SCIENTIFIC TEST
& ANALYSIS
TECHNIQUES
CENTER OF EXCELLENCE

Elements of a Mathematical Framework for Model Validation Levels

August 2022

Kyle Provost, Contractor
Corinne Weeks, Contractor
Nicholas Jones, Contractor
Maj Victoria Sieck, PhD

The STAT COE provides independent STAT consultation to designated acquisition programs and special projects to improve Test & Evaluation (T&E) rigor, effectiveness, & efficiency.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, CoE@afit.edu
Call, 937-255-3636 x4736

Scientific Test & Analysis Techniques Center of Excellence
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio

The views expressed are those of the author(s) and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government.

Version: 2, FY22

Modernizing *the* Culture *of* Test & Evaluation

Executive Summary

As the Department of Defense (DOD) seeks to increasingly leverage Modeling and Simulation (M&S) in the development of weapons systems, it is becoming progressively more important for models to be well understood and well vetted for use in order to mitigate the risks posed by relying on inaccurate, insufficient, or incorrectly applied models. Validation is intended as an assessment process to establish trust in models. However, validation criteria is often subjective, and is defined only to support a pass-fail understanding of validity at static points in time. This fails to provide a proper understanding of validity as requirements change, mission scope is redefined, new data is collected, or models are adapted to a new use. The metrics discussed here provide a rigorous, objective method to evaluate the trust that can be placed in a model according to fidelity, appropriate referents, and the specific intended use. These three pillars of validation will factor into a Model Validation Level (MVL), which will enable M&S to be employed in the engineering of complex defense systems with full comprehension of model capabilities and risk.

Keywords: Digital Engineering, Modeling & Simulation, Validation

Table of Contents

Introduction

Model validation in digital engineering suffers from the lack of a clear definition. Simply put, model validation compares a model to some basis to determine whether a model is "good enough", but there is no standard for how to conduct such a comparison, what constitutes an acceptable basis for comparison, or even what constitutes "good enough". Currently, these questions are frequently left to subject matter experts (SME) to answer on a case-by-case basis, resulting in subjectively defined criteria for validating a model. The lack of rigor leads to uncertainty by decision makers on the trustworthiness of a model, and uncertainty by model developers on whether a model is relevant under different use cases. In this paper, we discuss the mathematical considerations which must be addressed in a rigorous validation approach and present metrics and mathematical constructs for use in a Model Validation Level (MVL) according to the conceptual framework laid out in A Conceptual Framework for the Establishment of Model Readiness Levels (Ahner, 2021). A MVL is a single metric for the evaluation of model trustworthiness, and it enables an objective approach to model validation which numerically assesses the authority level of a model's outputs in its intended use case rather than a binary "pass/fail" criterion. In doing so, a MVL establishes not only a scalable measure of trust to be placed in a model, but also a basis for continual assessment that can grade a model consistently as new data is obtained, requirements change, or the model is adapted for new use. Through this improved validation framework, MVLs will enable DOD program managers and M&S (Modeling & Simulation) developers to effectively employ M&S in the engineering of complex defense systems with full comprehension of model capabilities and risk.

The remainder of this paper examines the major conceptual elements, or pillars, which must be addressed in validation: model-referent fidelity, referent authority, and specific intended use (i.e. scope). For each element, this paper addresses the mathematical considerations for validation and presents metrics and constructs useful for model validation. This paper additionally discusses necessary considerations for incorporating the mathematical constructs and metrics associated with the separate pillars of validation into a combined MVL. Finally, this paper reiterates the current state of M&S in the Department of Defense (DOD) and the value provided by the MVLs for rigorous, consistent, and repeatable validation.

Background

M&S is becoming increasingly integral to the system development process in the DOD. The DOD established a goal for Enterprise-level Digital Engineering (DE) across the DOD to reduce the length of the acquisition lifecycle (Deputy Assistant Secretary of Defense for Systems Engineering (DASD(SE)), 2018). DE is reliant on the use of M&S, and the trustworthiness of models is an important consideration in the design and test of complex systems.

Increased usage of DE has led to a need for rigorous methods of model validation. Such validation should be consistently applied at an enterprise level and emphasize continuous validation that allows model reassessment in response to changing data availability or changes in scope as models are adapted for new uses. Rigorous validation requires the examination of a model in terms of three pillars: sufficient fidelity, appropriate referents, and a specific intended use (Ahner et al., 2021), hereafter referred to as fidelity, referent authority, and scope, respectively.

The first pillar of fidelity addresses whether a model accurately matches reality and is perhaps the most fundamental pillar in determining the trustworthiness of a model. Second, referent authority addresses the extent to which our knowledge of reality is reliable. In practice, we do not perfectly know reality, and a model must be assessed against information sources, or referents, that themselves are inaccurate. The model inherits trust from the referents it is assessed against. In examining referent authority, we assess how much trust can be placed in these referents, and by extension how much trust can be placed in a model. Finally, the pillar of scope compares the intended use of a model to the actual realization of a model. That is, we must assess whether a model reflects the entire range of operational, environmental, or system factors that need to be modeled to support the mission. Below, we further define each pillar and the mathematical concepts that must be considered in their assessment. Furthermore, we provide metrics and mathematical constructs which provide quantitative assessments of each pillar.

Model-Referent Fidelity

An assessment of fidelity makes a direct comparison of model outputs to referent information, which represents the reality a model is intended to reflect. This comparison is the central goal of validation, and proper assessment of fidelity forms the backbone of our MVL assessment.

The detailed considerations and construction of a fidelity metric are discussed in the Scientific Test and Analysis Techniques (STAT) Center of Excellence (COE) Best Practice, Constructing a Metric for Fidelity in Model Validation (Weeks, 2022). Here, we will present a condensed look at the mathematical considerations for a fidelity metric.

To adequately capture the dimensions of accuracy, repeatability and resolution, the fidelity quantification approach uses two sub-metrics. The first sub-metric assesses accuracy, and the second sub-metric addresses repeatability and resolution as a single-variability assessment.

The accuracy sub-metric f_a is defined in Equation 1.1 and assesses the model's accuracy with respect to the referent.

$$f_a = e^{-\frac{1}{2}\left(\frac{\bar{x}_m-\bar{x}_r}{s_r^*}\right)^2} \tag{1.1}$$

In Equation 1.1, $\bar{x}_m$ and $\bar{x}_r$ are the sample means of the model and referent, respectively; s_r^* is the resolution-modified standard deviation for the referent.

The variability sub-metric f_v is defined in Equation 1.2 and assesses the similarity in variability between the model and the referent. In this case, variability refers to both the aleatory and epistemic uncertainty of the model and referent.

$$f_v = e^{-\frac{(s_m^*-s_r^*)^2}{s_m^* s_r^*}} \tag{1.2}$$

In Equation 1.2, s_m^* and s_r^* are the resolution-modified standard deviations of the model and referent, respectively. The calculation for these modified standard deviations can be seen in Equation 1.3, where s represents the standard deviation of the referent or model, and δ represents the resolution. Resolution is the degree of granularity with which a variable can be determined and is present due to any number of issues that would obfuscate our knowledge, such as simplifying assumptions in physics estimates, measurement error, or even rounding error.

$$s^* = \sqrt{s^2 + \frac{\delta^2}{12}} \tag{1.3}$$

The overall metric for fidelity is given in Equation 1.4, where the accuracy and variability sub-metrics are multiplied together to form a single metric for assessing the level of consistency between the model and the referent.

$$f = f_a f_v = e^{-\frac{1}{2}\left(\frac{\bar{x}_m - \bar{x}_r}{s_r^*}\right)^2} e^{-\frac{(s_m^* - s_r^*)^2}{s_m^* s_r^*}} \tag{1.4}$$

The multiplicative combination of metrics means that the overall fidelity metric may only be as high as the fidelity of either of its components. That is, the overall fidelity metric is strictly less than or equal to its accuracy and variability components. If either the accuracy or variability is low, the overall fidelity metric will be low as well. This results in a metric bounded from 0 to 1 where 1 implies perfect fidelity in terms of both accuracy and variability and scores close to 0 imply poor fidelity in at least one aspect.

This fidelity metric only compares a model to a single referent at a single point in the mission space or scope. As we discuss the remaining pillars of validation, we will further detail how to extend this measure to examine a model against multiple referents of varying authority as well as how to apply this measure as we assess the entirety of the model scope.

Referent Authority

Referent authority reflects the amount of trust that can be placed in a source of information, or referent, to reflect reality. Model validation aims to compare a model to the reality that the model is intended to reflect, but in practice, true reality is unknown and can only be approximately measured. Assessing trust in our information sources is a critical step in validating a model.

Unfortunately, there is no accepted standard for assessing the trustworthiness of a referent. Trust itself is not observable or otherwise directly measurable. However, referent authority can be understood in terms of relative comparison. That is, by comparing different types of referents, we can understand which referents are more or less trustworthy relative to each other. For example, expert opinion and physics predictions are generally less trusted than observed data, and observed data from a prototype is less trusted than observations of an operationally-ready system.

This understanding of trust informs the mathematical assessment of referent authority in validation. The first aspect we will discuss is a hierarchal weighting scheme allowing us to weight the fidelity assessment of a model according to the authority of the referents that the model is validated against. The second aspect we will discuss is a Bayesian method of pooling information from various referents to validate a model against the entire body of knowledge while still emphasizing the most trusted data sources.

Referent Authority Hierarchal Weighting

A referent authority hierarchy enables the weighting of fidelity by the amount of authority it inherits from the referents considered in the metric. A model that is highly consistent with a less trusted referent should carry less authority than a model that is highly consistent with a highly trusted referent. A weighting scale assigns a weight, or ceiling value, such that a model can have, at most, as much authority as the referent that it is measured against. A weighting scale should also be intuitive, with levels that represent apparent differences in authority. Finally, it should be consistently applied between different models to allow for authority comparison between models.

To meet these desired qualities, an authority weighting scale should be bounded between 0 and 1 such that a referent that best reflects reality confers an authority weight of 1 and a referent with no bearing in reality confers a weight of 0. This combines multiplicatively with the fidelity metric to yield an authority-weighted fidelity metric whose maximum value is determined by referent level as seen in Equation 2.1. Since each element of the equation is bounded between 0 and 1, we can state that $f^* \leq w$, or the authority-weighted fidelity of the model, is no greater than the authority of the referent it is measured against.

$$f^* = \mathrm{w} f_a f_v = w * e^{-\frac{1}{2}\left(\frac{\bar{x}_m - \bar{x}_r}{s_r^*}\right)^2} e^{-\frac{(s_m^* - s_r^*)^2}{s_m^* s_r^*}} \qquad (2.1)$$

We propose a 9-level weighting scale in the manner of Technology Readiness Levels (TRLs) with weights defined in Table 1 according to a geometric series descending from level 9, operational real-world data.

Table 1
Referent Authority Level Weights

AUTHORITY LEVEL	RELEVANT REFERENT	WEIGHT
1	SME Judgement	0.0183
2	First Principles/Physics Predictions	0.0302
3	Subcomponent Lab Test Data	0.0498
4	Component Lab Test Data	0.0821
5	Lab-Scale System Test Data	0.1353
6	Full Scale Prototype Test Data	0.2231
7	Production HW/SW-in-the-loop Data	0.3679
8	Live System Test Data	0.6065
9	Operational Real-World Data	1.0000

The choice to include 9 levels is largely arbitrary in terms of mathematical implications but is convenient for interpretation. Using a geometric series gives a consistent rate of reduction so that the amount of authority between levels can be consistently understood. The rate of reduction is defined as $e^{-1/2}$, and is set such that a difference of one standard deviation between the mean response of a model and a referent results in a one referent level reduction in authority. The weighting for a 9-level referent authority scale is given in Equation 2.2 where i is the referent authority level.

$$w_i = e^{-\frac{1}{2}(9-i)}, \quad i = 1, \ldots, 9 \qquad (2.2)$$

A benefit of the geometric series weighting is the heavier emphasis on higher referent levels for assessing a model. That is, consistency with high-level referents will be significantly more impactful to an MVL score than consistency with low-level referents.

Referent Pooling with Normalized Power Priors

The referent authority weighting scale enables us to consider fidelity and referent authority of a model, but there remains a question of how to measure fidelity when we have multiple referents to assess a model against. A common response may be to measure fidelity against the most trusted referent while discarding the rest. However, since no referent presents a truly perfect or complete picture of reality, a better approach is to use a method that can leverage every source of information available. Emphasis should be placed on the referents known to be the most trustworthy while still allowing other sources of relevant knowledge to refine our understanding of reality.

Bayesian statistics provides a method referred to as Normalized Power Priors (NPP) which incorporates information from various referents while weighting the impact of each referent. The output of this method will be a single mean and standard deviation, $\bar{x}_r$ and s_r, based on the referent information for input into Equation 1.4.

Bayesian statistics is a branch of statistical methods which aims to mathematically describe our existing knowledge, referred to as the prior distribution, and update that knowledge with data. This results in a new distribution referred to as the posterior. In the case of referent authority, our referents are the sources of data we will use to update our knowledge.

The general form of a Bayesian method for updating information is provided in Equation 2.3 where $p(\theta)$ is a prior probability distribution modelling our existing understanding of some parameter or set of parameters, θ. In the equation, $L(\theta|\text{data})$ is a likelihood function that assesses the likely values of a parameter of interest according to the observed data. The result, $p(\theta|\text{data})$, is our updated understanding that takes into account both our prior knowledge and our data.

$$p(\theta|\text{data}) \propto L(\theta|\text{data}) * p(\theta) \tag{2.3}$$

To accommodate the fidelity calculation, we define θ as (μ, σ^2), the mean and variance, respectively, of the response of interest. The posterior distribution then provides inputs to $\bar{x}_r$ and s_r respectively for Equation 1.4.

By using NPP, the prior data is discounted depending on how commensurate the prior data is with the current data being considered. This method allows us to mathematically model our understanding of a system based on multiple referents which have differing levels of referent authority. In our approach, the referent with the highest referent authority according to Table 1 provides a baseline for how we expect the system to behave and acts as the “current” data. All other lower-level referents are weighted according to their agreement with the highest-level referent, as quantified by τ, ranging from 0 to 1. The more agreement that is seen between referents, the more weight or trust is given to them. If there are multiple referents that share the highest referent level, the user must determine which referent is more trusted, or pool the data to be considered as a single referent. If no determination is made, the method defaults to using the referent with the most data.

The NPP method, including how τ is determined, is further described in Ye et al. (2019).

Equation 2.4 gives the posterior distribution, now incorporating the general form of NPP (the product term) and depending on all available data: the most authoritative referent, y_c, and all n

lower-level referents, $y_{h1}, \ \dots, y_{hn}$, each with their own weight, τ_i, where $i = 1 \dots n$. The fidelity inputs $\bar{x}_r$ and s_r are the expected value $E(\mu|y_c, \dots, y_h)$ and the square root of the expected value $\sqrt{E(\sigma^2|y_c, \dots, y_h)}$ of the posterior distribution, $p(\mu, \ \sigma^2|y_c, y_{h1}, \ \dots, y_{hn})$, in Equation 2.4 (Ye, 2019).

$$p(\mu, \ \sigma^2|y_c, y_{h1}, \ \dots, y_{hn}) = L(\mu, \ \sigma^2|y_c) \prod_{i=1}^{n} \left[\frac{\left[L\left(\mu, \ \sigma^2\middle|y_{hi}\right)\right]^{\tau_i} p_0(\mu, \sigma^2) p_0(\tau_i)}{\int \int \left[L\left(\mu, \ \sigma^2\middle|y_h\right)\right]^{\tau_i} p_0(\mu, \sigma^2) d\mu d\sigma^2} \right] \tag{2.4}$$

Here, $L(\mu, \ \sigma^2|y_c)$ is a likelihood function of the parameters given the most trusted data set y_c, and $L(\mu, \ \sigma^2|y_{hi})$ is a likelihood function of the parameters given a lower-level referent y_{hi}. Initial priors, represented by $p_0(\mu, \ \sigma^2)$ and $p_0(\tau_i)$, are chosen to reflect all possible values of our parameters, and the possible values of w_i, which ranges from 0 to 1. These initial priors are uninformative priors whose distributions reflect no previous information and minimally impact the posterior distribution (Zellner, 1971).

This paper assumes that the random variation of referent data is normally distributed within a given set of factor inputs. While this is often not true, the method will perform well for most cases where the data is continuous and roughly symmetric. It may also perform well with non-symmetric distributions given sufficient data. Given the assumption of normal data, we use a normal inverse-gamma prior distribution. However, this structure is easily modified to account for other data types. For categorical-response data or number-of-success measures, we can assume a binomial distribution for data with a beta distribution as a conjugate prior to describe our parameters. For highly-skewed data such as time-to-failure data, we can assume a Weibull distribution for data with an inverse-gamma distribution as a conjugate prior. In these alternate cases, the output parameters are probability of success, p, in the binomial case or alpha and beta in the Weibull case. These values would then need to be converted into μ and σ^2 according to the distribution's properties for the sake of fidelity calculation.

Scope

Scope is the set of model inputs, outputs, assumptions, and limitations representing the mission-relevant system parameters, environmental conditions, constraints, and requirements, and their allowable values. For a model of a physical system, this would include operational parameters and states that the system is intended to operate under, such as, different speeds or with stealth capabilities active or inactive for a plane. Scope also reflects the intended use by including parameters that define the operational environment of the modeled system such as cloud cover, operational altitude, or the presence of jamming signals. Proper validation of a model requires that fidelity be assessed across the entire scope. There are infinitely many possible input combinations within the scope in the presence of continuous inputs, and as a result it is impossible to directly validate a model everywhere in the operational domain. Instead, we can assess how well covered the scope is by our referents and model.

Issues of scope in model validation can often be seen as one of four issues; a.) poorly defined scope, b.) differing fidelity across the scope, c.) referents not extending across the entire scope, and d.) scope regions supported by sparse information. Properly defining the scope, or mission space, of a model is critical to validation. No metric will state if the scope is defined correctly, but an improperly defined scope will often lead to poor validation metrics in the case where the scope covers too much space or irrelevant factors. In addition, poor scope definition can lead to a model validated only for a portion of its mission space in the case that the scope does not include all important factors. The issues of differing fidelity, unsupported scope regions, or sparsely supported scope regions can be seen in Figure 1.

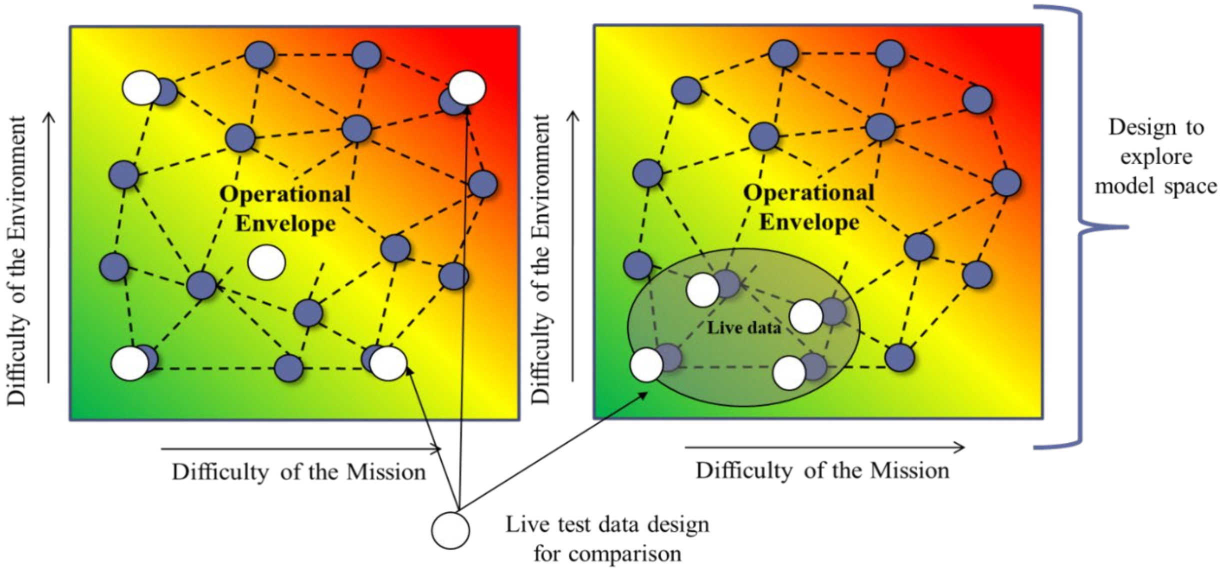
Difficulty of the Environment
Operational Envelope
Difficulty of the Mission
Difficulty of the Environment
Operational Envelope
Live data
Difficulty of the Mission
Design to explore model space
Live test data design for comparison

Figure 1

Notional Live Data Coverage of an Operational Space

Note. Figure from (Institute for Defense Analysis, 2019)

In Figure 1, the red section in the top right represents an operational region that is difficult to gain information about whether due to danger, cost, or other factors restricting test. The picture on the right represents likely situations where a large portion of the mission region is unaccounted for by our referents (the white dots). This may result in a model that seems to be high fidelity due to agreement with the available data, but may differ dramatically from reality in the unsupported sections. Likewise, in the left picture, we may have referents extending across the scope, but with large gaps between referent information. In other words, we may still have regions where the model doesn't capture reality, but fidelity cannot be properly assessed due to unavailable data. To account for these issues, we need to not only assess fidelity at various points across the scope, but we must also assess how well covered the scope is by our referents.

A metric to assess mission region coverage needs to assess both volume coverage and density coverage. Volume coverage describes the volume of mission space covered, while density coverage describes the density of points covering a defined scope region. The metric will be a multiplicative combination of separate volume and density sub-metrics. These metrics should be bounded between 0 and 1, representing no coverage and perfect coverage, respectively. Both high-volume coverage and high-density coverage are needed for a high-overall coverage score, yet a poor score for either volume coverage or density coverage is sufficient to result in a poor overall-coverage score. Finally, a coverage metric should normalize all mission factors to have the same weight. This normalization rescales any numerical factors to a scale from 0 to 1. This ensures that the impact of a factor isn't tied to the measurement scale used. Our overall metric then takes the form seen in Equation 3.1, where C_V is a volume metric and C_D is a density metric.

$$C = C_V C_D \tag{3.1}$$

The following sections will address the considerations for the volume component and the density component of the coverage metric in turn. Notably, these metrics are defined below for assessing continuous factors. To calculate these metrics with categorical factors, we must calculate the metrics individually within each factor level setting and then average the results.

Volume Metric

The simplest conceptual structure for a volume metric is to measure the ratio of the volume covered by data points to the volume of our mission space, where volume is understood not in three dimensions but as a d-dimensional volume where d is the number of factors that define our mission space. This metric is seen in Equation 3.2 (Hemez, 2010) where V_{data} is the volume of the convex hull around our data and V_{domain} is the entire volume of the domain, or mission space.

$$C_V = \frac{V_{\text{data}}}{V_{\text{domain}}} \tag{3.2}$$

The domain volume is computed with the understanding that each mission factor is rescaled and bound from 0 to 1, with 0 representing the minimum factor value of mission interest and 1 representing the maximum value of mission interest. The domain volume is then simply 1 unless some additional constraint is added where the system will not operate. The computation of data volume is more demanding. For this, we draw a convex hull around all points containing both referent and model information, which is the smallest shape that fully encapsulates our information while having no concave surface geometry. However, this construct presents computational problems and conceptual problems. Computationally, this calculation is straightforward and easily automatable, but as the number of dimensions increase, the time requirement for such a calculation increases exponentially. Conceptually, this metric is not consistent for interpretation between models with differing amounts of factors. As the number of dimensions increases, an increasing proportion of a geometric object's volume is found near the edges. With models, this means that a model with many factors would need data focused much closer to the edges of the mission space to cover the same volume proportion as a model with few factors. These issues drive the need for a computationally less demanding and more consistently interpretable volume metric.

Our proposed volume metric, seen in Equation 3.3, will still compare the d-dimensional volume of the convex hull around our information to the d-dimensional volume of our mission space only when our mission space is defined by 5 or less dimensions. In cases where our space is defined by more than 5 dimensions, we compensate for the growing issue of dimensionality by looking at 5 dimensional projections of our space. Specifically, we consider all $k = (d \text{ choose } 5)$ combinations of 5 factors from among the d factors that define our space and calculate the coverage volume with only those 5 factors in mind. Upon calculating volume for all k 5-factor spatial projections, we average the coverage volumes to calculate our volume metric. We define $V_{\text{data},i}$ and $V_{\text{domain},i}$ as the 5-dimensional volume of the data and domain, respectively, in the i^{th} unique 5-factor projection, where $i = 1 \ldots k$. We raise the calculation to a factor of 1/5 or 1/*d* for models with less than 5 dimensions, which has the benefit of providing a similar scale regardless of the total number of dimensions involved.

As a final consideration, while calculating the volume of each 5-factor dimensional projection, we calculate the hull around a random sample of our points containing both referent and model data rather than the full set of points. The random sample size is the total sample size, n, scaled down to be approximately proportional to the reduction in dimensions, or $n_{\text{sample}} = \text{ceiling}(n * 5/d)$. Keeping the data amount scaled with the number of dimensions prevents our revised volume metric from overcompensating and overstating the volume.

$$C_V = \begin{cases} \frac{1}{k}\sum_{i=1}^{k}\left(\frac{V_{\text{data},i}}{V_{\text{domain},i}}\right)^{1/5} & \text{for } d > 5 \\ \left(\frac{V_{\text{data}}}{V_{\text{domain}}}\right)^{1/d} & \text{for } d \leq 5 \end{cases} \tag{3.3}$$

The time complexity for computing a convex hull with the Qhull algorithm is $O(n\text{^floor}(d/2))$ (Barber, 1995), where n is the total sample size. The time complexity for the proposed volume metric has much more sustainable scaling, allowing computation of the volume coverage in high dimensional spaces. The comparison in time complexity is shown in Figure 2 for $n = 100$.

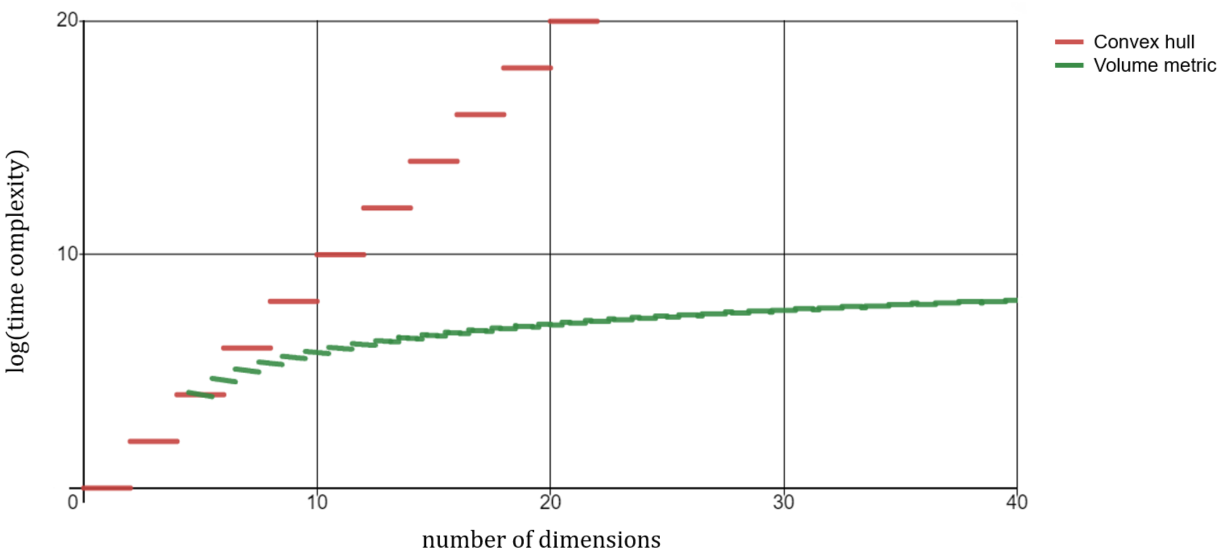

Figure 2
Time Complexity Comparison between Convex Hull and Volume Metric

Density Metric

With density, as with volume, we consider a straightforward case and then a more nuanced case to correct for shortcomings. In our straightforward density metric, we must keep in mind that we are assessing the information density evenly across the whole region of interest. This straightforward metric uses a similar approach to that of Atamturktur et al. (2009). As with our volume metric, we normalize our factors to a scale from 0 to 1 to ensure consistent presentation and interpretation of our metric. Then, to evenly evaluate across our space, we generate assessment points using a Latin Hypercube method which generates points distributed across the region of interest. All assessment points must be well covered for high coverage to be achieved. Coverage of each point is assessed in turn, which together provides a complete picture of density across the scope. To assess an individual point, we can use a nearest neighbor metric where we calculate the distance from the assessment point to the nearest referent point. Looking across the accumulation of assessment points, we will see that an informationally dense mission region will on average have data closer to each point while an informationally poor region will on average have data further away from each point. To compensate for differing scales as the number of dimensions increases, we will scale down the measured distance by the maximum distance attainable between two points in the mission space. Finally, as increasing distance from assessment points implies a reduction in coverage, we must present the average distance as a reduction from the maximal score of 1. The resulting metric is Equation 3.4 where r_i is the nearest neighbor distance from assessment point i to a referent point, q is the number of assessment points, and $\max(r_d)$ is the maximum distance possible in a mission space of d dimensions with normalized factor ranges, and is equal to $\sqrt{d}$ when no constraints on the mission space are present.

$$C_D = 1 - \frac{\sum r_i/q}{\max(r_d)} \tag{3.4}$$

While this metric presents a picture of density scaled from 0 to 1, it suffers from some undesirable qualities. Notably, it grants considerable coverage even when data is far from an assessment point. For instance, an assessment point in the middle of our space will have a scaled distance value no smaller than 0.5 if there is a data point anywhere in the scope region. This contributes to sparse referent coverage resulting in inflated density scores. Similarly, even data that is very close to an assessment point results in a penalty which causes very dense, well-filled spaces to receive understated density scores. For example, data spread by a Latin Hypercube method, which fills the scope by design, results in mediocre density metrics. The correction for these two problems is to establish maximum and minimum distances, M and L, where M is the distance beyond which no coverage is awarded and L is the distance within which perfect coverage is awarded. For maximum distance, the proposed distance is $M = \sqrt{d}/2$, which is the distance to ensure that the center of the scope inherits no coverage from corner points. For minimum distance, the proposed distance is $L = \sqrt{d}/6$, which was chosen by simulation as a favorable balance where Latin Hypercube designs receive favorable density metrics, and full-factorial designs, where data is located only at the corners, receive unfavorable density metrics. Between these minimum and maximum points, distance is awarded a coverage score according to a linear scale. Then, the coverage scores are averaged among q assessment points for an overall coverage metric, as seen in Equation 3.5.

$$C_{\mathrm{d}} = \frac{1}{q}\sum_{i=1}^{q} c_i \quad \text{where} \quad c_i = \begin{cases} 1 & \text{for} \quad r_i \le L \\ \frac{r_i - M}{L - M} & \text{for} \quad L < r_i \le M \\ 0 & \text{for} \quad r_i > M \end{cases} \tag{3.5}$$

Considerations for a Combined Metric

The separate pillars of fidelity, referent authority, and scope are inherently related, but their mathematical considerations are disjoint. Fidelity is measured at single points in space and separate measures of fidelity must be made and accumulated in a discussion of scope. Referent authority is assessed independent of spatial considerations but must be examined against scope as some referents are only available within a limited mission space. Fidelity and referent authority have interplay in their calculations as we accumulate multiple referents to generate a single-fidelity calculation at a given space, but only part of the referent authority considerations are incorporated in such calculations. Care must be taken to address all considerations for fidelity, referent authority, and scope in a single MVL measure.

A successfully combined MVL must have several properties. It must:

- assess fidelity of a model using the full body of knowledge or the whole set of relevant referents available across the mission space
- weigh the fidelity assessment according to the level of trust or authority that can be placed on the referents that validation will consider

- penalize a model if the entire mission space is not supported by referent knowledge or if there is an insufficient amount of support across the mission space
- be presentable in a concise, interpretable, and actionable manner

The proposed method of implementing the metrics and mathematical constructs in this paper in a single MVL assessment metric will be discussed in a future Best Practice. Along with this best practice, the STAT COE will provide a coded tool to automate the MVL process as well as a guide to using the tool and a Case Study of MVL implementation in a specific program.

Conclusion

As the DOD engineers defense systems of increasing complexity and comes to rely on M&S to understand and develop those systems, it is imperative that the models developed are well understood and trustworthy to minimize any risk introduced by the use of models in place of physical articles. Validation is the process which establishes the level of trust that can be placed in a model to represent the associated physical system. However, in practice, validation is often a subjective process resulting in a binary indicator of validity which grants the model validity for its entire lifetime without reassessment. MVLs address these problems and provide a rigorous validation framework that can be quickly, repeatedly applied to the wide variety of M&S in the DOD. MVLs will enable DOD program managers and M&S developers to effectively employ M&S in the engineering of complex defense systems with full comprehension of model capabilities and risk.

References

Ahner, D. K., Jones, N., Key, M., Adams, W., Burke, S., & Weeks, C. (2021). A Conceptual Framework for the Establishment of Model Readiness Levels. White Paper, Scientific Test & Analysis Techniques Center of Excellence (STAT COE).

Atamturktur S, Hemez F, Unal C, William B (2009) Predictive maturity of computer models using functional and multivariate output. In: Proceedings of the 27th SEM international modal analysis conference, Orlando

Barber, C. B. (1995, September 25). *Qhull Manual*. Qhull. Retrieved August 9, 2022, from http://www.qhull.org/html/index.htm#description

Deputy Assistant Secretary of Defense for Systems Engineering (DASD(SE)). (2018). DoD Digital Engineering Strategy. United States of America, Department of Defense, *Office of the Secretary of Defense*.

Hemez F, Atamturktur S, Unal C (2010) Defining predictive maturity for validated numerical simulations. Comput Struct J 88:497–505

Institute for Defense Analysis (2019). Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation, *Office of the Director, Operational Test and Evaluation*

Weeks, C., Jones, N., Key, M. (2022). Constructing a Metric for Fidelity in Model Validation. White Paper, Scientific Test & Analysis Techniques Center of Excellence (STAT COE).

Ye, K., Han, Z., Duan, Y., Bai, T. (2019). Normalized power prior Bayesian analysis. Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX, USA

Zellner, Arnold (1971). "Prior Distributions to Represent 'Knowing Little'". An Introduction to Bayesian Inference in Econometrics. New York: John Wiley & Sons. pp. 41–53. ISBN 0-471-98165-6.

Appendix

Key Definitions

To ensure a common understanding of the subject, the following definitions will be used throughout this paper:

accuracy: the degree to which a parameter or variable, or a set of parameters or variables, within a model or simulation conforms exactly to reality or to some chosen standard or referent (Modeling and Simulation Enterprise, 2021).

aleatory uncertainty: uncertainty arising from an inherent randomness in the properties or behavior of the system under study (Helton, 2011).

convex hull: the smallest possible convex space that contains a set of data points

epistemic uncertainty: uncertainty derived from a lack of knowledge about the appropriate value to use for a quantity that is assumed to have a fixed value in the context of a particular analysis (Helton, 2011).

fidelity: the level of consistency between a model and a referent, defined in the three dimensions of accuracy, repeatability, and resolution.

model: a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process (DoDI 5000.61).

modeling and simulation (M&S): the use of models, including emulators, prototypes, simulators, and stimulators, either statically or over time, to develop data as a basis for making managerial or technical decisions (Modeling and Simulation Enterprise, 2021).

referent: a trusted representation of reality.

referent authority: the strength of credibility of a referent’s claim to be a high-fidelity representation of reality.

repeatability: the similarity of the results obtained from the same model (or referent) over multiple observations under the same input conditions.

resolution: the degree of granularity with which a parameter or variable can be determined (Pace, 2015).

scope: the set of model inputs, outputs, assumptions, and limitations representing the mission-relevant system parameters, environmental conditions, constraints, and requirements, and their allowable values.

simulation: a method for implementing a model over time (DoDI 5000.61).

specific intended use: the set of dimensions, ranges, and assumptions of the model inputs and outputs needed to represent a system’s relevant mission parameters, environmental conditions, constraints, and requirements, combined with the additional constraints imposed by the target modeling environment and the required level of fidelity for the specific stage of program development.

validation: the process that determines whether a model has sufficient fidelity relative to an appropriate referent(s) for a specific intended use.

validity: the fidelity of a model over a pre-specified scope relative to an appropriate referent(s).