial form and a set of polynomial coefficients (such as in Eq. 1), the deterministic portion of the truth model is evaluated for perhaps 100 randomly chosen factor values; these samples can then be averaged to determine an estimate of the overall functional mean value (μ). Alternatively, other analytic means can be used to obtain an estimate of the overall functional mean value. This overall functional mean value is then multiplied by the COV in percent, and divided by 100 to obtain an estimate of the standard deviation (σ), as shown in Eq. 2 below

$$\sigma = \mu * \frac{COV\%}{100} \qquad (2)$$

The data uncertainty or variability of the truth model is then added to the deterministic portion of the truth model as shown in Eq. 3 below

$$\text{Uncertain Response} = \text{Deterministic Response} + \text{Normal}(0, \sigma) \qquad (3)$$

The deterministic response in obtained from an equation such as Eq. 1. Upper and lower bounds are prescribed to define the domain over which the truth model (including uncertainty or variability) will be evaluated. Typical bounds for a factor in this study might be [0, 1] or [-1, 1]. Other bounds have also been considered.

B. Testing Strategy Selection

The testing strategy defines how the test points will be distributed within the domain defined by the factor upper and lower bounds. In defining the testing strategy, there are a few competing elements to consider: the number of factors involved in the test, the expected polynomial functional form (if any form is expected), the total number of test samples that can be obtained, the need for global coverage of the domain, and the need for replicate points; replicate points are those obtained at the same factor values (or combination of factor values) as one or more other previous points. The number of factors and the expected polynomial functional form work together to establish a lower bound on the number of samples that need to be acquired. Alternatively, the number of factors and the number of total samples that can be acquired work together to establish an upper bound on the polynomial forms that can be created. The need for global coverage of the domain is greater for nonlinear models than for linear models. The need for replicate points is to establish an important independent process uncertainty metric for the overall repeatability of results.

Domain coverage can be achieved in one of several ways, including:

- Completely random distribution
- D-optimal DOE distributions[17] and non-optimal DOE distributions
- Test points obtained at a pre-defined list of factor values or combinations, either with or without prior knowledge of the expected functional behavior; a uniformly-spaced rectangular grid of points is a common implementation of this option.

The various test strategies are illustrated in Figure 5 with different numbers of test samples for a one-factor domain. Beginning from the top (line marked “1”), seven points are randomly generated to cover a certain domain. Notice that the domain covered is of a lesser extent than the remainder of the point sample sets. This is because, for randomly generated samples, there is no guarantee that the end points of the domain will be covered. Moving downward, the second data set illustrates a D-Optimal DOE[17], also with seven points; notice how the data points are spaced irregularly and more toward the bounds of the domain, as is typical of these designs. The bounds of the domain are automatically covered. The third curve illustrates uniform spacing of seven points in the same domain with bounds coverage. The next three data sets (labeled 4, 5 and 6) illustrate the effect of increasing the sample size with D-Optimal designs; sets with 9, 11 and 13 data samples are shown, respectively. Again, the points are arranged irregularly and more toward the bounds of the domain. Data set labeled 7 illustrates a random distribution of data points with 15 samples; coverage of the domain bounds has been enforced. The next data set (labeled 8) illustrates a D-Optimal DOE with 15 samples; notice the more uniform, but still irregular coverage of the domain, compared to that of set 7. The last set (9) show a uniform distribution of 15 data points. One can observe that as more points are added to the D-Optimal DOE, that the distribution becomes more like a uniform distribution. These data sets are all shown with zero replicates; each data point occupies a unique position in the domain. One can image overlaying upon these original distributions sets of replicate points which duplicate prior test conditions. A replication strategy must be defined. The replication strategies are to randomly distribute the replicate points across the existing unique test conditions, to use some D-Optimal strategy, to uniformly distribute the replicates, or perhaps to distribute the replicates with some predefined strategy that might place more points in one part of the domain compared to another part of the domain.

Figure 6 illustrates the various test strategies with 25 samples each in a two-factor domain: (a) random, (b) D-Optimal and (c) uniformly-spaced rectangular grid patterns. The arrangement in Figs. 6(a) and 6(c) precludes accurate estimation of effects and interactions compared to that of Fig. 6(b)[3]. Again, the degree of replication can range from zero to compete. In the former scenario, no replicates are obtained. In the latter scenario, every test point is a replicate of at least one other test point. These different testing strategies, coupled with the degree of replication, can be alternatively considered as simply distributed [Fig. 6(b)] or clustered.

C. Testing Design Creation (DOE)

Design of experiments is the design of any information-gathering exercises where variation is present. Formal planned experimentation is often used in evaluating the effect that one or more factors have on physical objects, structures, components, and materials. The DOE methods rely upon statistical methods: ANOVA and RSM. Design of experiments has very broad application across all the natural and social sciences and engineering.

Testing design creates the objects of comparison: distributed, clustered, and in-between testing designs. First, the total number of data points available for testing is chosen, which in this study varied from 10 to 1000. Keeping the total number of data points constant, the way these data are distributed in the testing domain is varied. Completely distributed deigns do not contain replicated experiments; in other words, distributed designs contain clusters of one sample each. Completely clustered designs consist of clusters at the extreme points of the testing domain, in other words, at the maximum and minimum value of interest in each factor. The number of clusters is gradually increased to create a spectrum of in-between designs, equally replicating each experiment (clusters of equal numbers of samples). For example, for 10 total test points in one factor, the designs tested are 2 clusters of 5 experiments, 3 of about 3, 5 of 2, and 10 of 1. Gradually varying the number of clusters in the test designs allows for the amount of clustering to be evaluated, instead of only comparing the two extremes. The testing conditions (i.e., cluster conditions) are D-optimally distributed throughout the design space using the built-in MATLAB function **cordexch**[21], which for high numbers of clusters is approximated by uniformly spacing the clusters [e.g., Fig. 2(c)]. For a certain total number of data points, each design is simulated against the same ensemble of 1,440 truth curves in order to ensure any differences in the validation step are due to the design and not the underlying truth model.

When performing testing in a laboratory, some knowledge of the truth model may be available, for example, when testing a physical behavior that is known to be quadratic with respect to a certain factor. Using this knowledge can be advantageous, because in order to produce a D-optimal design, some guess of the truth model polynomial order is needed. By keeping the truth model creation separate from the DOE creation, these simulations represent the worst case scenario: when no outside knowledge is available. Similarly, knowledge of the polynomial order of the truth model may influence the choice of the polynomial order of the surrogate curve fit to the data by RSM (discussed subsequently), but such influence is not present in these simulations.

While the D-Optimal design uses the minimum number of test points to define a specific functional form through RSM, the value of the D-Optimal DOE is somewhat limited when only a single factor is present. The true value of D-Optimal designs increases exponentially as the number of factors increase. For example, to define a cubic RSM form in one dimension requires four data points. To define a cubic form in two dimensions using a D-Optimal DOE requires 10 and in three dimensions requires 20 points. If a rectangular grid of points were instead used to define the cubic form, it would require 16 points in two dimensions and 64 points in three dimensions. Thus, the D-Optimal approach requires just 62.5% of the testing for two factors, and just 31.25% of the testing for three factors, compared with a more traditional rectangular grid pattern of testing.

A two-dimensional D-Optimal DOE is shown previously in Fig. 6. In the following example, two factors are present: temperature and a material property ratio. The factor ratio represents the numerical ratio of two dimensional parameters. Just nine test samples were to be used; this enables only a quadratic RSM form to be defined, but provides three data samples to be used as lack-of-fit tests to increase the user's confidence in the resulting model. A traditional testing approach might simply request a uniformly-spaced rectangular grid pattern for the test. Further discussion with a subject matter expert (SME) for this test revealed a very limited interest in acquiring test samples near the lower bound of either factor. Hence, a constrained two-factor D-Optimal design, shown in Figure 7, was created to more evenly spread the test points over just the desired test domain. Here, both physical factors have been replaced with normalized factors that each occupy a range from zero to unity, but this normalization is not required. The color contours in Figure 6 are meaningless and are simply shown to illustrate a potential nonlinear physical behavior. The reader should note how the requested test points are skewed away from the lower left hand corner of the test domain because the DOE was constrained to do so.

The constrained, two-factor D-Optimal DOE was then further modified by the DX9 square root transformation. For values between zero and unity, the square root transformation moves data points closer to unity. For example, a requested test point at NormTemp = 0.5 and NormRatio = 0.3 was replaced by one with coordinates of about (0.71, 0.55) in the transformed and normalized space; these values are the square root of each component. The reader should notice how the proposed test points are now even more skewed toward the upper right corner of the figure and away from the lower left corner of the figure. This arrangement was found to be highly desirable by the SME for the test being considered. Again, the color contours in Figure 8 are meaningless and are simply shown to illustrate a potential nonlinear physical behavior.

D. Data Collection

Simulated test response data is generated from the uncertain truth model (the mean value truth model curve with normal distribution of the specified COV wrapped around the truth model curve). This is an intrinsically random process, which approximates the inherent variability in a batch of actual test material. To account for random effects, the data collection step (and the subsequent surrogate model creation) was repeated ten times for each combination of the truth model and the testing design.

E. Surrogate Model Creation

Surrogate model creation relies on information only from the data collection step, the response data. Two fits were created for each simulated data set: a least-squares polynomial regression fit and a piecewise linear interpolation fit. The built-in MATLAB function **stepwise**[21] was used to select the terms in the polynomial to be fit using regression. This function was initialized with a quartic polynomial, to which the function subsequently adds or removes terms in the polynomial based on the sum of squared error between the data and the fit curve. To create the linear interpolation fit, the mean of each cluster was calculated and a piecewise linear fit was created through these means. This method is commonly employed for aerospace database generation and interrogation[27].

F. Validation of the Surrogate Model

The validation step quantifies the closeness of the surrogate model to the truth model via several metrics. These measures consist of: 1) comparing how close the surrogate curve is to the truth curve, 2) comparing the sample population COV with the true population COV, and 3) comparing the properties of allowables with their definition.
The primary measure of validation is the root mean squared error (e_{RMS}) between the surrogate curve and the truth curve[28]. For surrogate curve $\hat{y}$ and truth curve y over a domain A, this is given exactly and approximately (for N evaluation points) by:

$$e_{RMS} = \sqrt{\frac{1}{|A|}\int (\hat{y}-y)^2\, dA} \approx \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \qquad (4)$$

The e_{RMS} is always nonnegative, and naturally a better surrogate model has a lower error. The units are the same as the response y, but e_{RMS} may be non-dimensionalized by dividing by the standard deviation of the truth model. Loosely, the e_{RMS} can be thought of as the average distance from the surrogate curve to the truth curve, but squared to always be nonnegative. A secondary nondimensional measure is the error in the estimated variability (COV):

$$e_{COV} = \frac{|COV_{surrogate} - COV_{truth}|}{COV_{truth}} \qquad (5)$$

A better surrogate model will have near zero COV error, e_{COV}. The estimation of the COV in the surrogate model is important because it is used to create confidence intervals and tolerance bounds, such as the A-Basis and B-Basis allowables. The properties of the A-Basis and B-Basis allowables are also checked directly. The fraction of the true population above each allowable is checked using the COV of the truth model. The confidence in each allowable is checked by comparing the population requirement with the actual population along each surrogate curve. In addition, the fraction of simulations in which the order of the surrogate polynomial matches the order of the truth polynomial is recorded.

G. Data Augmentation

Data augmentation is the process of adding new data points to an existing data set. Again, using the ANOVA technique to create a distributed template of new points is favored, unless specific analyses are performed which suggest that the new data points should be added in specific regions of the domain to accommodate deficiencies in the existing data set. Such deficiencies might be found where: (1) RSM gradients are large, (2) RSM residuals (the difference between actual and predicted data values) are large, or (3) where RSM leverage values (the degree of influence a particular data point has on the resulting response surface) are large. The residuals can take on any positive or negative values; these should be small in absolute value. The leverage of a given point is a positive definite number, usually between zero and unity; it is best if all the leverage values for a given data set are nearly equal to unity. A very successful means to identify where new data points should be added to an existing data set is also enabled through the use of ANOVA and RSM as follows. For each data point, multiply the absolute value of the residual [Abs(R)] times the leverage value (L) plus unity. This provides a metric (M) for each data point that gets larger as the need for new data increases, M = Abs(R) * (L+1). The data points are then ranked from largest to smallest in the value of M. New data points should be added between the cases where M is the largest. For example, if there are twenty data points in an existing data set, each corresponding to a unique factor value, suppose the metric value M is largest for point 7 and 8. New data point should first be added into this gap region of the domain. A more complete discussion of this technique is beyond the scope of this document, but has been used successfully in the past[10, 27].

III. Results

The preliminary results for this study (shown in Section I) used specific clustering strategies for relatively small sample sizes (tens to hundreds of points). This work was later extended and generalized through parameterization to consider a variable degree of clustering over much larger sample set sizes (up to 1000 data samples). In addition, several validation metrics, including those described by Eqs. 4 and 5, were considered.

The e_{RMS} (Eq. 4 above) for each response curve was calculated, and compared between the two analysis methods. Figure 9 shows the percent of simulations where the regression curve had a lower e_{RMS}, graphed against the number of clusters in the testing design. Each colored line represents the total number of samples in a given design. When only two clusters are present in a testing design, the two methods of analysis are equivalent because both fit a straight line, which is graphed as 50%. For any other number of clusters, and any number of total samples, the regression analysis is always on average better than the linear interpolation analysis, as shown by a value greater than 50%. For designs with a high number of clusters, greater than 10, regression is better than linear interpolation more than 90% of the time. The advantage of using regression is more pronounced with a larger number of total samples, as can be seen from the line representing 1000 points quickly approaching 100%, while 10 and 30 point designs have a regression fit that is better around 70% of the time.

Notable exceptions where linear interpolation in fact was better more often are designs with 3 to 6 clusters with high COV and relatively low total number of points, such as 10 points in 5 clusters at 25% COV and 100 points in 4 clusters at 100% COV. These cases are noisy and difficult to fit with a polynomial surrogate model in general.

Observing Figure 10, the primary validation measure e_{RMS} is graphed, summarizing the all the simulations. The total number of samples is indicated by the different colors sets. Dashed lines shown represent the linear interpolation surrogate model, and the solid lines the regression. A better surrogate model will have fit curve that is near the truth curve, which is equivalent to a small RMS error. Since the underlying truth model in all cases was smooth, the interpolation surrogate model is better than linear interpolation. This trend holds not only for the average over varying truth order and varying COV, but for any given combination of truth order and COV. In sum, given any testing design, the average regression fit is better than the average linear interpolation fit.

The secondary measure of validation e_{COV} is shown in Figure 11. The total number of samples is indicated by the different colors sets. The dashed lines shown represent the linear interpolation surrogate model, and the solid lines the regression. A better surrogate model will be able to estimate the true COV with small error on average. As for e_{RMS}, the average regression estimate of the variability is less erroneous than the average linear interpolation estimate. This conclusion holds for the average over varying truth orders and varying COVs, and for any specific combination.

Allowables, such as the B-Basis and A-Basis[1], are intended to provide a reliable bound on a material property in the presence of natural variability. Allowables are typically applied to the strength of materials and might be applied in other situations. A typical allowable for material strength, the B-Basis, states that 90% of the material samples exceed the allowable value to 95% confidence; similarly, the A-Basis (used for human rated designs) states that 99% of the material samples exceed the allowable value to 95% confidence. The allowable statement has two portions: the population percentage (90% or 99%, respectively) and the confidence level with which the population meets the allowable (95%).

To minimally demonstrate compliance with a given B-Basis allowable value, one might use 20 sample sets of ten samples each (200 total material samples). It would be expected that for 19 of the 20 sample sets, the measured strength of nine samples in each sample set, would exceed the stated allowable value. Hence, the measured strength for nine of the ten samples (90%) exceeds the allowable value 19 out of 20 times (95% confidence). A much more meaningful demonstration of compliance with the same B-Basis allowable value might employ 200 samples sets of 100 samples each, for which it would be expected that the measured strength exceeds the allowable value for 90 samples from each sample set, 190 out of 200 times. Appropriate statistical factors have been defined to allow for estimation of A- and B-Basis values with many fewer samples[1]. The more samples in the experiment, the more information that is available.

Figure 12(a) illustrates the percent of the population above the B-Basis value. The total number of samples is indicated by the different colors sets. The solid lines represent the regression surrogate model and the dashed line the linear interpolation surrogate model. For distributed designs, as the total number of samples increase, the average population decreases towards 90%. Notice that this limiting effect is not reliably present for linear interpolation fits. For highly clustered designs, the population above the B-Basis of linear interpolation fits approaches a value lower than desired. And for distributed designs, an extremely high number of samples would be needed to see this limiting effect.

Test planning is the process of choosing how and under what conditions to take data. This process of allocating data collection should address the objectives of taking the data. Common objectives may be to identify the conditions where the response is minimized, or to identify if a factor has a significant effect on the response. When more than one factor is being investigated, DOE provides a systematic method for planning testing. As demonstrated above, the RSM of least squares regression produces a more accurate surrogate model. Discussed below is how to best allocate data collection to produce the most accurate regression surrogate model.

Observing Figs. 10 and 12, the validation measures graphed for regression exhibit drastic behavior for designs with a low number of clusters. These clearly indicate that highly clustered designs are far worse than distributed testing designs. In Fig. 10, the average RMS error of designs with less than eight clusters is greater than the error in distributed designs of the same total number of samples, with a drastically greater error in designs of two or three clusters. In Fig. 12(a), the average population above the B-Basis levels off for designs of 8 clusters or more. For clustered designs, the average population was lower, for most below 90%, than the intended value. Similarly in Fig. 12(b), the confidence in the B-Basis is lower for clustered deigns of less than eight clusters and levels off for distributed designs. The linear interpolation procedure incorrectly overstates the confidence level.

The testing design limits what order surrogate models can be fit to the data. If only a few unique response values are sampled, only low order models can be fit using regression. For example, if the underlying physical behavior is a cubic polynomial with respect to two factors ($y = C_0 + C_1a + C_2b + C_3ab + C_4a^2b + C_5ab^2 + C_6a^3 + C_7b^3$, for factors a and b), in order to fit a cubic polynomial to the response data, the minimum number of required distributed clusters is just eight, one for each of the polynomial coefficients (C_0, C_1, C_2, …, C_7). The total number of test samples, however, must be at least one greater than the minimum number of clusters, since one additional data point is needed to estimate the variability.

Using the minimum number of clusters to fit the expected order model is dangerous because it inhibits any higher order behavior to be identified. Design-Expert, for example, always recommends including lack-of-fit points (also called exploratory points) in an optimal testing design[17]. Lack-of-fit points are data taken under new experimental conditions other than the minimum needed to fit the expected model. Not distributing data collection results in poorly fitted curves with a large e_{RMS}, unacceptably low population above allowables, and unacceptably low confidence in allowables and confidence intervals. In Fig. 13, the error and the allowable metrics are shown for a fourth order truth model. The total number of samples is indicated by the different colors sets. Note that for designs with less than five clusters, these statistics indicate a far worse fit. This indicates that regression could not fit a fourth order polynomial and missed major information about the truth curve, which is evident in both graphs. The RMS error is substantially greater for testing designs with two, three or four clusters than designs with five or more clusters, which have a stable average RMS error for a given total number of samples. Similarly the average population above the B-Basis is stable for designs of the same total number of samples and greater than 5 clusters. Designs of two, three or four clusters produce a B-Basis with an average population of about 90% or less, which as a result have confidence far below 95%. However, the statistics for the linear interpolation fit are also inappropriately valued at low numbers of clusters, indicating that regression is not the problem. In fact, there is simply not enough information to properly fit an accurate curve with a highly clustered testing design.

When a model of a lower order than the true underlying behavior is fit to a data set, the full curvature of the truth curve cannot be captured. As a result, the surrogate curve may differ from the truth curve by omitting peaks or valleys or by straightening curvature present in the truth curve. When too few clusters are present to identify these higher order behaviors, no indication of this missed information is available. Whereas when using a distributed design, many more unique factor values are sampled than the minimum needed to fit the order of the truth curve. However, especially when using a low number of total samples in a distributed design, identifying higher order behaviors is the process of discerning the spread of the response values due to the intrinsic population variance and that due to the changing mean value. When some of the intrinsic variability is misidentified as higher order behavior (over-fitting), the error in the COV is increased (and the expected value of the COV is decreased (not shown)). This effect is seen in Fig. 11 for the distributed designs of 10 and 14 total number of samples. In practice, human inspection of the data sets should reduce the COV error. For low total number of samples, distributed designs are still preferable because the RMS error is significantly lower.

IV. Conclusions

Computational simulations of two distinct testing strategies, distributed and clustered, are compared. For clustered testing, many test points are acquired at a few different conditions. For distributed testing, only one or a few test points are requested at many different conditions. The statistical techniques of Analysis of Variance (ANOVA), Design of Experiments (DOE) and Response Surface Methods (RSM) are applied for the distributed test planning, data analysis and test augmentation. A statistically-based argument is advanced to suggest that the distributed testing approach advocated in this paper is superior to the clustered testing approach for most applications. The evidence to support the claim is based upon the work of numerous researchers using different metrics to achieve the same results.

In preliminary results, the distributed approach outperformed the clustered approach in recovering the underlying truth model at least 93% of the time. The 95% confidence bands around RSM fits to distributed data sets with as few as 9 points converged quickly to the underlying truth model defined by 75 data points obtained via the clustered approach. The B-Basis curves defined by the distributed approach are smoother and have less offset from the mean value curve than those obtained using the clustered approach.

In later work, the percent of data set simulations where the regression surrogate curve has a lower e_{RMS} than the linear interpolation surrogate curve rapidly approaches 100% as a more distributed sampling approach is employed. A lower e_{RMS} value is obtained as the designs become more distributed. The percent error in COV is smaller for a given design using the distributed approach, compared to the clustered approach. The percent of population meeting the allowable definition is better for the distributed approach than the clustered approach. In short, all the experiments performed support the conclusion that the distributed testing or sampling approach is superior to the clustered approach.

Acknowledgements

This work was performed under the sponsorship of Dr. David Glass. The project was conceived by Lawrence L. Green and carried out by students, Weston Barger, Mylan Cook and Katherine Maxwell, spanning a more than a year of effort. Each student emphasized various aspects of the project; the process and insights evolved over time. The results from each student's work have been assimilated to create this paper. Dr. Wallace Vaughn has been a continuing consultant throughout the project.

References

1. Department of Defense, "COMPOSITE MATERIALS HANDBOOK, VOLUME 1. POLYMER MATRIX COMPOSITES GUIDELINES FOR CHARACTERIZATION OF STRUCTURAL MATERIALS", MIL-HDBK-17-1F, 17 JUNE 2002, Superseding MIL-HDBK-17-1E, 23 January 1997.
2. The Composite Materials Handbook website: https://www.cmh17.org/HOME/Intro.aspx (accessed 05/11/2015).
3. The one-factor-at-a-time method of testing wikipedia webpage: http://en.wikipedia.org/wiki/One-factor-at-a-time_method (accessed 05/13/2015).
4. Top500 List, November 2002: http://www.top500.org/list/2002/11/ (accessed 05/11/2015).
5. Top500 List, November 2014: http://www.top500.org/lists/2014/11/ (accessed 05/11/2015).
6. The Analysis of Variance wikipedia website: http://en.wikipedia.org/wiki/Analysis_of_variance (accessed 05/11/2015).
7. Fisher, Ronald A., "Statistical Methods for Research Workers", ISBN 0050021702, Oliver and Boyd, Edinburgh England, 1925.
8. John, Robert, "Elementary Statistics (7th Edition)", Duxbury Press / Wadsworth Publishing Company, Cincinnati, OH, 1996.
9. Carnahan, Brice; Luther, H. A.; and Wilkes, James O., "Applied Numerical Methods", John Wiley & Sons Publishing, New York, NY 1976.

10. Green, L. L.; Rickman, S. L.; Remark, B. J.; Vander Kam, J. C.; Kowal, J.; Johnson, K. L.; and Bouslog, S. A.: Orion Thermal Protection System Margin Study, Phase 2, NASA/TM-2014-218262, May 2014.
11. Green, L. L.: The Challenges of Credible Thermal Protection System Reliability Quantification, Paper #77, Session 8B in Cross-Cutting Technologies III, The 10th International Planetary Probe Workshop (IPPW-10), San Jose, CA June 2013.
12. Green, L. L.: Uncertainty Analysis of Historical Hurricane Data. AIAA 2007-1101. Presented at the 45th AIAA Applied Aerospace Sciences Conference and Exhibit, Reno, NV, January 8-11, 2007.
13. Green, L. L.; Cruz, J.: Uncertainty Analysis for a Jet Flap Airfoil. Paper Number SISO-06F-SIW-010. Published at the Simulation Interoperability Standards Organization, 2006 Fall Simulation Interoperability Workshop, Orlando, FL, September 10-15, 2006.
14. The Design of Experiments wikipedia website: http://en.wikipedia.org/wiki/Design_of_experiments (accessed 05/11/2015).
15. The Response Surface Methodology wikipedia website: http://en.wikipedia.org/wiki/Response_surface_methodology (accessed 05/11/2015).
16. The Stat-Ease, Inc. website: http://www.statease.com/ (accessed 05/11/2015).
17. The Design-Expert (version 9) website: http://www.statease.com/dx9.html (accessed 05/11/2015).
18. The R Language Project website http://www.r-project.org/ (accessed 05/11/2015).
19. The Java Language website: https://www.java.com/en/download/faq/whatis_java.xml (accessed 05/11/2015).
20. The Python language website: https://www.python.org/ (accessed 05/11/2015)
21. The Matlab language website http://www.mathworks.com/products/matlab/ (accessed 05/11/2015).
22. The Optimal Design wikipedia website: http://en.wikipedia.org/wiki/Optimal_design (accessed 05/11/2015).
23. The Engineering Statistics Handbook website (5.5.2.1 D-Optimal Designs): http://www.itl.nist.gov/div898/handbook/pri/section5/pri521.htm, accessed 10/14/2015.
24. Naus, D. J.: A Compilation of Elevated Temperature Concrete Material Property Data and Information for Use in Assessments of Nuclear Power Plant Reinforced Concrete Structures. NUREG/CR-7031, ORNL/TM-2009/175, Oak Ridge National Laboratory, 2010.
25. Barger, Weston; Glass, David E., and Green, Lawrence L.: Using Optimal Experimental Design to Improve Testing Procedures. Talk presented at the 39th Annual Conference on Composites, Materials, and Structures, Cocoa Beach, FL, January 26, 2015.
26. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.
27. Pamadi, Bandu N., et. al., “Aerodynamic Modeling and Database Development of the Space Launch System Booster Separation”, AIAA 2015-0779. Presented at the AIAA SciTech 2015 Conference, San Diego, CA, January 4-8, 2015.
28. Picheny, V., Improving Accuracy and Compensating for Uncertainty in Surrogate Modeling, Ph.D. thesis, University of Florida, 2009.

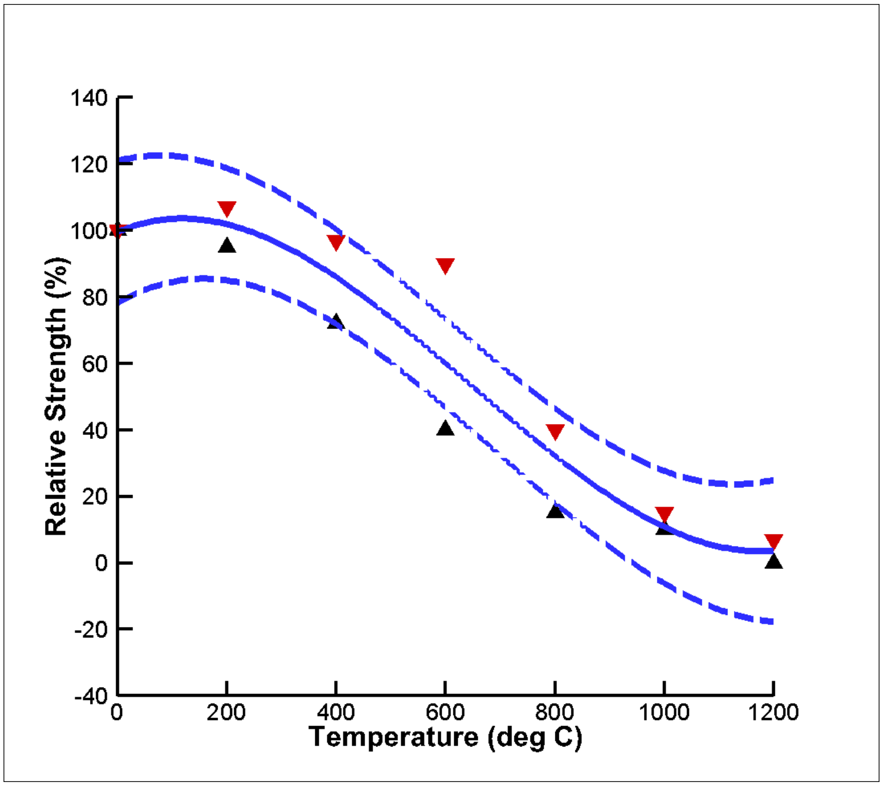
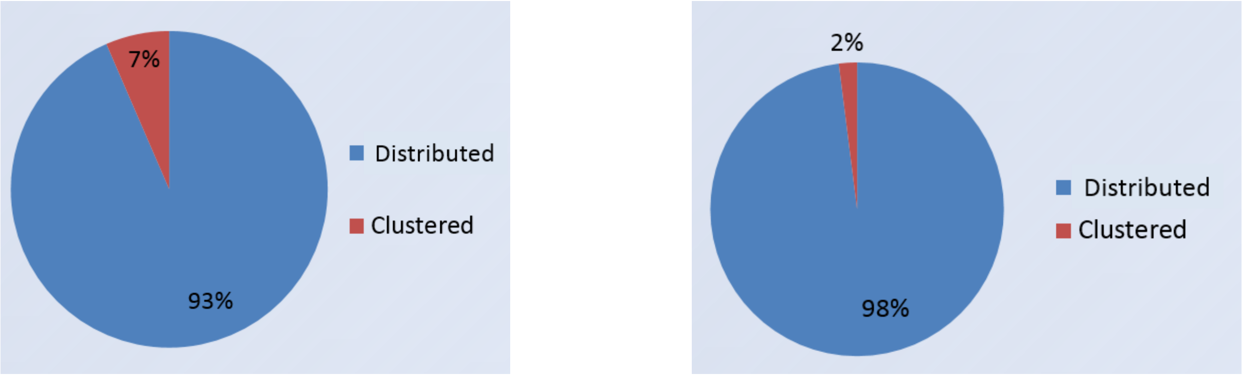


Figure 1. Sample data sets representing the relative strength of concrete as a function of temperature.



(a) One-Factor Example

(b) Five-Factor Example

Figure 2. Relative performance of the distributed and clustered testing approaches in predicting the underlying truth models.

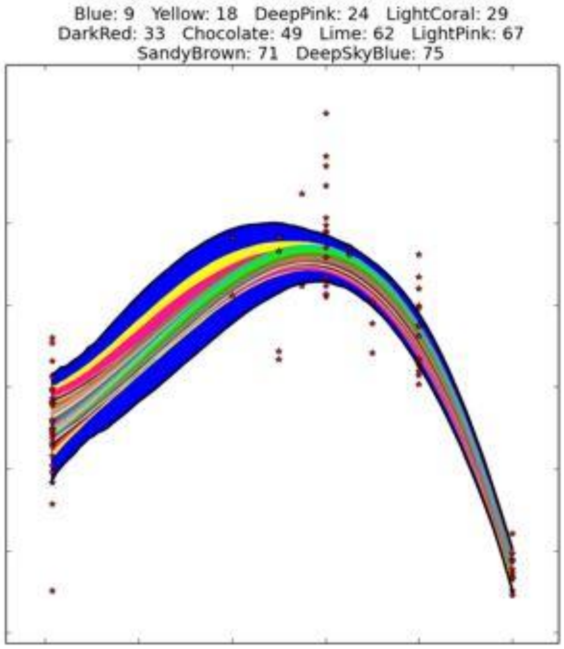
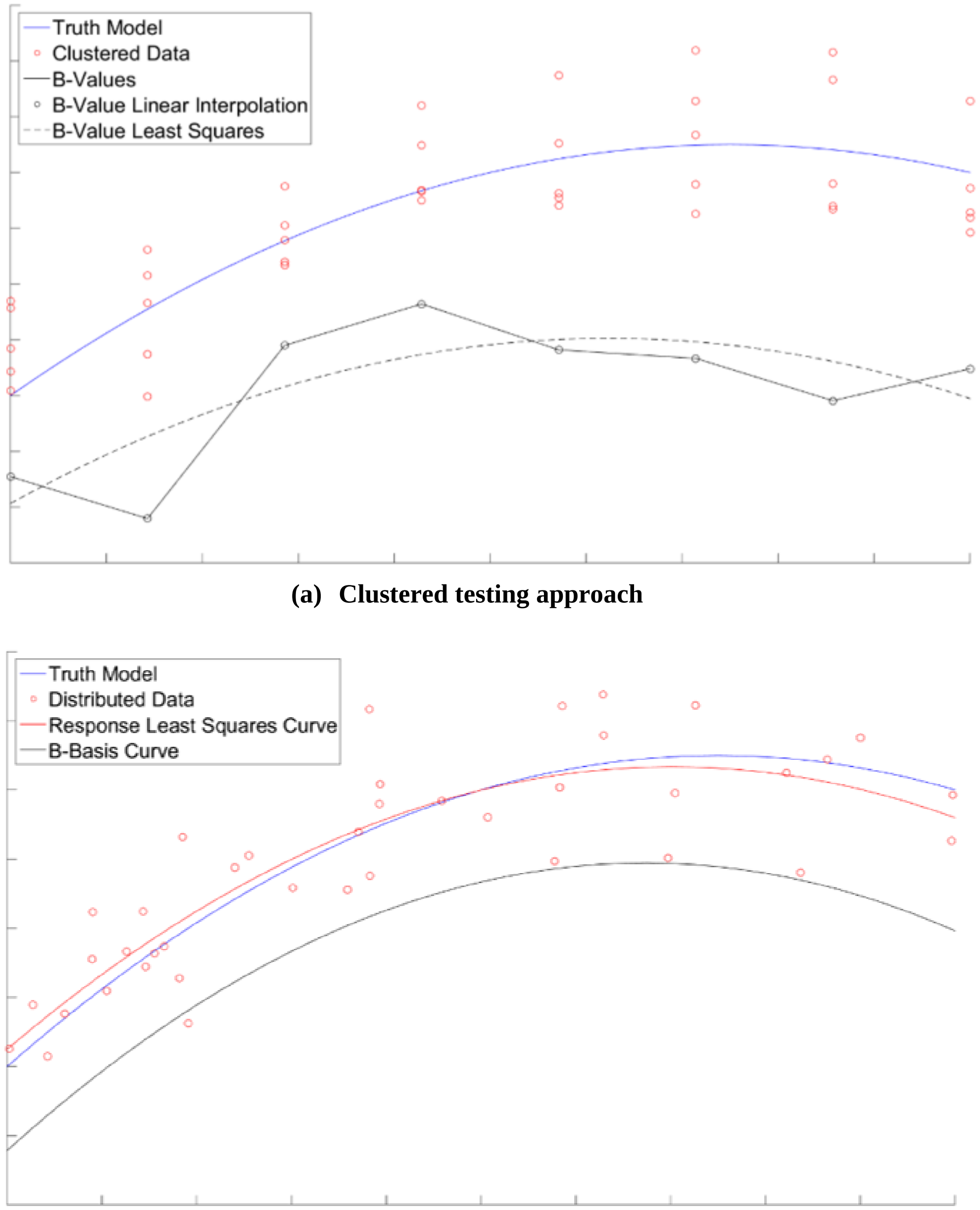


Figure 3. Comparison of the 95% RSM confidence bands for various numbers of data points obtained using the distributed testing approach, compared to the original clustered testing approach.



(a) Clustered testing approach

(b) Distributed testing approach

Figure 4. Comparison of B-Basis allowables curves derived for the clustered and distributed testing approaches.

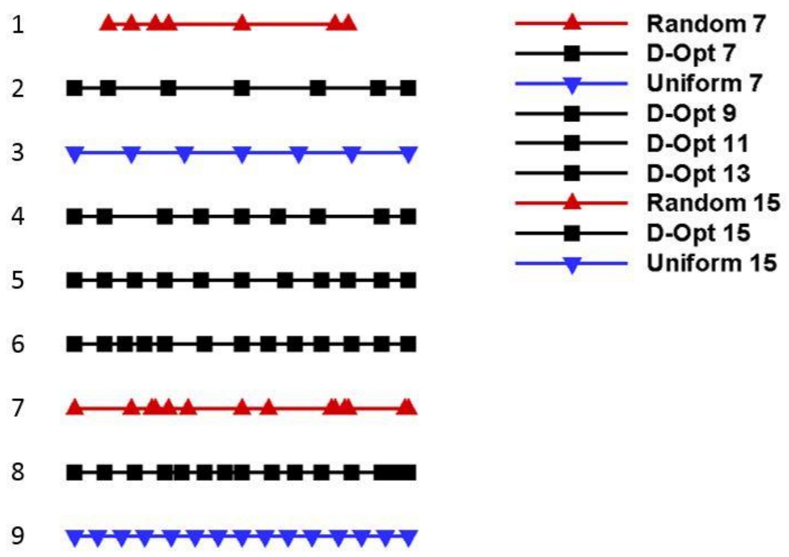
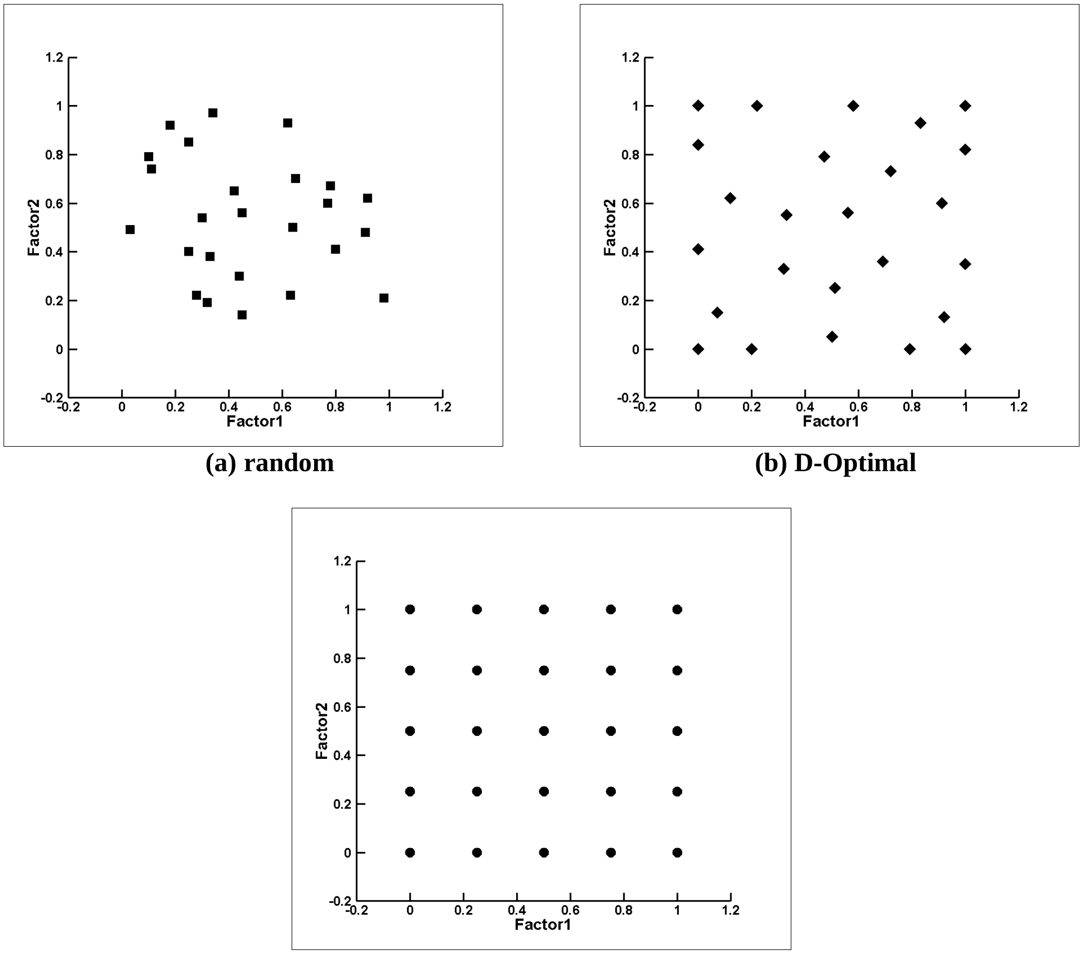


Figure 5. Illustration of various point distribution (test) strategies (one factor).



(a) random

(b) D-Optimal

(c) uniform rectangular grid

Figure 6. Illustration of three distribution patterns (two factors).

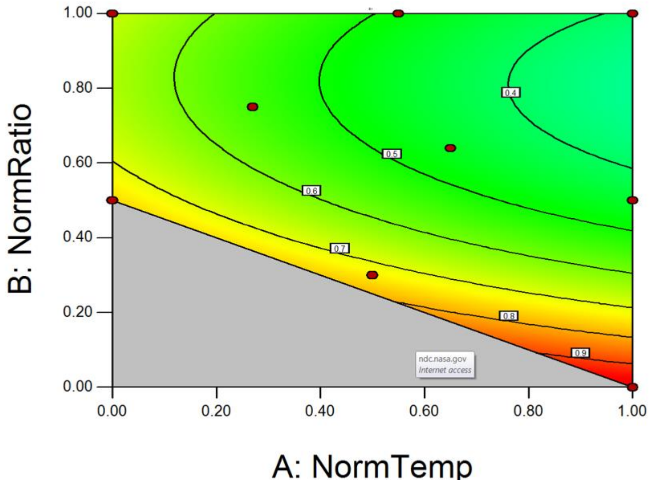


Figure 7. Constrained D-Optimal DOE for two-factor example.

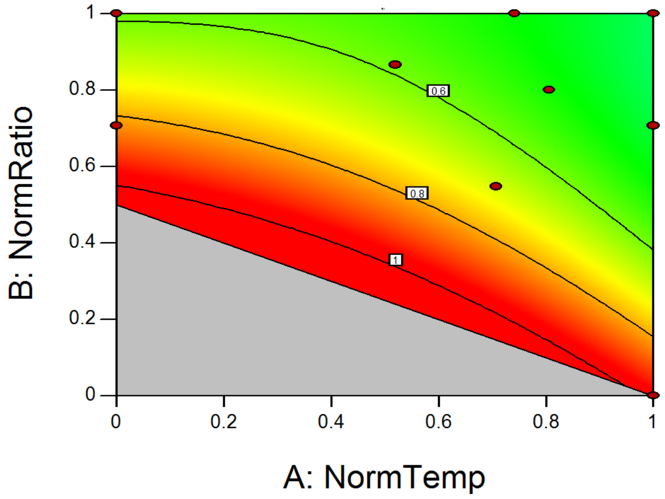


Figure 8. The transformed, two-factor, D-Optimal DOE.

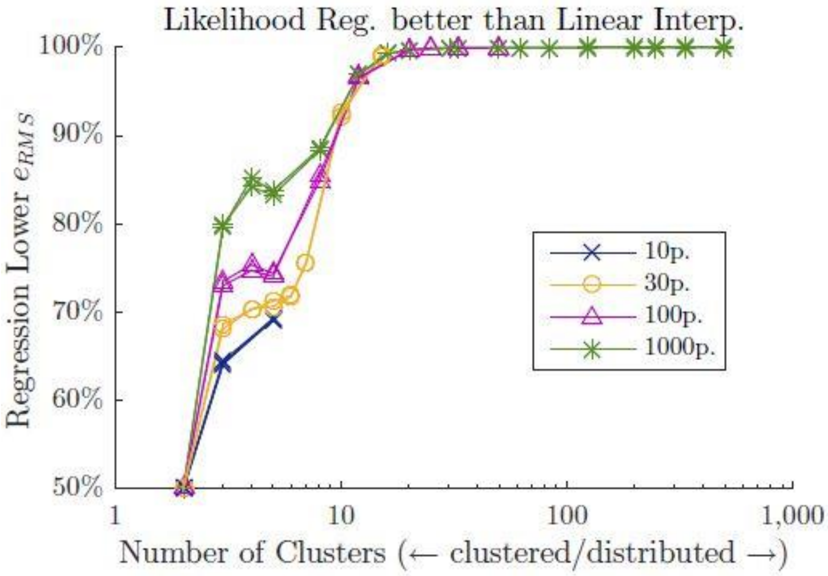


Figure 9: Percent of data set simulations where the regression surrogate curve has a lower e_{RMS} than the linear interpolation surrogate curve.

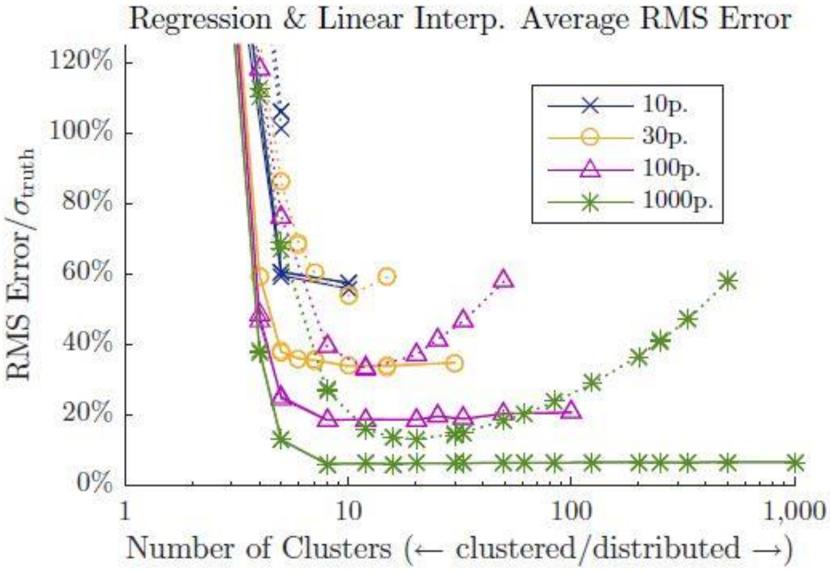


Figure 10: The e_{RMS} shown plotted against number of clusters for several sample size cases.

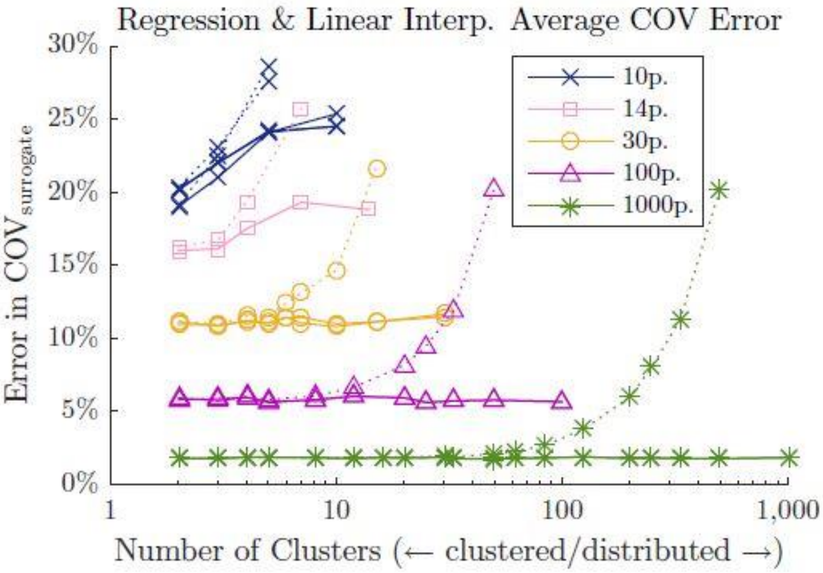
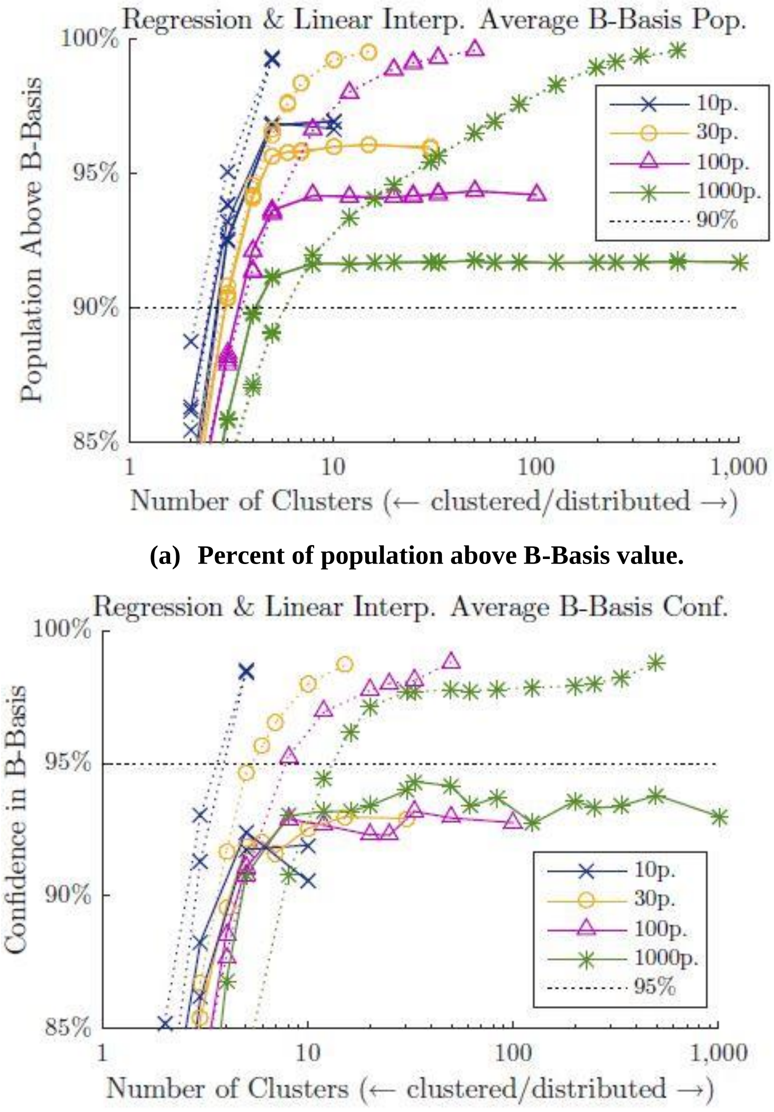


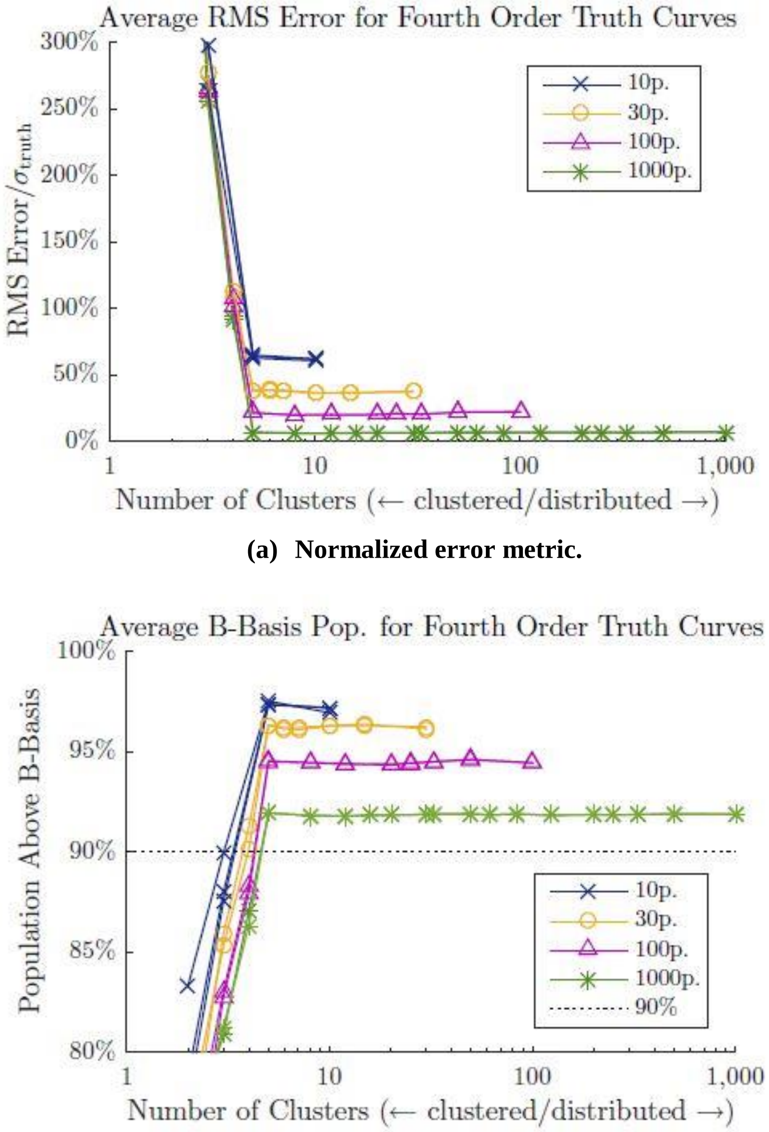
Figure 11: The percent error in the COV shown plotted against number of clusters for several sample size cases.



(a) Percent of population above B-Basis value.

(b) Confidence in B-Basis estimate.

Figure 12: Average B-Basis population and confidence shown plotted against number of clusters for several sample size cases.



(a) Normalized error metric.

(b) Percent of population above B-Basis value.

Figure 13: Graphs for fourth order truth curves.