

Implications of Multi-Resolution Modeling (MRM) and Exploratory Analysis for Validation

James H. Bigelow and Paul K. Davis
RAND, 1700 Main Street, Santa Monica, CA 90401

Keywords: Multiresolution, variable resolution, family of models, exploratory analysis, model validation, metamodel

INTRODUCTION

Objective

Over the last decade, we and colleagues have done considerable theoretical and applied work involving multiresolution modeling and exploratory analysis,¹ in part to connect the worlds of strategic planning (e.g., development of the defense guidance and defense programs)² with the world of more detailed analysis (e.g., development of military forces, weapon systems, and doctrine).³ We have also found multiresolution work essential in attempting to model adversaries in studies of deterrence and compellence.⁴ Our work has not been about model validation per se, but many implications for model validation have emerged as byproducts. Our objective in this paper is to discuss some of those implications. In doing so, we draw heavily on examples from our past work.

Some Prefacing Philosophy

It is often believed that if one has a respectable high-resolution model and a respectable low-resolution model, then the low-resolution model can and should be calibrated to be consistent with the high-resolution model, and that if the high-resolution model is considered authoritative (at least organizationally accepted), then doing so constitutes validation of the low-resolution model. By and large, we reject this notion. Our experience has been that in such a case it is quite

¹ Davis and Bigelow (1998), Davis, Bigelow, and McEver (2001), and Davis and Bigelow (forthcoming). Our work has been sponsored by the Air Force Research Laboratory, Office of the Secretary of Defense, and the Defense Advanced Research Projects Agency.

² Davis (1994) and Davis (2002).

³ Defense Science Board (1998) and Gritton, Davis, Steeb, and Matsumura (2000).

⁴ See "Improving Deterrence in the Post-Cold War Era," in Davis (1994).

likely that *both* models will have strengths and shortcomings (substantively, not just in terms of ease of use) and that the correct procedure is to draw on all information available, at all resolutions, including results from running the models, so as to end up with a pair of *mutually calibrated* models. Ideally, the models' structures will be clearly related and, to greater or lesser extent, the two models will represent an integrated family. Even if the structural relationships are muddy, however, mutual calibration may be possible.

We also observe that, independent of model "validity," it is crucial that model-based analysis be *comprehensible*, but the explanations needed vary greatly with context. Sometimes, in analyst-to-analyst discussion, or in responding to a probing spot-check question by a client, the explanation needs to be detailed (e.g., "Well, the limiting factor in such cases is the number of weapons that can be carried on the <platform>, which is actually much less than you might think because of a deeply embedded software limitation in the <fire-control system>, which can't readily be fixed"). Other times, and indeed in most discussions with higher-level clients, explanations need to be more abstract and lower in resolution, as in, e.g., "It really comes down to a tradeoff between the speed of deployment versus the nature of what can be deployed. For the same dollars, we can either have terrific fast-response capability against small-to-moderate threats or slow-response capability against a larger threat." Further, "accreditation" (essentially official acceptance for a very specific purpose, such as drawing particular conclusions in a particular study) may depend strongly on the quality and persuasiveness of explanation to different audiences. It follows, then, that we need combinations of low-and-high-resolution models for explanation purposes also.

Figure 1 tries to capture this idea by using double-headed arrows in a multiresolution family.⁵ The arrows may be regarded as information flow in the sense of mutual calibration, or in the sense of explanations. This paper is about how we have turned these ideas into reality in a number of studies. We believe strongly in using a range of models, from simple, low-resolution models reducible to a formula, to entity-level agent-based simulations with detailed physics. For us, the family-of-models approach is not just a "nice idea," but something that can and should be more routinely adopted.

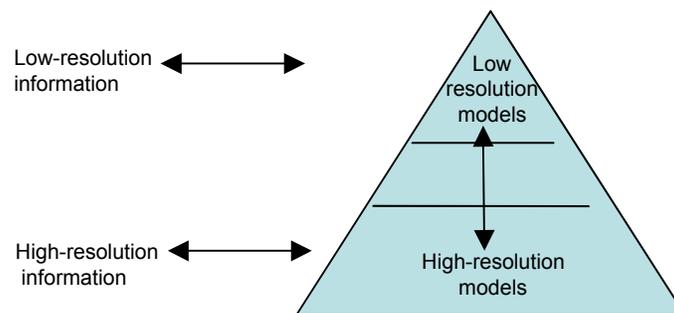


Figure 1: Mutual Calibration of Multiresolution Models Using All Information Available

⁵ Adapted from National Research Council (1997).

Some Definitions and Distinctions

For the sake of clarity in our paper, we simplify discussion by imagining that we deal with models at only two levels of resolution, low and high. However, we distinguish among a number of low-resolution models. Empirical models are models developed strictly from data (e.g., curve fitting). Metamodels, or models of models, are developed by applying statistical methods to data generated from high-resolution models (referred to as the object models). Theory-based low-resolution models are models built from the viewpoint of phenomenology, but with low-resolution concepts (e.g., the volume of a pond might be estimated roughly as the product of its average breadth, width, and depth). We distinguish further between purely statistical metamodels (often referred to as response surfaces) and theory-motivated metamodels. The former are obtained by applying statistical methods to the output of “experiments” conducted with the high resolution model. No physical insight is necessary and may not even be desired by the analyst: he is just doing “data analysis.” Theory-motivated metamodels, to which we shall subsequently refer as just “motivated metamodels,” are obtained by using physical and behavioral reasoning to suggest the *structure* of the model, after which coefficients, exponents, and correction factors are determined using statistical analysis of high-resolution model runs.

Another term appearing in the title of our article is “exploratory analysis.” Exploratory analysis is an analysis strategy that focuses on breadth and a synoptic view, rather than depth. It is best done with low-resolution models. Exploratory analysis is used when dealing with massive uncertainty. It is used, e.g., to develop *robust* strategies.⁶

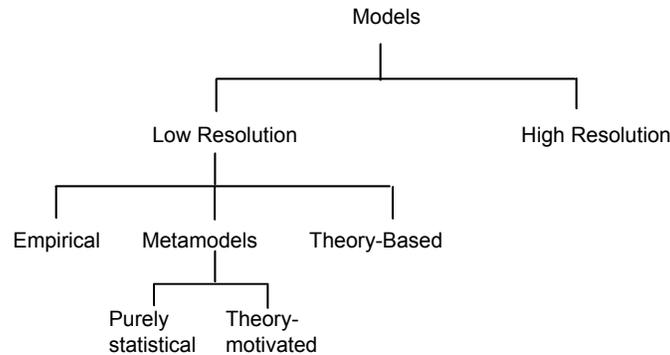


Figure 2: Taxonomy Used In Paper

Finally, we need a definition for validation itself. According to the glossary on the website of the Defense Modeling and Simulation Office (DMSO), validation is defined as follows:

Validation: (1) The process of determining the degree to which a model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model. (2) The process of determining the fitness of a model and its associated data for a specific purpose.

⁶ Exploratory analysis has developed from a desire and concept (Davis and Winnefeld, 1983) to a reality, due to a combination of technological developments and theoretical work. See citations in Davis, Bigelow and McEver (2001) and Davis and Bigelow (forthcoming).

We shall address both aspects of this definition in what follows. As we shall demonstrate, the definition's subtleties matter.

Structure of Paper

With this background, our paper proceeds as follows. The first section provides validation-related reasons for working at multiple levels of resolution. The second section discusses what it means to maintain *consistency* across levels of resolution (consistency between models is related to model validation). The third section argues for the importance of having a low-resolution model (perhaps a metamodel) that provides a credible and insightful story or theory about causal relationships. It also suggests that the form of metamodels should be "motivated," by drawing upon knowledge of the subject area to assure such a story is built in, even if only crudely.

Appendices provide extensive examples to illustrate points made in the main text.

VALIDATION-RELATED REASONS TO WORK AT MULTIPLE LEVELS OF RESOLUTION

Uses of MRM When Comparing Weapon Systems or Forces

We have discussed broad motivations for multi-resolution, multiperspective modeling (MRM and MRMPM) elsewhere. Here, let us focus on its value for validation. A starting point is to realize that analysts *often* work at multiple levels of resolution, but may not realize it. Consider a client who faces the question: "Given a specified amount of funding, should the military buy weapon system A or B or C?" We perform the study, which involves the heavy use of models, and find that the answer is: "Buy system B." The client, of course, will not be content with this unsupported advice, so we need to construct a persuasive argument that supports this conclusion. The legs of the argument may be:

Leg One: System B is more combat effective than A or C in scenarios 1, 2, 3, and 4.

Leg Two: Scenarios 1, 2, 3, and 4 together constitute an adequate test of the combat effectiveness of these systems. That is, the conclusions drawn from these can be generalized to all scenarios for which the new capability is needed as can be seen from an abstracted (low-resolution) depiction of results in which the scenarios are merely representative cases (e.g., Figure 3).

Leg Three: System B has adequate non-combat performance (e.g., logistics requirements, mobility, cost, production schedule).

We shall refer extensively to these legs in what follows.

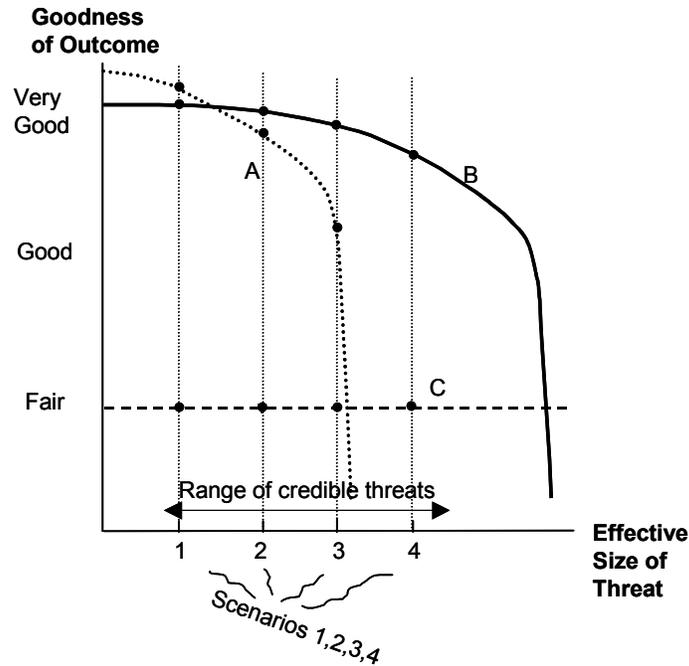


Figure 3: A High Level Capabilities Comparison for which Detailed Analysis was Done for Four Scenarios

Consider Leg One above. Analysts typically assess the combat effectiveness of each weapon system in each of a number of scenarios, using a high-resolution simulation model. Then, however, they construct credible but simplified stories *explaining* why the high-resolution model behaves as it does. This helps to establish the face validity of the high-resolution model and analysis. It is often a relatively small step to turn such a story into a low-resolution model, but even if this step is not taken, the use of simple stories in these ways is an example of working at multiple levels of resolution. Appendix A provides an example. In this case, the low-resolution work not only explained puzzling results from high-resolution work, it also pointed out potential flaws and provided an easy mechanism for extrapolating results to other circumstances. However, the low-resolution work was motivated by discoveries of the high-resolution work. Thus, in the terms of Figure 1, there was a good deal of back-and-forth between levels.

Next, consider Leg Two. At the end of the study the analyst must explain his results to the client. Even a combat simulation of modest size will have tens of megabytes of inputs and outputs. No analyst will show the client the raw inputs and outputs – at least, not more than once. Instead, he will select and summarize. He will present only the amount of information that can comfortably fit on two or three briefing charts, with the claim that this small amount of information captures the essence. This requires him to extrapolate and generalize beyond the relatively few high-resolution cases that were run (12 in the example of Figure 3). Even if the high-resolution model is considered valid, it is necessary to establish that the generalizations are correct. The analyst usually appeals to the simple stories he constructed to develop confidence in the high-resolution model and its data sets. A better way, however, is for the analyst to be able, simultaneously, to discuss results *broadly* (in an abstracted, lower resolution way) and to go into details selectively –

for purposes of both credibility and explanation. Figure 3 shows both depictions. Each point represents a case worked through in detail with the high-resolution model, but the *curves* for A,B, and C were generated by a lower-resolution model mutually calibrated with the first. One sees the “big picture,” along with results of detailed analysis. The multiresolution analysis described in Appendix A can be used to generate such capability curves.

Look finally at Leg Three, where analysts frequently supplement detailed work with very low-resolution, back-of-the-envelope calculations in the service of validation. For example, suppose we perform a detailed computation to estimate the unit cost of a proposed aircraft. We include costs for the engines, fuel system, airframe, landing gear, avionics suite, system integration, etc. The result is detailed, precise, and impressive. However, we (or competitors down the hall) may also calculate the aircraft’s cost per pound, and compare it to historical experience for comparable technology (the back-of-the-envelope, low-resolution approach). If the estimates more-or-less agree, we have reasonable confidence in the detailed calculation. However, if they don’t agree, we are likely to look for errors in the *detailed* computation.

Appendix B presents an extended example of using a low-resolution computation (in this case, an empirical model) to check the results of a high-resolution calculation. Interestingly, the high-resolution results is shown to be wrong, even though it was made with a validated model! The error was that the model had been validated for a very different purpose.⁷ This example again illustrates one of the points referred to in Figure 1, that low resolution information and models can greatly inform judgments about the validity of a high-resolution model for a particular analysis. On the other hand, the more detailed analysis makes it possible-with caution-to consider effects not present in historical cases.

Top-Down, Multiresolution Thinking in Studies and Experimentation

Let us next discuss the role of multiresolution work in conceiving and executing a *program* of study and experimentation, perhaps in the context of force transformation. Here one should worry about validating the study or experimentation campaign: will the results, when they are in, be sufficient to justify decisions about acquiring a new system, adopting a new operational concept, and so on? The answer depends on the validity of the war games, models, simulations, and field experiments used, and, ultimately, the “campaign plan” for pulling things together.

One aspect of such a campaign must be to identify the broad “scenario space” within which the prospective capabilities are to be assessed. A second is to identify specific cases for which the assessment would be worked through in detail, with high-resolution simulation, or large-scale

⁷ As a second example, suppose that we are building a detailed mobility model simulating strategic lift by a combination of various types of aircraft from different bases to different destinations. We may build a simple spreadsheet model that estimates tons moved as a function of the number of “equivalent C-10s,” the average utilization rate, the average load, etc. The inputs are all familiar from past work. We then compare results between the detailed and simple models. If there is a large error, it is likely that the error is in the detailed model—whether because of a verification-related problem (e.g., typos) or something more serious, such as omitting real-world factors, such as maintenance times and crew rotation factors, that reduce the fraction of the time that aircraft can be used.

field experiments, or perhaps a combination. In developing such a plan, one should be concerned about which instruments (e.g., a simulation model or a field experiment) is appropriate for which purposes. Although military experimentation is often driven by the demands of an anticipated big event (a particular large-scale field experiment), that is a poor way to develop a campaign plan. As discussed in Appendix B, we recommend a top-down approach.

In the context of force transformation, for example, we have argued (Davis, Gompert, et al., 1998; Davis, 2002) that the driving impulse for initiatives should be the *important and stressful* operational challenges that might be faced in the future. Five examples for projection forces are as listed below:

- Early halt of a classic armored invasion given depth (e.g., in Kuwait or Northern Saudi Arabia);
- Early shallow halt of a fast invasion on multiple axes, without depth (e.g., Korea);
- Early offensive action without first building up a massive force (e.g., Kosovo or Afghanistan);
- Effective low-risk interventions (e.g., Bosnia);
- Effective peacemaking in urban environments (e.g., Kabul).

In a sense, specifying such challenges is akin to specifying the *broad* dimensions of a name-level scenario space (“name-level” suggests that we’re talking about classes of scenario, rather than all of the many details that would apply in a specific war at a specific time). U.S. forces must be prepared for many other challenges, but the requisite capabilities may come along naturally (e.g., who doubts that the Air Force will be able to achieve air-to-air superiority or that the Navy will be able to achieve control of the seas, even if the Secretary of Defense provides no specific guidance about related scenarios?).

Given a set of broad challenges to work with, we recommend a hierarchical approach to fleshing matters out. Briefly (see also Appendix C), we recursively break down each challenge into the need for various building block capabilities. For example, to accomplish an early halt given depth, we must quickly (a) establish an effective C4ISR capability in the theater *and* (b) secure bases or operating locations *and* (c) rapidly deploy forces, *and* (d) rapidly employ those forces effectively. We break down each of these into smaller building block capabilities, and those into still-smaller blocks (hence the term “recursive”). We arrive at a list of high-value building block capabilities, which may suggest the sorts of forces and doctrinal innovations we ought to consider developing. The process of generating the building block needs highlights the circumstances in which they have value, thus suggesting the kinds of scenarios in which they should be tested and metrics that should be used. These tests can be carried out by some combination of model-based analysis and experiments in the field.

To elaborate, compare (a) and (c) in the above list of necessary building-block capabilities. There may be no reason to conduct expensive large-scale exercises to investigate deployment for diverse assumptions: models and simulations can do well and are driven largely by the laws of

physics and factors such as decision delays. In contrast, the ability quickly to establish an *effective* C4ISR capability is a wholly different matter. Even if current models and simulations did a good job of representing C4ISR systems in substantial detail, they would be utterly unreliable about major factors such as “How quickly can a newly assembled group of officers, brought in from diverse assignments worldwide, understand and master their new context, learn to work together well, and develop and direct appropriate tactics, techniques, and procedures?” As of 1998 (Defense Science Board, 1998; Davis, Bigelow and McEver, 1999) the answer was that it might take weeks, but no one really knew. Since that time, OSD and US JFCOM have put great emphasis on developing standing command and control capabilities that would greatly reduce the “spinup time” in conflict, but even today much of what is being learned is of the “soft” variety in which human experiments – and even some large-scale experiments – are necessary.

What the low-resolution top-down analysis demonstrated for the context of “the halt problem” (the first challenge in the above list) was that the spinup time (which, of course, would depend on a number of identifiable factors) was of first-order importance to the overall capability, and yet was not amenable to pure modeling and simulation: laboratory and field experiments were crucial. This said, *most* of the halt problem – and the assessment of alternative programs nominally addressing that challenge – could not only be done with modeling and simulation, but could be done with *low-resolution* modeling and simulation. In other instances, the analysis identified topics that cried out for detailed simulation and small-scale human experiments.

This top-down hierarchical approach is an example of multi-resolution thinking. Although the term “top-down” is used frequently, the approach is used too sparingly. Studies typically devote space to describing the scenarios used, but few studies make more than a feeble, ad hoc attempt to show why those scenarios reflect the range of circumstances for which we want assured capabilities.

This process can perhaps best be seen as a matter of *design*. An architect or systems engineer must understand and define his “design space” in relatively low-resolution terms so that he can see issues and address tradeoffs. At some point, he or others will need to use much more detailed descriptions to assess issues in depth, but both the broad view and the detailed view are necessary.⁸

Broad, Exploratory Analysis and Choosing Cases for Detailed Study

We have talked a good deal about broad analysis. Let us now elaborate. The term we prefer here is *exploratory analysis* (Davis, et al., 2002, 2001 and citations therein). Exploratory analysis is very different from the narrow and more traditional *predictive analysis*. In the predictive analysis strategy, the analyst identifies a base case and a few excursion cases. He runs his model for each one, and takes the results to be predictions of what would happen if the modeled circumstances were replicated in the real world. In a concession to uncertainty, he runs a handful of sensitivity

⁸ Such matters are discussed extensively in Davis (2002), which was written to put more meat on the bones of “capabilities-based planning.”

cases as well. In the exploratory analysis strategy, the analyst recognizes that he has no prior reason to single out a base case. Because of massive uncertainty or ignorance, he must explicitly examine hundreds or thousands of cases. But exploratory analysis can only be done rigorously with models that have but a handful of uncertain parameters to be varied (perhaps up to a dozen). The number of cases one must consider explodes exponentially as the number of parameters increases and the ability to comprehend and explain both inputs and the results diminishes accordingly. Faster computers will not solve these problems.

Some will argue at this point, claiming that detailed models can be used for exploratory analysis, which is quite true. We, and RAND colleagues Carl Jones and Dan Fox have done a good deal of this with a theater-level model called JICM, and other colleagues have used experimental-design methods to explore with complex models. However, the dimensionality of input datasets will depend on the resolution of the model. Suppose, for example, that the analyst is using high-resolution simulation models for entity-level analysis of force-on-force encounters. For each case the analyst must specify the *detailed* scenarios, specifying, e.g., terrain, build-up schedules of Blue and Red forces, Blue and Red tactics, and so forth. He must also estimate a host of parameters describing its capabilities relative to capabilities of weapons that may be arrayed against it. There will be dozens, if not hundreds or thousands of parameters. Varying them all, and in combinations, can rapidly generate astronomical numbers of cases. Yes, clever experimental design may reduce the number of cases needed, but not to single digits. The alternative, of varying only a small number of variables while leaving others constant is difficult to defend unless the model is very well understood (as, for example, in RAND's JICM work).

The way to avoid this "curse of dimensionality" is to work at multiple levels of resolution. Assume that we can build a low-resolution that is reasonably consistent with behaviors of the analyst's high-resolution model (see Appendix C), and needs only a handful of parameters to do so. The space of input datasets will have relatively few dimensions (5-12 in our studies). We can generate and analyze (with the metamodel) a big enough sample of input datasets to be reasonably sure that we have identified all the tough tests. We can then select one of the low resolution model's input datasets, and build a number of corresponding input datasets for the high-resolution model. This is an extension of the top-down process outlined earlier. Start with a broad, low-resolution view, identify areas where adding detail will add value and not just increase the workload, and drill on down to paydirt.

Validation of a Model for Exploratory Analysis

To our knowledge, there has not previously been discussion of what constitutes "validity" for a model intended for exploratory analysis, rather than prediction. The DMSO definition of validity refers to the combination of model and its data base, but in exploratory analysis, some or all of the input data is regarded as highly uncertain, not something to be proven correct.

We suggest that validation be understood in this context as requiring that the model be *structurally* valid and that the domain for exploratory analysis includes all cases of interest. Exploratory analysis, then, greatly reduces the burden of model validation – although assessing

structural validity can be very difficult and assessing the validity of conclusions from exploratory analysis depends, as always, on the quality of the analysis itself.

CONSISTENCY AND VALIDATION

In the previous discussion, we have relied on an intuitive notion of consistency between models with different levels of resolution. Intuitively, a multi-resolution family displays consistency if employing its family members at different levels of resolution generates no contradictions. In this section we develop a more precise notion of consistency.

We do this for two reasons. First, there is a close relationship between validation and consistency. To be valid, according to the DMSO definition, a model must be an accurate representation of the real world from the perspective of its intended uses. We don't expect it to provide a complete, fully detailed picture of the real world. So we interpret its "accurate representation" as one that is "consistent with" – i.e., does not contradict – the real world. More mundanely, suppose we have a high-resolution model that we consider to be valid. If we build a metamodel of it, then demonstrating that the two are consistent will presumably validate the metamodel. That is, a transitivity principle applies.

Second, low-resolution models are often deterministic, while high-resolution models frequently have stochastic components. But as an examination of the notion of consistency will show, one ought to expect low-resolution models to be more stochastic than high-resolution models. Each model has a representation of the thing modeled that leaves out details. Thus, an ensemble of many states of the thing modeled map into each state of the model. The stochastic elements in the model must represent not only the stochastic elements in the thing modeled, but the variation of outcomes over the ensemble as well. We believe the failure to fully represent this variability is a major reason that low-resolution models have gotten such a bad name in some circles. This bad name is unfair, because low-resolution models are often used in connection with sensitivity analysis or even exploratory analysis, in which case having a stochastic model may be unimportant, but in any case many analysts distrust and dislike low-resolution models.

Resolution

To discuss consistency, we first need to elaborate on "resolution." Resolution is the ability to distinguish one thing from another. We measure the resolution of an optical system by the smallest (angular) separation two points must have before their images are distinct. By analogy, every model has a representation of the system it models. We measure the model's resolution by the separation two states of the modeled system must have before they are represented differently in the model.

Of course, resolution is a much more complex notion in models than in optics. An optical image is a straightforward transformation of the scene viewed. By contrast, a model can use representations that are complex, obscure, and abstract. Let us look at some examples.

Terrain is often represented digitally in models, with grid cells ten kilometers (low resolution), or one kilometer, or a hundred meters (relatively high resolution) on a side. Similarly, time can be measured in days (low resolution), hours, or minutes (relatively high resolution). The lowest resolution might be adequate for representing the placement and movement of divisions in the theater. The highest might be needed for the simulation of target-seeking by a precision weapon.

Entities in the simulation can be sorted into many categories or lumped into only a few categories. Each kind of aircraft and missile could be represented as a different class of object, or they could all be lumped together as “long range shooters.”

Processes can be represented in detail as consisting of many different steps or activities. Or they can be represented by one or two factors. For example, resupply of ammunition could be modeled as:

- Communicating the need for ammunition to a storage site;
- Loading bulk ammunition on large trucks;
- Driving it to a transshipment point;
- Breaking the bulk truckloads into “combat configured loads” (i.e., consisting of the precise mix of ammunition types needed by the recipient); and
- Delivering the ammunition to the consumer.

More simply, one could represent ammunition resupply by a delay time. (Even more simply than that, one might choose not to represent ammunition resupply at all. Most combat models place no limits on ammunition consumption.)

Representations can be statistical. Air quality is measured by the concentrations of pollutants at measuring stations in an air basin. The raw data will show the concentration as a function of time. But an air quality standard is expressed in terms of an averaging time for a pollutant, a threshold concentration, and a frequency. For example, in 1970 the air quality standard for oxidants would be exceeded if the oxidant level, averaged over one hour, exceeded 0.08 parts per million during more than one hour per year. So it is convenient to calculate hourly averages from the raw data, and then to describe the set of those averages as a frequency distribution, without regard for the times at which a particular concentration was observed. Further, Larsen & Zimmer (1965) observed that this frequency distribution could be fit pretty well by a Log-Normal distribution, and hence described by two parameters. The statistical distribution, of course, is a lower resolution representation of air quality than the original time series of measurements, because there are many time series that have the same statistical representation.

Finally, abstraction plays a role in resolution. In Davis, Bigelow and McEver (2000a), (see also Appendix A) our low-resolution model used the concept of a “clearing.” For this model, a clearing was an open stretch of roadway where munitions could acquire targets unobstructed by foliage. The only attribute a clearing had was its width. In the high-resolution model, the terrain was modeled as an array of cells 100 meters on a side. Each cell either had trees or did not have

trees in it. To identify clearings in this terrain, we had to decide about ambiguous cases. Was a stretch of road a clearing if trees lined the road closely but the road itself was clear? Was an open stretch of road one clearing or two if it was interrupted only by a very short stand of trees? Must the treeless area be large enough that a distant reconnaissance platform could identify it, or only large enough so trees didn't interfere with the terminal search algorithm of a munition?

Aggregation and Disaggregation

"Aggregation" and "Disaggregation" are often used in discussing changes of resolution. With *aggregation* we move from high to low on some dimension of resolution. By *disaggregation* we do the reverse, moving from low resolution to high. When we aggregate we lose information, and we don't retrieve it when we disaggregate. This is inherent in the definition of resolution given earlier, namely that it involves the ability to distinguish among states of the modeled system. Many high-resolution states correspond to each low-resolution state. We can move from a high-resolution state to a unique corresponding low-resolution state, but when we try to move back – when we disaggregate – we must be content to deal with an ensemble of high-resolution states, all of which will aggregate back into the same low-resolution state. Here are some examples.

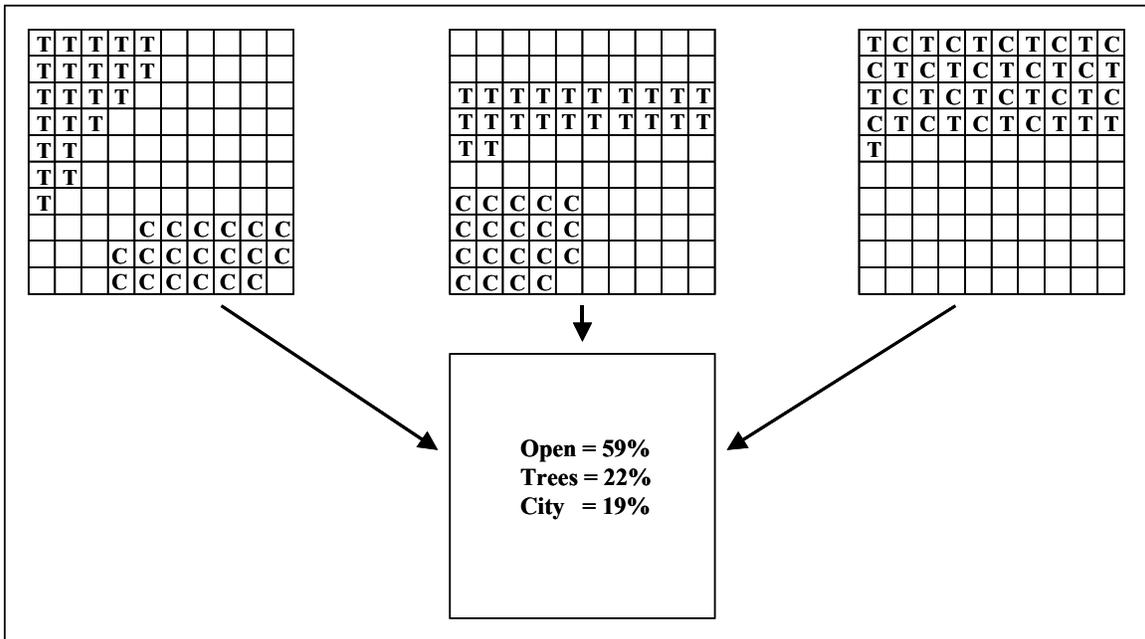


Figure 4: Aggregation Causes Loss of Information

Consider terrain represented as a rectangular array of cells 100 meters on a side (Figure 4). Each cell, let us say, has an attribute with one of three values. The cell can be (1) open, (2) covered with trees, or (3) covered by buildings. Now we aggregate this description to an array of cells 1 kilometer on a side. Each of the new cells consists of 100 old cells. We can characterize each new cell with three numbers: (1) the percentage of the old cells that are open; (2) the percentage covered with trees; and (3) the percentage covered with buildings. If we now try to disaggregate,

we cannot recover the original, high-resolution description. As shown in Figure 4, there are many 10×10 arrays of small cells that would aggregate to the same large cell.

We conceive of the aggregation of a process as truncating a network of variables such as depicted in Figure 5. Describing the process in detail requires many variables, but an aggregated description requires only a few. In the figure, “Y” represents an outcome of the process, the variables X_1, X_2, X_3 comprise the low-resolution description, and the variables X_4, \dots, X_{10} comprise the high-resolution description.

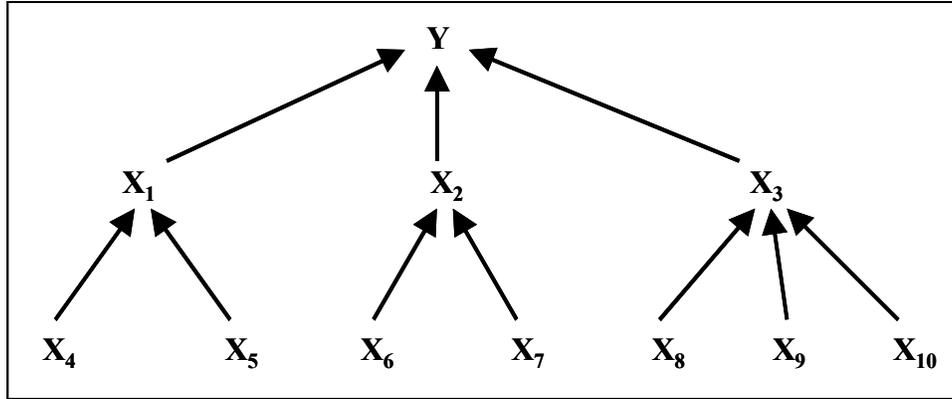


Figure 5: Notional Network of Variables

Algebraically, we write the low-resolution description as:

$$Y = F(X_1, X_2, X_3) \quad (1)$$

We write the high-resolution description as:

$$Y = F(G_1(X_4, X_5), G_2(X_6, X_7), G_3(X_8, X_9, X_{10})) \quad (2)$$

The links between them are the definitions of the low-resolution variables in terms of the high-resolution variables:

$$\begin{aligned} X_1 &= G_1(X_4, X_5) \\ X_2 &= G_2(X_6, X_7) \\ X_3 &= G_3(X_8, X_9, X_{10}) \end{aligned} \quad (3)$$

That is, we can substitute the expressions (3) into the low-resolution description (1) to obtain the high-resolution description (2).

Clearly, the high-resolution description contains more information than the low-resolution description. If we specify values for all the high-resolution variables, we can calculate unique values for the low-resolution variables using Equations (3). But if we are only given values for the low-resolution variables, we can find many high-resolution values that satisfy Equations (3).

Consistency

Figure 6 depicts the usual definition of consistency (e.g., see Davis & Bigelow, 1998). Starting in the upper left corner with high-resolution inputs, one can follow either of two paths. Going first to the right and then down, one uses the high-resolution model to produce high-resolution outputs, and then aggregates them to the lower level of resolution. Going first down and then to the right, one aggregates the high-resolution inputs to the lower level of resolution, and then uses the low-resolution model to produce low-resolution outputs. The two models are consistent if the results are the same (or nearly so) regardless of which path you follow.

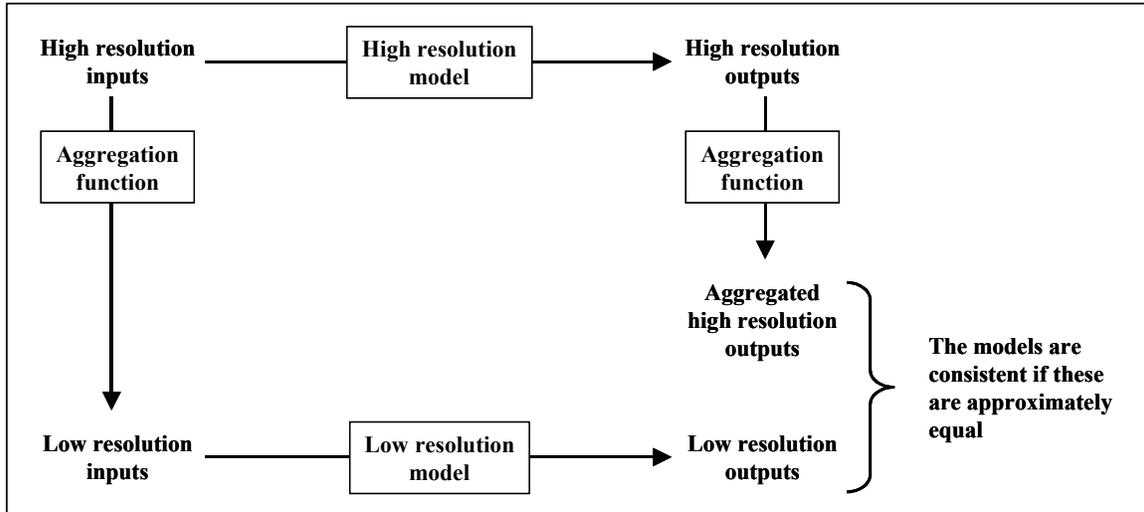


Figure 6: The Usual Definition of Consistency Between High- and Low-Resolution Models

However, as we stated earlier, we lose information when we aggregate. Aggregation moves us from a high-resolution state to a unique corresponding low-resolution state, but when we try to move back—when we disaggregate—we must be content to deal with an ensemble of high-resolution states, all of which will aggregate back into the same low-resolution state. In some instances, it may be adequate to select a single, typical high-resolution state from the ensemble, in which case the consistency definition just outlined is adequate (and there is a lot of redundancy in the high-resolution representation). In other instances we may need to acknowledge the variation within the ensemble.

In such cases we need the definition depicted in Figure 7. In this definition, we start in the *lower* left corner, with low-resolution inputs. Going first upwards, we disaggregate those inputs into an ensemble of high-resolution inputs. We use the high-resolution model to generate high-resolution outputs for all the inputs in the ensemble, and then aggregate each set of outputs. Taking the other path, we use the low-resolution model to generate low-resolution outputs from the low-resolution inputs. The two models are consistent in this new sense if the ensemble of aggregated high-resolution outputs is “comparable” to the low-resolution outputs.

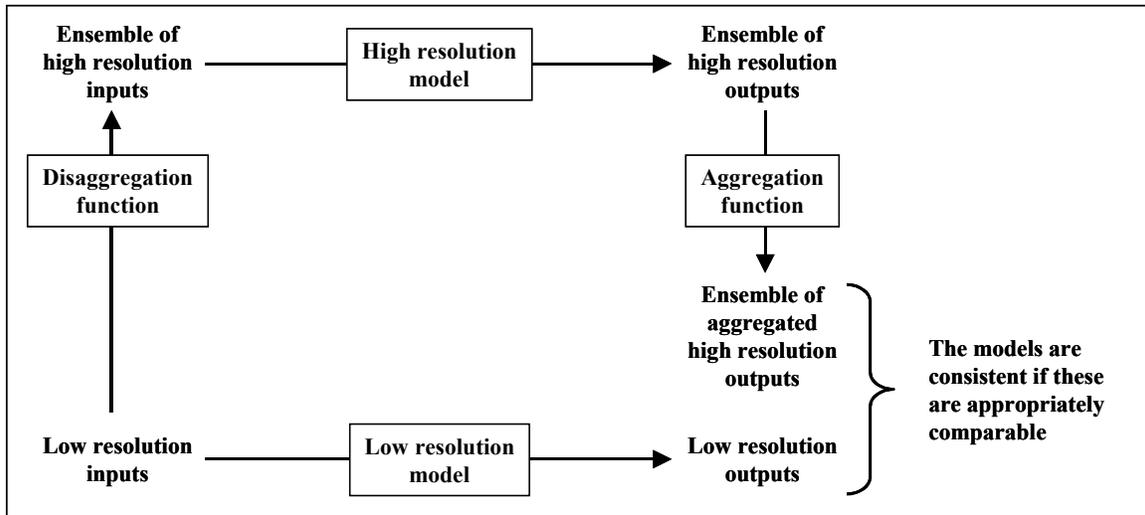


Figure 7: A More General Definition of Consistency Between High- and Low-Resolution Models

In principle, we should use an even more general definition of consistency. Since resolution is multi-dimensional, one model of a pair could have higher resolution in one dimension, and the other higher resolution in a second dimension.

Often, though, this extra complication will not be necessary. Perhaps the most common example of models with mixed resolutions concerns models with different scope. For example, Model 1 may represent long-range missile strikes in detail, and take as inputs the positions and velocities of target vehicles in the missile’s footprint. Model 2 might represent missile strikes as a simple kill probability multiplied by the number of targets in the missile’s footprint. But it might calculate vehicle movements in detail, including movements of vehicles far from any particular missile’s footprint. So Model 1 has greater resolution than Model 2 in the immediate neighborhood of a missile strike, but lower resolution (in fact, no information at all) at a distance. The analyst might well assume in this case that the positions and movement of vehicles outside a missile’s footprint have no effect on the lethality of the missile strike, and omit them from the consistency definition altogether – even if they seem to have an effect on the missile’s effectiveness as estimated by Model 2.

Right and Wrong Answers

The key, of course, is what we mean by “comparable” in Figure 7. Following the figure, let:

y = high resolution output
 \vec{xhr} = vector of high resolution input variables
 \mathbf{D} = domain of high resolution inputs
 $HR(\cdot)$ = high resolution model
 \vec{xlr} = vector of low resolution inputs
 $AggIn(\cdot)$ = function for aggregating high resolution inputs to low resolution inputs
 $LR(\cdot)$ = low resolution model

For simplicity, we will assume that the high-resolution model is deterministic. For any low-resolution input vector, the ensemble of high-resolution outputs is a set of values for y , and the output of the low-resolution model should be one or more attributes of that set, i.e.:

$$LR(\vec{xlr}) = \text{Attribs} \left\{ y \mid y = HR(\vec{xhr}), \vec{xhr} \in \mathbf{D}, AggIn(\vec{xhr}) = \vec{xlr} \right\} \quad (4)$$

But what attributes of this set should the low-resolution model estimate? If the set is a small interval—i.e., if there is little variation in the ensemble of high-resolution outputs—it is enough to estimate any y in the set. The first example in Appendix D provides an illustration.

The situation becomes so simple when the ensemble is a small interval (or even a single point) that it can be tempting to make it happen. The usual way to do this is to limit the domain \mathbf{D} of high-resolution input vectors that one considers. For example, the analyst may assume certain of the inputs to the high-resolution model are constant. Doubtless, this simplifies the low-resolution model, but it greatly limits its applicability. And when one conveniently forgets the limitation and applies it anyway, it gives spuriously precise results. We suspect this is one of the major reasons that low-resolution models have a bad reputation in some circles.

There are, of course, “honest” ways to limit the domain \mathbf{D} . For example, the object model may contain five or ten different kinds of Blue shooters. The analyst may have information that constrains the mix, either because a particular mix is infeasible or because it would be silly to choose it. Some shooters may require airfields with long runways, and there may not be many such airfields in the theater. There may be limited numbers of other shooters in the force structure. Some shooters may not be rapidly deployable. Such constraints may reduce the size of the ensemble of high-resolution cases that corresponds to each low-resolution case.

Even if the ensemble of high-resolution outputs is not a small interval, one can sometimes make do with a low-resolution model that estimates a single output. The second example in Appendix D outlines a low-resolution model that estimated the maximum size an algae bloom could attain under certain conditions. In the notation above, this corresponds to estimating the maximum value of y in the ensemble, i.e.,

$$LR(\vec{xlr}) = \text{MAX} \left\{ y \mid y = HR(\vec{xhr}), \vec{xhr} \in \mathbf{D}, AggIn(\vec{xhr}) = \vec{xlr} \right\} \quad (5)$$

Much, perhaps most, of the time, the ensemble of high-resolution outputs is not a small interval. Appendix D provides an illustration of this. In these instances we can’t capture all the

information we need about the ensemble in a single output, so the low-resolution model will have to estimate two or more outputs. The obvious candidates are a mean and a standard deviation, or the upper and lower bounds of a confidence interval, but these make no sense unless we can assign a probability distribution to the high-resolution input vectors. Without such a distribution, the most we can do is estimate the range of y in the ensemble.

But such a distribution will generally not describe the probabilities that various high-resolution cases would occur in the real world. Rather, the analyst will specify a distribution that captures the analytic importance of various regions of the domain D , or equivalently, he will design a sample of high-resolution cases. This problem is covered under the topic of experimental design in numerous statistics texts (e.g., see Saltelli et al, 2000). Given such a sample (or distribution), one can readily calculate means, standard deviations, and confidence intervals for the low-resolution results.

MOTIVATED METAMODELS, OR THE IMPORTANCE OF A GOOD STORY

Nobody denies the importance of basing a model on—or calibrating it to—data. In our view, however, a good story is equally important. The model cannot be a “black box;” it must tell a story about how things work in the relevant portion of the world. It must express a set of logical relations, cause-and-effect mechanisms on which to base inferences. The term “story” suggests that the explanation may be an ad hoc invention, and one might therefore prefer to use another term such as “theory” or “phenomenological explanation.” Sometimes the story can be quite speculative, however, and then alternative terms suggest that we know more than we actually do. But no matter how skimpy our knowledge, we consider it important to support a model with an explanatory and motivational story.

Not everyone agrees with this. Some consider it to be enough to fit equations to data, e.g., by statistical methods such as regression. In such a model, an independent variable is said to “explain” some part of the variation of the dependent variable, but this does not mean that a change in the independent variable causes the dependent variable to change. Rather, it means that the two variables are correlated, with possibly no causal relation at all. Before we would find such a model useful, we would need to identify the cause-and-effect relations—i.e., we would need to discover or construct the story.

There are several reasons that a good story is important. First, as mentioned earlier, it will be necessary to explain to the client why the model yielded the results it did and why those results are generally true, and not just true in the specific cases run. A persuasive story is invaluable for this task. The cause-and-effect aspect of the story is essential here, because the client (presumably a decisionmaker) wants to take actions that will cause the desired consequences.

Second, one step in validating the model is establishing its face validity. Face validation is the process of persuading subject matter experts that the model behaves reasonably, i.e., that for

reasonable inputs it produces reasonable outputs. How will they judge that the model is reasonable, if not by determining that it conforms to a good story or theory?

Third, a model is used to estimate things that cannot be observed directly.⁹ This means that the model will be used to extrapolate beyond the data one has in hand. We are acutely aware that extrapolation has a bad name. However, as Law & Kelton (1991) observe:

“The greater the commonality between the existing and proposed systems, the greater our confidence in the model of the proposed system. There is no completely definitive approach for validating the model of the proposed system. If there were, there might be no need for a simulation model in the first place.”

So extrapolation is where the action is. Somebody had to extrapolate beyond historical experience to suppose that precision-guided munitions would increase the capability of the force that used them. Today, we have to extrapolate beyond past experience to suggest that information technology will revolutionize warfare. The authors agree with both of these extrapolations, although the jury is still out on the second.

Extrapolation is never based on data alone; it is based on a function that one fits to the data. Extrapolated results depend crucially on the form of the function chosen. Appendix E illustrates this with a simple example. The form of the function, in turn, is suggested by plausible stories (or theories) about how the real world behaves. Thus the story influences how we build the model in the first place.

A story is just as important for a metamodel—a model of a model—as for any other model. Ideally, the story would be enough by itself. The low-resolution models in a family could be derived or inferred from physical considerations, or derived explicitly from the theory embodied in valid high-resolution models, or obtained by algebraically reducing or simplifying the high-resolution model. In practice, however, one must augment the story by applying statistical methods to data from high-resolution experiments.

A common method of building a metamodel is to run the high-resolution object model many times, collect the results in a dataset, and simply fit a response surface to the data. Appendix F describes an experiment in which we examined ways of constructing a metamodel of a given high-resolution object model, and the first method we tested was this “most common” approach. By judiciously selecting independent variables, we were able to achieve a fairly good overall fit (e.g., as measured by root mean square error).

However, we found ways do much better than to blindly regress the outcome(s) on the inputs. We defined transformations of the inputs, and various functions of them, to use as additional independent variables in the regression model. When these additional variables were chosen based on knowledge of the object model’s underlying theory, they vastly improved the regression results. This knowledge could be a description of how the object model works,

⁹ It might be possible in principle to observe these things directly, but very costly, or dangerous, or not possible soon enough, or otherwise inconvenient.

perhaps from documentation. Or it could be a coherent story about how the model works. Or, the knowledge could be a thought experiment about how the target model “ought” to work, i.e., how it would work if we had built it. We coined the term *motivated metamodel* to describe a low-resolution model whose structure is based on knowledge of these kinds (Davis and Bigelow, forthcoming).¹⁰

Appendix A provides another example of constructing a metamodel, this time a metamodel to estimate the effectiveness of missile salvos fired from long range at groups of armored vehicles. The object model was extremely large, and the analysis dataset was constructed from only about a dozen cases. But each case provided data on hundreds of missile salvos, so we had plenty of observations in our analysis dataset.

Each case from a large model provides thousands of data elements. To build a metamodel one must select which data elements should be included among the independent variables. In the example of Appendix A, we assumed that the number of vehicles killed by a missile salvo depended on the vehicles that were within a specified distance of the impact point (a footprint) and not obscured by foliage at the time of impact. It seemed only reasonable to exclude such variables as the number of vehicles very far from the impact point, and locations of vehicles at times other than the impact time. We offer these exclusions as examples of using knowledge of how the large model works (or how it should work) to help structure the metamodel. By reducing the set of candidate independent variables in this way, we vastly simplify the task of building the metamodel.

Having a story was also important because the model has parameters that remained constant within a case, but that we wanted to be able to vary when we employed the metamodel. In the high-resolution model of Appendix A, all cases assumed the same failure probabilities of the missiles and their warheads. Some parameters did vary, such as the transparency of the foliage and the delay time between selecting an aim point and a salvo’s impact. But few combinations of these parameters could be explored within the dozen or so cases we had available. To estimate the effects of these unvaried or hardly varied parameters required that we extrapolate and interpolate in ways that the data did not support.

Motivating the metamodel can be strategically important. For example, suppose one is dealing with a system that could fail if any of several critical components fail. Naïve (unmotivated) metamodels may fail to reflect the individual criticality of such components and may therefore be quite misleading if used for policy analysis. Naïve metamodels may be correct “on average,” but give misleading results on the relative importance of inputs, thereby skewing resource allocation decisions. Naïve metamodels may also fail in “corners” that seem to be obscure, but that can actually be focused upon by adversaries. Motivated metamodels can ameliorate such problems.

All in all, a good story (or theory, or phenomenological explanation) is an essential element of any model. It helps immensely in:

¹⁰ The manuscript has been reviewed and revised, but publication will likely not occur until late 2002 or early 2003..

- Building the model;
- Establishing its face validity;
- Generalizing and extrapolating; and
- Explaining and justifying results to the client.

To those who argue that data alone provide a sufficient basis for modeling, we offer this quotation. It was written about world affairs and international relations, but we believe it applies in the present context as well.

“[I]nvariably we operate with some kind of theory. It is sheer myth to believe that we need merely observe the circumstances of a situation in order to understand them. Facts do not speak for themselves; observers give them voice by sorting out those that are relevant from those that are irrelevant and, in doing so, they bring theoretical perspective to bear. Whether it be realism, liberalism, or pragmatism, analysts and policy makers alike must have some theoretical orientation if they are to know anything. Theory provides guidelines; it sensitizes observers to alternative possibilities; it highlights where levers might be pulled and influence wielded; it links ends to means and strategies to resources; and perhaps most of all, it infuses context and pattern into a welter of seemingly disarrayed and unrelated phenomena.”
(Rosenau, 1997, p. 92)

Appendix A: A Simple Story Contributes to Face Validity of a Complex Model

RAND has a suite of high-resolution simulation models for high fidelity analysis of force-on-force encounters. Our colleagues have used this suite of simulation models to assess the combat effectiveness of numerous weapon systems, including ATACMS (Army Tactical Missile System) with the BAT (Brilliant Anti-Tank) munition. In a 1996 study for the Defense Science Board (Matsumura et al, 1997), four cases were simulated in which a total of 88 ATACMS were fired, killing 283 enemy tanks, or about 3 tanks per missile. Over time, this figure of three kills per BAT warhead gained status as a reasonable estimate of the effectiveness one should generally expect from this weapon. “Three tanks killed per BAT” had graduated from a mere statistic to a rule of thumb (a particularly simple metamodel).

The transition to rule of thumb came about in this way. People were disappointed that BAT did not perform better in the 1996 study. Each BAT has 13 submunitions, which spread out over a huge footprint (4 kilometer radius) and independently use acoustic sensors to home in on targets. People had hoped that substantially more than a quarter of the submunitions would score hits. So the 1996 results came to be interpreted as a lower bound – “you can’t count on more than three tanks killed per BAT.” It seemed natural to begin using it as a conservative estimate of BAT’s effectiveness in general.

Clearly, granting rule of thumb status to this statistic was unjustified. Yet analysts do it all the time. We feel that we have no choice. If we merely show the client the results for a handful of scenarios, the client will surely ask us how those results generalize. We can say, “I think three kills per BAT is a reasonable general estimate.” Or we can say, “I can’t generalize from these results, and I strongly recommend you don’t either.” The latter would not be responsive to the client’s needs.

With this background, in a 1998 study for the Defense Science Board (Matsumura et al, 1999), six new cases were simulated in which 324 missiles were fired killing 142 armored vehicles (tanks and BMPs) were killed, or only about 0.44 kills per missile. At first, it seemed that we didn’t have to look far to rationalize the differences. In DSB ‘96 the terrain was entirely open, while in DSB ‘98 the terrain had considerable tree cover. In DSB ‘96 almost all the Red vehicles were armored fighting vehicles (AFVs), while in DSB ‘98 less than 20 percent of the Red vehicles were AFVs. And in DSB ‘96 the Red vehicles were in dense formations (50 - 100 meter spacing), while in DSB ‘98 vehicles were much more dispersed (150 - 600 meters).

But for each of these explanations there was a plausible counterargument. Thus in DSB ‘98, missiles were aimed only at clearings. One could argue that this should reduce the number of shots but not necessarily the effectiveness per shot. The BAT submunition preferentially homes in on AFVs, so the presence of trucks should have made little difference. And the fact that

ATACMS/BAT has such a huge footprint (a radius of at least four kilometers) should have negated the large separations between vehicles. So, what were the real reasons that ATACMS/BAT performed so much more poorly in DSB '98 than in DSB '96? Although the rationalizations were presented at the time ("rationalization" is the name for "story" when the story is not well supported by analysis), we were troubled by the questions.

Over the next year, as part of a small project for OSD, we used this puzzling experience as an opportunity to demonstrate what could be accomplished with multiresolution analysis of the phenomena. We examined the performance of BAT in every missile salvo simulated in the two studies. We found that BAT's performance differed in the two studies because, as Figure A.1 shows, there were generally many fewer AFVs in BAT's footprint in DSB '98 than there were in DSB '96. It was a rare salvo (of two missiles) in the DSB '96 study that had as few as 20 AFVs in its footprint, and often there were 40-100, while there were rarely more than 15 AFVs in the footprint of a DSB '98 salvo. As one would expect, then, salvos killed fewer AFVs when there were fewer AFVs in the footprint.

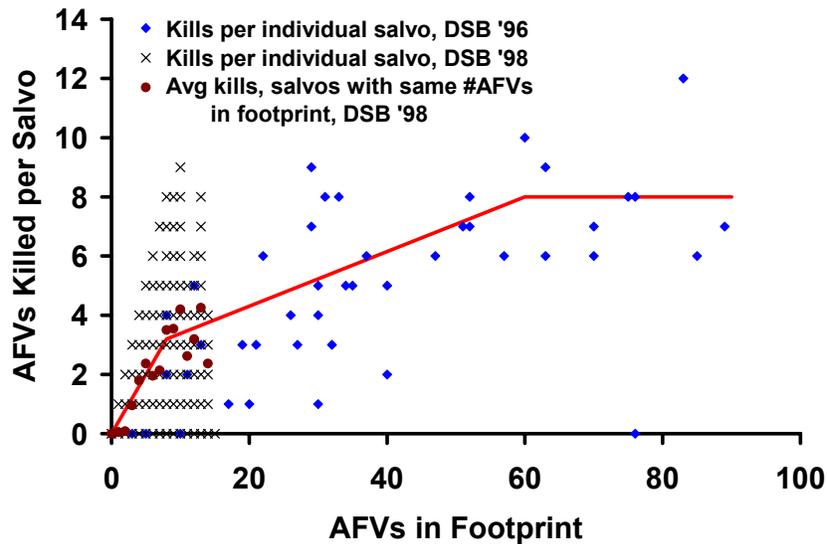


Figure A.1: Kills per Salvo (of two missiles) vs. AFVs in Footprint

This may seem like an obvious conclusion, but it was not easy to arrive at. The RAND suite of models, like most high-resolution models, consumes a great deal of input and generates an enormous amount of output. Producing a graph like Figure A.1 is a major undertaking, almost as great an undertaking as producing a similar graph from field experiments. One must manipulate a lot of data. One must define concepts such as "AFV" and "Footprint" in terms of the entities in the high-resolution model. Thus several different kinds of vehicles fall into the class "AFV," and while almost all kills occurred within four kilometers of the impact point, there were a few scattered kills at distances of six or seven kilometers. In our experience it is rare for an analyst to devote this level of effort to the examination of the inputs and outputs from a high-

resolution model. Yet without such an effort, why should one have confidence that the high-resolution suite was treating BAT in a reasonable way?

In addition, however, this analysis is not yet complete. After DSB '98, we knew that “three tanks killed per BAT” was not a good rule of thumb, and after the data analysis, we knew that the line in Figure B.1 was better – a kind of simple statistical metamodel. But to use it, we had to be able to estimate the number of AFVs in the footprint from variables under the control of Blue and Red. That created a new challenge.

With this in mind, we built a low-resolution theory-based model called PEM (PGM Effectiveness Model) (Davis, Bigelow, and McEver, 2000b). We structured it to be consistent with the functions of RAND’s high fidelity suite of models that are involved in interdiction, and calibrated it to the high-fidelity results. PEM is not quite small enough to be written on the back of an envelope, but it is nonetheless small and simple. We implemented it in Analytica™, a very flexible visual modeling tool.

PEM assumes that a column of Red vehicles is traveling along a road and moves through a clearing of width W . Rather than being uniformly spaced, the Red vehicles are grouped into packets, perhaps representing platoons. Each packet has N AFVs separated from one another by a distance S . Successive packets are separated by a distance P , which is larger than S . This column of vehicles moves through the clearing at a velocity V .

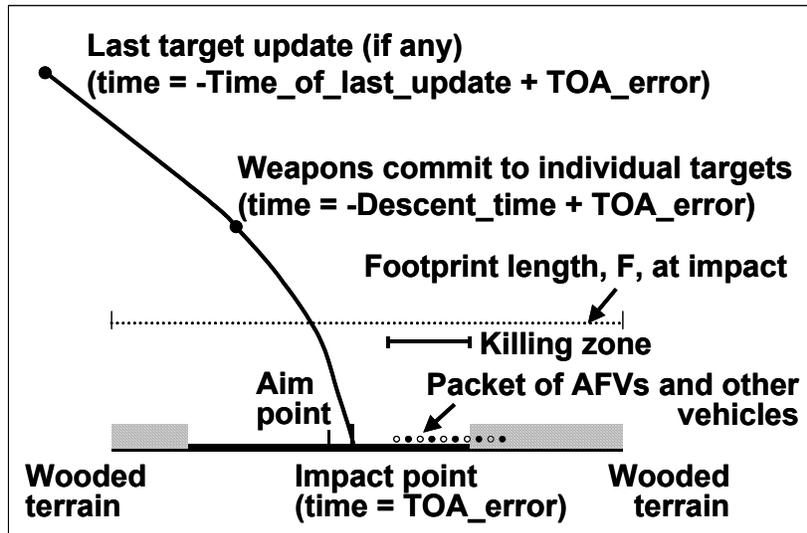


Figure A.2: PEM Concepts

Blue attacks the column by firing a salvo of one or more missiles at the clearing, timed to arrive when a selected packet is expected to be in the center of the clearing (see Figure A.2). But there is a random error in the arrival time (TOA_error) whose mean is proportional to the time since the missile last received information about the position of the target packet ($Time_of_last_update$). If TOA_error is too large, the target packet may have passed completely through the clearing, or beyond the weapon’s footprint F , whichever is larger; or (if the missile arrives early) the target packet may not have entered the clearing or the footprint. A smaller error will find the target

packet not centered in the clearing, and part of it may be hidden in the trees on either end of the clearing. Depending on the various parameters, parts of the packets just forward and rearward of the target packet may be in the killing zone.

Except for a few details, this determines how many Red AFVs are in the killing zone of the weapon at its time of impact. We can then use a function like the line from Figure 3 to estimate the number of AFVs actually killed.

One can think of PEM as a scaling function that adjusts the effect of long-range precision fires for the influence of a variety of factors. These factors include the time of last update, which operates through the error in the missile arrival time; the footprint of the weapon; the openness of the terrain; and the formation (including the dispersion) of Red's vehicles. We have used the model for exploratory analysis, to identify both positive and negative interactions among the factors. We have also developed an even simpler version of the model that could be used as a subroutine to incorporate these factors in other models. Of course, we could use "three tanks killed per BAT" in all the same ways. But PEM is much richer and much more plausible.

Returning to Figure A.1, note that there is a large variation in number of kills per salvo, even when the number of AFVs in the footprint is held constant. Some of this variation is due to the fact that the high-resolution model performs Monte Carlo trials to determine which BAT submunitions fail at one point or another in the process of acquiring and killing a target. But some of the variation may be systematically related to variables we have not considered here. Further analysis of the high-resolution inputs and outputs might suggest ways to increase the average kills per salvo, even as AFVs in the footprint are held constant.

What does all this have to do with validation? First, by developing PEM and using it to reason about the data from high-resolution simulation, we constructed a satisfying explanation for the high-resolution results, which had previously been troubling. Again, that explanation was at a much lower resolution than the original models and was not easily have been inferred from those models. Had we been unable to construct PEM, we would have been left to argue that the result must be right "because the high-resolution model says so, and it must be correct because of all the careful work that's gone into it." The fact that we could construct a low-resolution explanation thus tended to validate the high-resolution model. Second, having implemented the low-resolution explanation as a simple model (PEM), we could then use it to explore scenario space looking for the kinds of scenarios that would challenge BAT most severely (Davis et al, 2000b).

In the course of the work, we also found instances in which the correctness of the high-resolution models was questionable, or at least not well understood. For example, we tentatively concluded that large area acoustic noise due to AFVs was apparently having a substantial effect on results. It was questionable whether the estimates of this effect had been realistically calculated for mixed terrain. Also, with the benefit of hindsight, we concluded that the tactics used in the DSB '98 study had probably been unrealistically extreme, thereby further emphasizing the importance of being able to scale results to other cases, as was possible once PEM was created.

Appendix B: Selecting a Good Test Set of Detailed Scenarios

One of the generic problems in combat analysis is picking appropriate test cases for study with high-resolution simulations. Another problem is how to use a combination of field experiments and constructive modeling and simulation. We addressed both such issues in a 1999 study advising OSD on how to think about analysis efforts within efforts to “transform the force.”¹¹ The motivations involve both opportunities and necessity. By exploiting modern technology and new operational concepts, the Department of Defense expects that U.S. forces can greatly increase their capabilities and, in some cases, do so while reducing costs. At the same time, major changes will also be *necessary* to mitigate difficulties that can be posed by even mid-level rogue states. These include short-warning attacks and other so-called “asymmetric” strategies involving weapons of mass destruction (WMD), missiles, mining, high lethality conventional weapons, exploitation of urban sprawl and innocent civilians, and coercion of regional states resulting in access constraints.

A key mission of the newly created U.S. Joint Forces Command was to conduct experimentation in support of force transformation. It was often observed that such experimentation should employ a combination of models, simulations, war games, and field experiments. This observation, however, was rather abstract and many interpreted it to mean only that in both the work-up phase and follow-up phase of a major field experiment, one should use models and simulations. We had and continue to have a sharply different view, a view in which field experiments are seen as merely one part of experimentation, and not even the most important part scientifically. Further, in our view, field experiments are rare and valuable opportunities that should be used to collect information that cannot be obtained more readily in other ways.¹²

In Davis, Bigelow, and McEver (1999), we outlined an approach to such matters. The purpose of this work was to assist the DoD in thinking about ways in which analysis could guide and supplement research and experimentation. We recommended a hierarchical analysis approach, combining top-down and bottom-up methods. “Top-down” means starting with a broad, low-resolution view of the issue, and identifying a succession of higher-resolution formulations of narrower problems. “Bottom-up” refers to combining solutions to the high-resolution problems into an integrated response to the top-level issue. (The common use of these terms is yet more evidence that analysts routinely work at multiple levels of resolution.)

¹¹ The DoD first highlighted the transformation issue in its 1997 Quadrennial Defense Review. See also Davis et al. (1998), which documented suggestions one of us made in late 1996.

¹² To be sure, field experiments have other important objectives as well, notably demonstrating to high-level officers and officials that certain capabilities can be made real. This can be crucial for achieving support and acceptance.

At the top of our illustrative application of the top-down approach was the “operational challenge,” of bringing about the *early* halt of an invading armored column, even under stressful circumstances related to asymmetric strategies.¹³ Working down, we would develop variations of the challenges that stressed U. S. forces in different ways, and from them identify high-value building-block operational capabilities, adaptive integration capabilities, and cross-cutting functional capabilities. The process of identifying these capabilities also highlights the circumstances in which they are valuable, and suggests the scenarios in which they should be tested. In some cases, detailed analysis or even experiments would be needed to do the testing.

Any challenge problem in our methodology is the outer shell of a multi-layer structure. To say on a viewgraph that the United States should be able to bring about an early halt is straightforward, but accomplishing the challenge requires many building-block capabilities as shown here (an incomplete depiction). For example, on the left of Figure B.1 we see the requirement to establish quick and effective command-control and theater missile defense. That in itself is an extraordinary operational challenge, but in this depiction the related capability is a subordinate building block. In many cases, success of the whole requires success of all the parts. That is, we have a complex *system problem*.

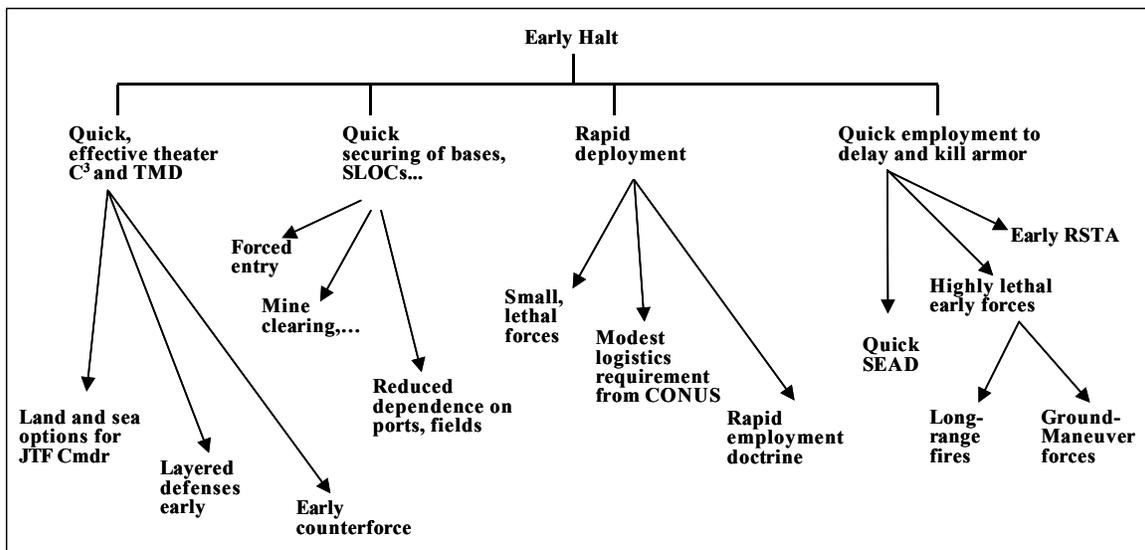


Figure B.1: Challenges Can Be Decomposed Hierarchically

Certain crucial features of the overall operational capability are crosscutting and therefore do not appear in any single branch of the decomposition tree. In particular, *all* of the subordinate operations are in our view likely to depend on network-centric command and control, long-range fires, effective operations with allies, forward presence, forward leaning during crisis, and mobility.

¹³ This problem was worked in much more detail in the context of asymmetric strategies in Davis, et al. (2002). It was also emphasized in that study that the same interdiction capability measured by a “halt distance” was quite relevant to counter-maneuver strategies that might be important in, e.g., an invasion of Iraq.

We can drill down further. We implemented a multi-resolution version of a halt-phase model called EXHALT (*exploring the halt problem*) in the Analytica™ modeling system (various versions are documented in Davis & Bigelow, 1998; McEver, Davis & Bigelow, 2000; and Davis, McEver & Wilson, 2002). Figure B.2 shows some of the factors EXHALT considers to determine the effectiveness of long-range fires (a building block on the “quick employment to delay and kill armor” branch) in halting the invading armored column. We measured effectiveness by how far the Red column penetrates. This depends on the time it takes to halt Red and his speed of advance. Moving down the left-hand branch, the halt time will depend on kills per day as a function of time, which in turn depends on the number of shooters of all types available to oppose the column and the effectiveness of each type of shooter (the underlining of these factors indicated they are vectors). As shown in the figure, these factors can be decomposed further.

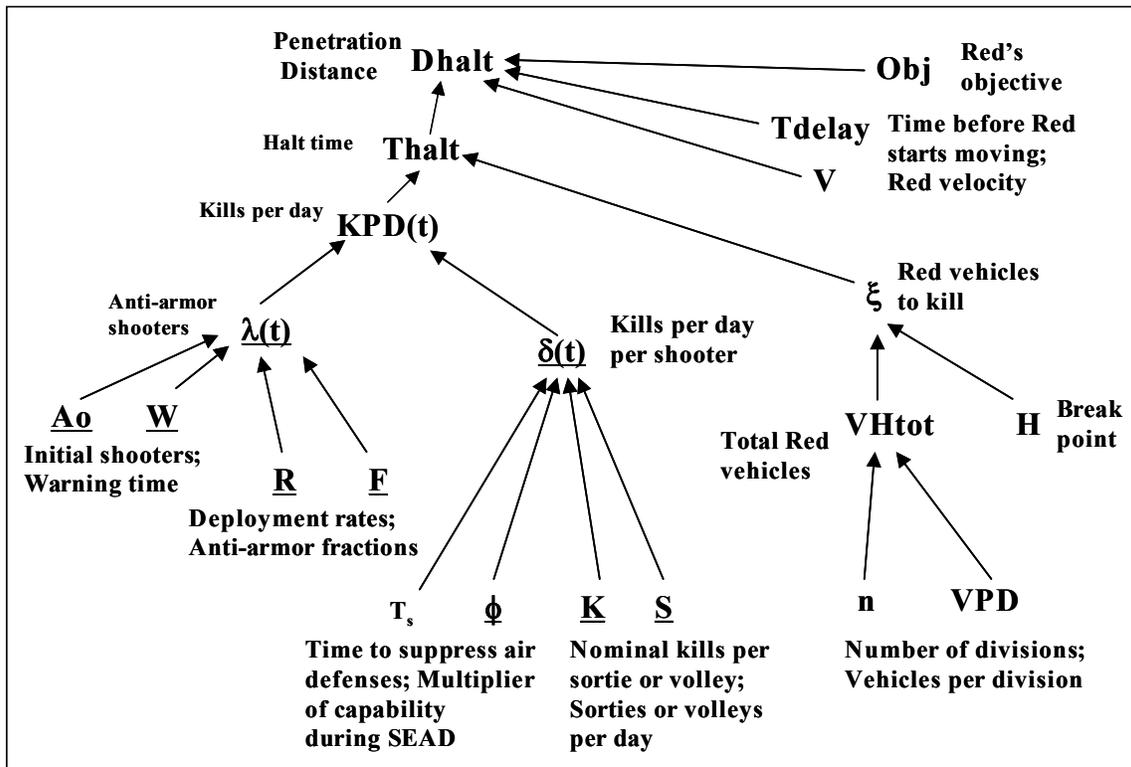


Figure B.2: Some Factors Considered in EXHALT

It is useful to express the factors in a hierarchy like this because one can visualize what combinations of factors are necessary to bring about a halt. To illustrate, let us truncate Figure B.2 at the level of “kills per day” on the left branch from $Thalt$, and at “Red vehicles to kill” on the right branch. If we make the approximation that $KPD(t)$ is zero until air defenses are suppressed and constant thereafter, we can then express the halt distance as:

$$D_{halt} = V \times \left((T_s - T_{delay}) + \frac{\xi}{KPD} \right) \quad (B.1)$$

Figure B.3 visualizes the combinations of Red velocity, wait time (the difference $T_s - T_{\text{delay}}$), and kills per day that are sufficient, under these assumptions, to kill 500 Red vehicles by the time the column has traveled 100 kilometers. Points on or below the surface will do the job; points above the surface will fail. Clearly, to succeed at this task, Blue must concentrate on measures that will either slow Red to a crawl, or reduce the wait time to a day or less. Note that in this figure, the wait time is the period between the time the Red column begins to move and the time Blue begins to attack it. So it may include warning time as well. Unless the wait time can be made small, the kills per day the Blue force can achieve hardly matters.

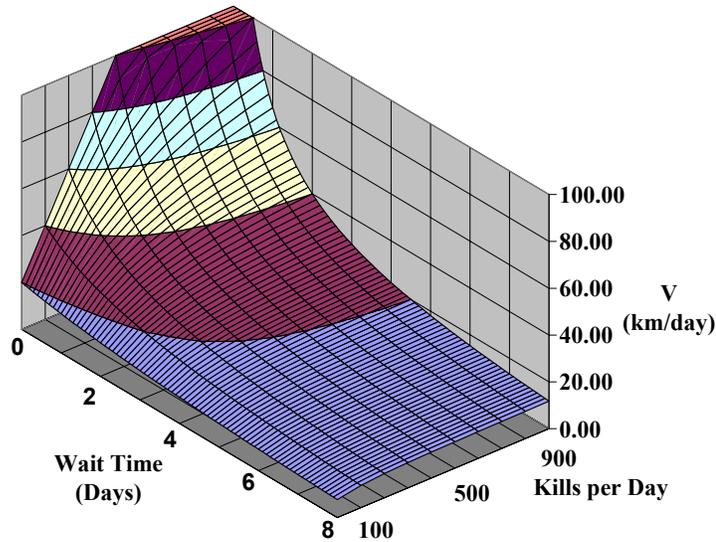


Figure B.3: Factor Combinations that Kill 500 Red Vehicles Within 100 Km

Next, consider the challenge of killing a much larger force (4000 vehicles), by only before Red has traveled 300 kilometers. Now (Figure B.4) it pays to improve any of the factors.

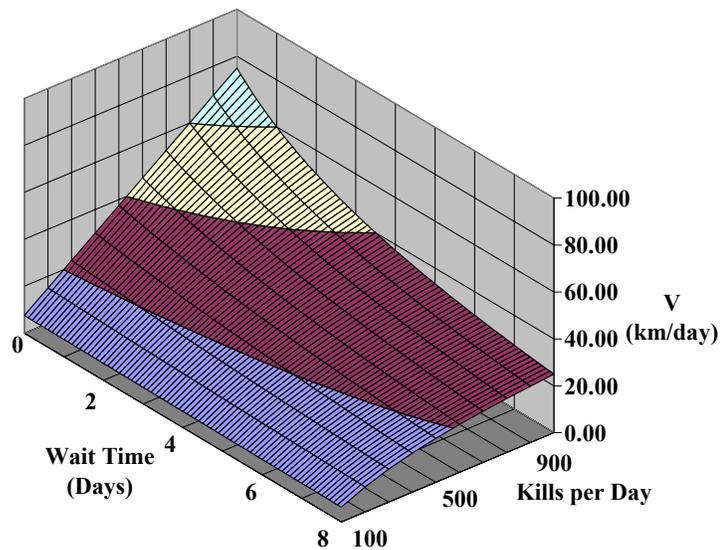


Figure B.4: Factor Combinations that Kill 4000 Red Vehicles Within 300 Km

So we find it is important to include in our test set of scenarios both a scenario that stresses the rapid accomplishment of a relatively small task, as well as a scenario that stresses the slower accomplishment of a very large task. Note that a single worst-case scenario won't do, i.e., one that requires halting a very large and fast force within a day or two. True, if we can accomplish the worst-case scenario, we can accomplish the two lesser scenarios. But the cost of a worst-case capability may be prohibitive. For example, it might require us to buy a force consisting entirely of very rapidly deployable, highly lethal weapons capable of operating from austere bases. Weapons with all of these capabilities will have very high unit costs. By contrast, to accomplish both lesser scenarios we might buy a couple of squadrons of not-so-lethal but highly deployable weapons, capable of operating from austere bases. These weapons would be first on the scene, and could accomplish the scenario that stresses the rapid accomplishment of a relatively small task. We would supplement them in the longer, bigger scenario with more lethal weapons that needed more time to deploy and operated only from better-prepared bases. This mixed force would doubtless be considerably less costly than the worst-case force.

Of course, the above discussion does not complete the top-down development of scenario and building blocks. Figure B.2 expands on the parameters we explored in Figures B.3 and B.4, and the other branches and building blocks in Figure B.1 remain to be explored. The result of this analysis will be a list of high-value building blocks, each described in terms of a combination of capabilities. The process of generating them also highlights the circumstances in which they are valuable, and suggests the scenarios in which they should be tested. Detailed analysis will be needed to do the testing.

Appendix C: Using a Low-Resolution Calculation to Check a High-Resolution Calculation

In this example we estimated selected logistics requirements of the Army's Light Helicopter, Experimental (LHX), which has since been re-designated the RAH-66 Comanche (see Smith et al, 1988). The requirements we estimate are the cost per flying hour of replenishment spares and the cost per flying hour of replenishment repair parts. Spares are parts that can be removed from the aircraft, repaired, and reused. Repair parts are used once and thrown away. These cost coefficients are labeled replenishment costs to distinguish them from initial costs. When any service buys an item of equipment, they buy initial spares and repair parts to stock the supply system. Subsequent operation of the equipment will break spares and repair parts, and those stocks will need to be replenished.

Traditionally, these cost coefficients are estimated from historical data (Figure C.1). The resulting lines correspond to linear empirical models. Such empirical models are a special case of low-resolution models. They are not "metamodels" as the term is used here: they are based on empirical data, not some more detailed model.

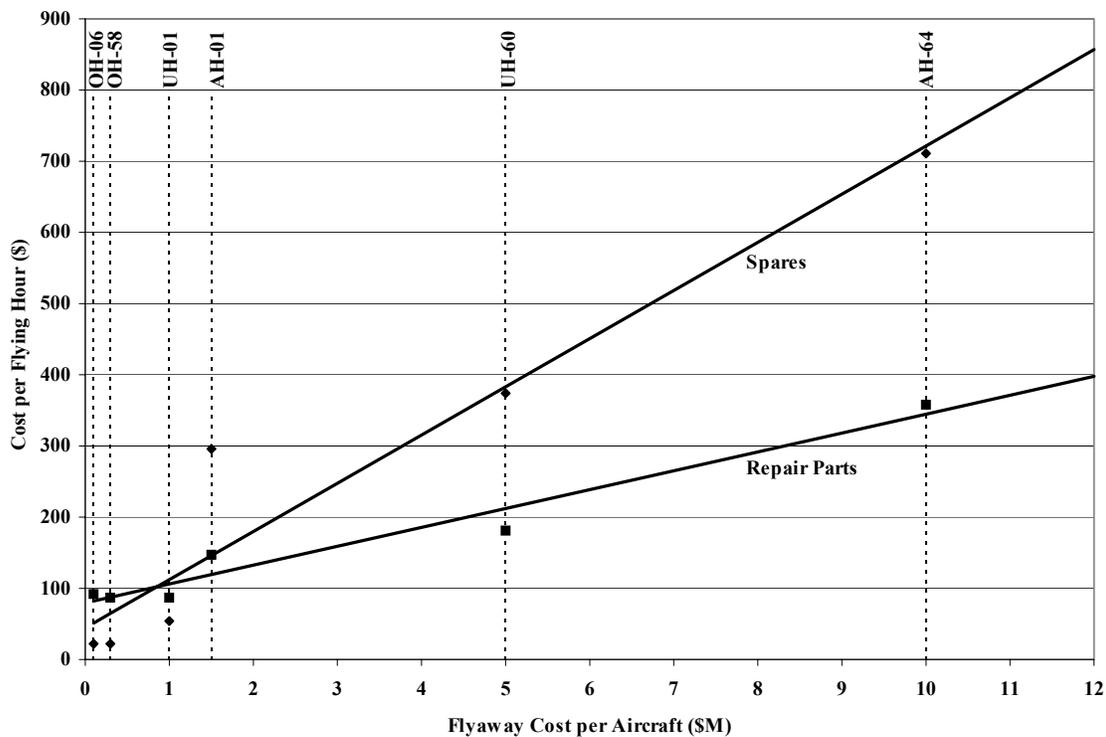


Figure C.1: Replenishment Cost Factors, Traditional (Aggregate) Method

In Smith et al (1988), one of us (Bigelow) and colleagues used data for six existing helicopters (the OH-6 Cayuse, the OH-58 Kiowa, the UH-1 Iroquois—better known as the “Huey,” the AH-1 Cobra, the UH-60 Black Hawk, and the AH-64 Apache). These were the helicopters whose missions the LHX was intended to perform. We plotted the costs per flying hour against the investment cost per helicopter, and found that the six historical points lay reasonably close to straight lines. When we plotted the LHX at its projected investment cost (\$10.6 million), the straight lines led us to expect the cost elements ought to be over \$700 and \$350 per flying hour, respectively.

The Army was unwilling to use this simple, aggregate model for estimating these cost factors. Army analysts had adopted a reliability improvement program that was intended to reduce these costs, and the traditional method has no “hooks” for including its effects. So in their baseline cost estimate (BCE) for the LHX, they estimated these two cost elements one spare and repair part at a time. They estimated a unit price for each item, a mean time between failures, and for spares the fraction of the item price that would be spent on repairs and the fraction of items that could not be repaired (the condemnation fraction). The Army was optimistic; they found a close analog for each item on their existing helicopters and used the one with the lowest cost in their LHX estimate. They also adjusted the costs downward to account for a reliability improvement effort that was part of the LHX program. The result put the two cost factors at about \$35 and \$16 per flying hour for spares and repair parts, respectively, or about a twentieth of the estimates made using the empirical model.

There were several reasons for the discrepancy. Part of it, of course, was the unbridled optimism of the Army analysts. A new weapon system is always cheaper and more reliable, not to say more effective, than anything you already have.

But most of it was due to the fact that the detailed model—the model that describes what happens to an individual item—was an unrealistic idealization. A high-level view of this multi-echelon inventory model, used by all the services, is based on the network shown in Figure C.2.

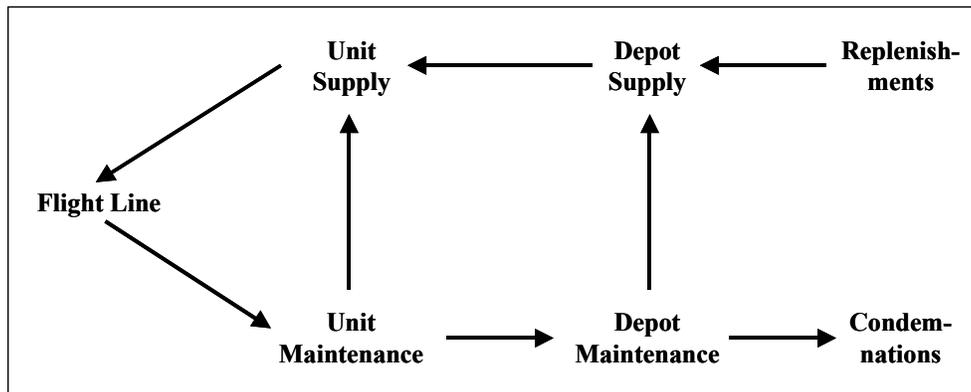


Figure C.2: Multi-Echelon System for Processing Recoverable Items

As helicopters fly, radars, fuel pumps, or engine parts may fail. When a part fails, it is removed at the flight line and sent to unit maintenance for repair. A replacement is drawn from unit

supply, if one is available. Unit maintenance tries to repair the item and, if successful, returns it to unit supply to replenish their stock. If the repair is too difficult for unit maintenance, personnel there return the part to the depot for repair and order a replacement from depot supply. Depot maintenance repairs the item and sends it to depot supply.¹⁴ As the system operates, the inventory of each item will be distributed among the transportation and repair pipelines. Items in transit between the unit and the depot in both directions are in transportation pipelines; items in repair at both the unit and depot are in repair pipelines. The number of items in a pipeline will depend on the transportation or repair time, and will equal the number that have entered during a period that equals the pipeline time.

In theory (i.e., assumed in the model), initial procurement is supposed to provide just enough items to fill the pipelines. Over time, the system will lose items due to condemnations. Replenishment spares and replenishment repair parts are intended to replace these condemned items through procurement of items from manufacturers. Consistent with this theory and model, the Army BCE calculated the costs per flying hour of replenishment spares and repair parts as the costs of replacing condemned items.

The theory, however, has many shortcomings. For example, initial procurement may provide incorrect inventories of some parts. If more items are provided to the system initially than are needed to fill the pipelines, the excess will migrate to intermediate and depot supply. There will be no need to replenish this item until the excess is used up. On the other hand, if too few items are provided initially to fill the pipelines, parts will migrate out of war reserve stocks or, in an extremity, be made up by leaving “holes” in helicopters (i.e., not replacing items that were removed). In this case, requirements for replenishment may vastly exceed condemnations.

Nor is misestimation of initial procurement the only cause for replenishments to differ from condemnations. Removal rates of items at the flight line vary for many reasons (Crawford, 1988). The military frequently modifies its aircraft, either to improve safety or increase capability. These changes can cause removal rates to increase or decrease. If an expensive recoverable item has a high removal rate, engineers will redesign or reinforce it, or design modifications to other portions of the aircraft to relieve stress on the item. If aircraft are redeployed to a location with extreme environmental factors (e.g., temperature extremes, sand storms), removal rates can rise. In all such cases, there are procurement demands that are not due to condemnations. Indeed, since condemnations tend to be very low, purchases to fill pipelines dominate replenishment.

Perhaps some individuals at the time expected the Law of Large Numbers to rescue the Army cost estimates. That is, we could consider the demand rates to be random numbers, and this would have made the contribution of each item to its cost factor a random number as well. Since

¹⁴ The network for repair parts is simpler. Repair parts are discarded when they are removed from the helicopter, and replaced from intermediate supply. Intermediate supply is replenished from depot supply. Repair parts are not cycled through maintenance facilities for reuse. But the model described here can represent repair parts as recoverable items with condemnation fractions of one. So our comments about the deficiencies of this model for estimating the cost per flying hour of replenishment spares also apply to estimating the cost per flying hour of replenishment repair parts.

there are thousands of items per weapon system, why shouldn't the sum of all these contributions converge on the right answer?

The answer is one we all know well. The expected value of a function of a random variable x is generally not equal to the function applied to the expected value of x . That is:

$$E(f(x)) \neq f(E(x))$$

We do find equality, of course, if $f(x)$ is a linear function, and it may seem at first sight as if we have a linear function in the present case. In fact, if we hold all the other parameters constant (i.e., pipeline times, fraction repairable at unit maintenance, and condemnation fraction), then the total required inventory of an item is proportional to the demand rate. But it isn't the total inventory that contributes to the cost factor; it is the incremental inventory. It is the amount we must buy *this year*, not the total amount we must have on hand. For this year's buy we must calculate the total inventory required, and then subtract the inventory we already had on hand. And then we must *truncate* the result, for if it happens that the increment is negative, then the contribution of this item to the cost factor is zero. We cannot sell any of our current inventory.

In addition, the other parameters don't in fact remain constant. If a spare (a repairable item) can't be repaired at unit maintenance, the engineers will try to redesign it, or to design the appropriate tools so that it becomes repairable at unit maintenance. The engineers will try to fix parts with high condemnation fraction as well.

So, why did the Army use this model to estimate the replenishment cost factors? Inventory models based on the multi-echelon structure of Figure C.2 are used by all the services and the Defense Logistics Agency (DLA) to manage items in their supply systems. They work very well for most items most of the time and they signal the item manager when an item begins to misbehave. At that point, the item manager takes action, for example by assigning a high repair or transportation priority to an item (reducing repair and transportation times), or redistributing the items in the system (Hodges, 1980).

Nor does the fact that the item manager must intervene from time to time mean that the model is not good for its management purpose. Herbert Simon (1982a, pg. 197) describes a study that devised rules for inventory and workflow smoothing. He later commented (Simon, 1982b, pg. 434) that he had simplified the description of the real world to make his model mathematically tractable, and he never expected that the optimal solution to the idealization would be optimal in the real world. He was relying on the people actually implementing the policy to intervene from time to time when the policy went off track.

In short, the detailed model meets the DMSO definition of validity in regards to its use as an item management tool. But it is not valid for the use the Army made of it in their cost estimating.

In a related effort, one of us tested the validity of the theory in another application, one using Air Force data for selected aircraft (Bigelow, 1984).¹⁵ The model calculated a very large requirement to buy spares in the first year, and much smaller requirements (about ten percent of the first year requirement) in each subsequent year. The first year's requirement sufficed to redress the inventory discrepancies, due to the difference between historical and current demand rates – the same problem that derailed the Army helicopter-cost estimates. Once redressed, the model assumed such discrepancies would never recur. Actual requirements have never shown this drop after the first year, suggesting that these discrepancies are a permanent fact of life. Empirical models, then, however simple, reflect real-world considerations that often do not appear in theory – even in meticulously developed high-resolution models.

Strictly speaking, this is not an example of metamodeling or MRM. But it does suggest that working at multiple levels of aggregation can reveal errors that would remain hidden if one worked only at a single level.

To cast this result in MRM terms, these two methods for estimating the cost factors are both based on widely used models. The low-resolution empirical method uses statistics (or less formal means) to fit historical data. Cost analysts are very comfortable with this method; they have been using it since the beginning of cost analysis and systems analysis in the 1950s. It is based on the supposition that the future will look like the past, or (more generally) can be extrapolated from past trends.

The higher-resolution method is based on an idealized description of the process that generates the demands for spares and repair parts. Because it looks explicitly at the process, it has the “hooks” for estimating the effects of changing aspects of the process. But because it is an idealization, not everything that can be estimated with this method is right.

These “hooks” provide the means for extrapolating from historical practices to new practices. In this case, the “hooks” are needed to represent an attempt to design reliability into the LHX. However, there is nothing in the historical records that represents attempts to do this for previous helicopters, so no variables can be included in a statistical model based on historical data. The “hooks” must be based on thought experiments (aka theory). This raises the question of just how effectively such a method with these hooks can be validated. The whole point of using the detailed method, the method with these “hooks,” was to allow extrapolation into an unobserved realm.

Unfortunately, the two methods are not mutually consistent. The exercise of making them consistent – which was never done – would presumably have yielded more defensible estimates of the annual cost of replenishment spares and repair parts, one which provided a plausible explanation of deviations of the estimates from the historical data.

¹⁵ The main purpose of this study was to build a metamodel of the item-level model described here, so that OSD could quickly estimate the impact of changes in flying programs on elements of the Air Force budget. The invalidation of the item-level model to predict the requirement for replenishment spares dollars was an unanticipated byproduct.

Appendix D: Illustrations of the Use of Consistency Definitions

The Ensemble of High-Resolution Outputs Can Be a Narrow Interval

In 1975, RAND conducted the Policy Analysis of the Oosterschelde (POLANO)¹⁶ study for the government of the Netherlands. In 1953 a flood, caused by a very large storm sweeping down the North Sea, devastated the Delta region of the Netherlands. After recovering from the immediate effects, the Government decided that they would build engineering works in the Delta region to protect the area and its population and resources from any future flood.

The Oosterschelde was the last estuary whose protection was undertaken. The original plan called for the Oosterschelde to be dammed off from the sea, creating a freshwater lake. By 1974, however, controversy had arisen over the damage this would do to the ecology of the Oosterschelde, and to its thriving oyster and mussel fishing industry. So the Dutch Cabinet directed the Rijkswaterstaat (RWS), the government agency responsible for water control and public works, to assess alternative protection plans. The RWS turned to RAND for help in this assessment.

Ultimately, three plans were considered. The original plan was to close the Oosterschelde off completely. One alternative was to leave it open and build massive new dikes around its perimeter. The third alternative, and most challenging to assess, was to build a storm surge barrier, a dam with gates that could be left open most of the time but closed if a large storm threatened. Different variants of the storm surge barrier had different aperture sizes, each size producing a different reduction in the tide and hence a potentially different effect on the Oosterschelde's ecology.

We needed to estimate peak water levels at each dike section around the Oosterschelde, for a wide range of storm scenarios and barrier apertures. (We created additional estimating relations to account for the incremental effect of wind and waves on peak water levels by dike section.) By comparing the peak water level at a dike section with the height of that section we could estimate the likelihood and severity of flooding for different scenarios. In the follow-on Barrier Control (BARCON) study,¹⁷ we needed test barrier control strategies, i.e., how quickly should the barrier be closed, what should be the target water level inside the basin while the barrier was closed, and what should be the relation of inside and outside water levels when the barrier was opened.

¹⁶ This description of the project has been paraphrased from the preface of Bigelow et al, 1977b.

¹⁷ Once the Dutch Cabinet had determined to protect the Oosterschelde estuary by building the storm-surge barrier, they commissioned RAND to investigate the question of what conditions ought to trigger the closing and reopening of the barrier.

The Rijkswaterstaat (RWS) had developed a model called IMPLIC that calculated water levels as a function of time throughout the Oosterschelde. It used an implicit finite difference scheme to integrate partial differential equations, hence the name IMPLIC. Its calculated water levels and flows throughout the basin matched measured water levels very closely.

We constructed a model we called SIMPLIC (Simple IMPLIC) to estimate the same water levels, but at much less cost (Abrahamse et al, 1977; Catlett et al, 1979). It consisted of a single ordinary differential equation, which could be integrated much more economically than IMPLIC's partial differential equations. Where IMPLIC required the user to input water levels over time at many points on a section across the mouth of the estuary, IMPLIC required only the average water level over time at the mouth. Similarly, IMPLIC required a detailed description of the basin, including depths at every point relative to a standard elevation. SIMPLIC merely needed the surface area of the basin as a function of average water level. So it was much easier to specify cases for SIMPLIC, and SIMPLIC was much less costly to run.

Needless to say, SIMPLIC did not reproduce IMPLIC's water levels exactly. IMPLIC successfully represented basin resonance phenomena (i.e., the phenomena that make tidal amplitude different at different locations) and the phase shift of the tide at different points within the basin. SIMPLIC predicted the water level at only a single reference point in the basin, and we had to introduce a calibration curve to adjust the SIMPLIC results for other locations around the basin. We actually did not care about phase shifts. We were interested in estimating the peak water level achieved at each point in the basin during each tidal cycle, and that had almost zero dependence on the phase. Once we calibrated SIMPLIC for the resonances of the basin, SIMPLIC's estimates of peak levels per tidal cycle matched IMPLIC's estimates within two or three centimeters.

In this example, the ensemble of high-resolution results would consist of all scenarios that produced a given peak water level at the reference point used in SIMPLIC. But even after accounting for resonances and phase shifts, IMPLIC's predictions of peak water levels at different locations are very highly correlated. Once the peak water level at one location is known, the peak levels at all other points can be predicted with very little error. This is equivalent to saying that if we were to look at all cases in the ensemble, we would find little variation in peak levels at any given location. Hence SIMPLIC is consistent with IMPLIC according to the first definition of consistency given in the main text.

Selecting the Maximum of the Ensemble of High-Resolution Outputs

During the POLANO project, we developed a model to predict the severity of algae blooms and the effectiveness of methods for controlling them. We used the model in that project and in a later project PAWN (Policy Analysis of Water Management for the Netherlands). F. J. Los later developed the model further at the Delft Hydraulics Laboratory (Bigelow et al, 1977 and 1982; Los, 1991).

Algae—otherwise known as phytoplankton—are single celled waterborne organisms that consume nutrients (nitrogen, phosphorus, and for some species silicon) plus energy from

sunlight in order to grow. Given enough of these resources, a population of algae can grow large enough to become a nuisance. They can cause the death of plants on the bottoms of water basins by shielding them from sunlight. They often produce substances that are toxic to fish and shellfish. They can clog filters in water systems. And when the population dies off, the process of decomposition can exhaust the dissolved oxygen in the water, causing fish kills and bad odors.

Most published models of algae growth simulate the population of algae over time, as it grows from a small to a large concentration. For our purposes, however, the peak concentration is a good measure of the severity of an algae bloom. So we built a model that estimates the peak directly, without trying to map the trajectory by which the bloom arrived at the peak. We formulated the model as a linear program, in which the variables were the concentrations of different species of algae, and the constraints ensured that the algae population did not outgrow the available nutrients and solar energy. The model calculated the mix of species that maximized total biomass.

We were aware that the model could overestimate the peak of a bloom. It takes time for algae to grow, and the conditions most favorable for algae might persist for too short a period. But we argued that sometimes overestimating the bloom was acceptable for policy purposes, so long as blooms reached or nearly reached their “theoretical” peaks with reasonable probability.

An Example That Requires Ensembles of High-Resolution Cases

Consider the rule of thumb that in order to succeed, an attacking force must have a superiority of 3:1 over the defending force. If this rule is valid for a small sector along the line of battle (i.e., in detail), what can we say about its applicability to the theater as a whole (i.e., in the aggregate)? (See Davis, 1995.)

Suppose that Blue has 1 unit of force in the whole theater, and Red has 2 units. Suppose there are two sectors on the battle line, and the 3:1 rule is true in each sector individually. Clearly, if Blue and Red both decide to allocate their forces equally between the sectors, the ratio in both sectors will be only 2:1 in favor of Red. This is smaller than 3:1, so a Blue defense will be successful. But if Blue allocates his forces equally, and Red “loads up” on Sector 1, Red will be able to create a 3:1 or greater advantage there while maintaining a safe 1:3 or better ratio on Sector 2. Now Red can attack successfully on one sector, and having defeated half of the Blue force, presumably fall upon the other half and finish the job. On the other hand, if Blue can detect Red’s concentration of forces on Sector 1, he can reallocate his own forces to counter Red’s move.

The situation becomes even less clear if we divide the line of battle into more sectors – ten sectors, for example. Let Red allocate his two force units however he chooses. Let Blue then allocate two-thirds of his one force unit to mirror the Red allocation. This is enough to defend successfully on every sector. Now let Blue put his remaining one-third force units on the sector where Red is weakest. Red must have no more than a tenth of his force, or 0.2 force units, on the weakest sector, so Blue will enjoy a superiority on that sector greater than 3:1. It appears Blue can attack successfully, and defeat the Red force in detail.

In this example, the low-resolution case is characterized by the theater-wide force ratio of 2:1 in Red's favor. The corresponding ensemble of high-resolution cases consists of all allocations of Blue and Red forces over the sectors that adhere to the 2:1 constraint. That is, if B_j and R_j are the Blue and Red forces allocated to sector j , then the high-resolution ensemble consists of all solutions to the constraints:

$$\begin{aligned} \sum_j B_j &= 1, B_j \geq 0 \\ \sum_j R_j &= 2, R_j \geq 0 \end{aligned} \tag{D.1}$$

Clearly, the fact that the theater force ratio 2:1 does not constrain the ratio on any given sector. There are many solutions to constraints (D.1) in which different sectors have force ratios that deviate from 2:1. Without additional constraints, we cannot predict the theater outcome from the theater force ratio.

Unless the force allocation to sectors was an input, a high-resolution model would have to include a force allocation algorithm. This might be modeled as an OODA¹⁸ loop. Each side would use its intelligence assets to observe the opponent's force allocation, and would estimate the opponent's intentions concerning changes to the allocation. Each side would then decide how to change its own allocation, and would proceed to do so. If one side had substantially better intelligence and mobility, that side might be able to largely dictate the sector-by-sector force ratios. If each side could quickly observe what was happening, but only slowly change its own allocation, neither side could readily change the initial allocations. If each side could move its forces very rapidly, but was unable to detect the opponent's movements, the theater campaign would be a crapshoot. In none of these cases would it be appropriate to apply the 3:1 rule at the theater level.

¹⁸ Observe, orient, decide, act.

Appendix E: Basing Extrapolation on a Story

In this Appendix, we illustrate the need to use a story or theory as a basis for extrapolating beyond the available data. We generated a data set consisting of ten points (y_i, x_i) , using the following model:

$$y = f(x) + \varepsilon \tag{E.1}$$

The residual ε is a Normal random variable with mean zero and standard deviation 6. We will withhold the function $f(x)$ for the time being. Table E.1 shows the values of the independent variable x and the dependent variable y . Because we generated the data set, we can provide the “true” value of the dependent variable, $f(x)$, as well.

Table E.1
Sample Data Set

x	y	$f(x)$
3.0	9.72	10.98
3.5	20.13	15.00
4.0	12.01	19.28
4.5	19.76	23.62
5.0	25.35	27.89
5.5	43.58	31.95
6.0	25.46	35.70
6.5	31.86	39.07
7.0	42.32	42.00
7.5	42.82	44.47

Given these data, we wish to estimate the value of y that corresponds to $x=15$. In addition, we want to state how uncertain we ought to be about our estimate. The standard method is to use statistical regression to fit a straight line to these data. When we regress y on x , the least-squares fit is:

$$\hat{y} = -10.1 + 7.12 \times x \tag{E.2}$$

Substituting $x=15$ into this equation yields our estimate $y=96.76$.

There are three sources of uncertainty in our estimate (we prefer the term *uncertainty* to *error*). First, there is a *residual uncertainty*. We know that Model (E.1) has a residual term, which by construction is a Normal random variable with mean zero and standard deviation 6. If this were the only source of uncertainty, the 90 percent confidence interval for our estimate would be $96.76 \pm 1.645\sigma$, or (86.83 – 106.63).

Second, there is a *calibration uncertainty*. Calibration uncertainty arises from the uncertainty in the values of the intercept and slope of the regression equation. To assess the size of this uncertainty, we generated 20 sets of residuals using the Excel RAND() and NORMINV() functions, calculated the least-squares linear function for each, and use them to make 20 estimates of the value of y at $x=15$. We estimated the 90 percent confidence interval for calibration uncertainty by dropping the minimum and maximum estimates. The range of the remaining 90 percent of the estimates was from 88.81 to 123.29. This interval is about 1.8 times the width of the 90 percent confidence interval for the residuals.

Third, there is *structural uncertainty*. We chose to fit a linear model to the data, and there is no clear indication from the data that it is a bad choice. But that is no guarantee that it is a good choice. In fact, we generated the data using a function:

$$f(x) = x^3 e^{-0.3x} \tag{E.3}$$

As shown in Figure E.1, this function looks very nearly linear for $x \in [3, 7.5]$, but it extrapolates to larger x in a very nonlinear fashion. Indeed, $f(15) = 37.49$, a much smaller value than any of the linear fits could suggest. Figure E.1 also shows the minimum, maximum, and nominal linear fits, and the data from Table E.1.

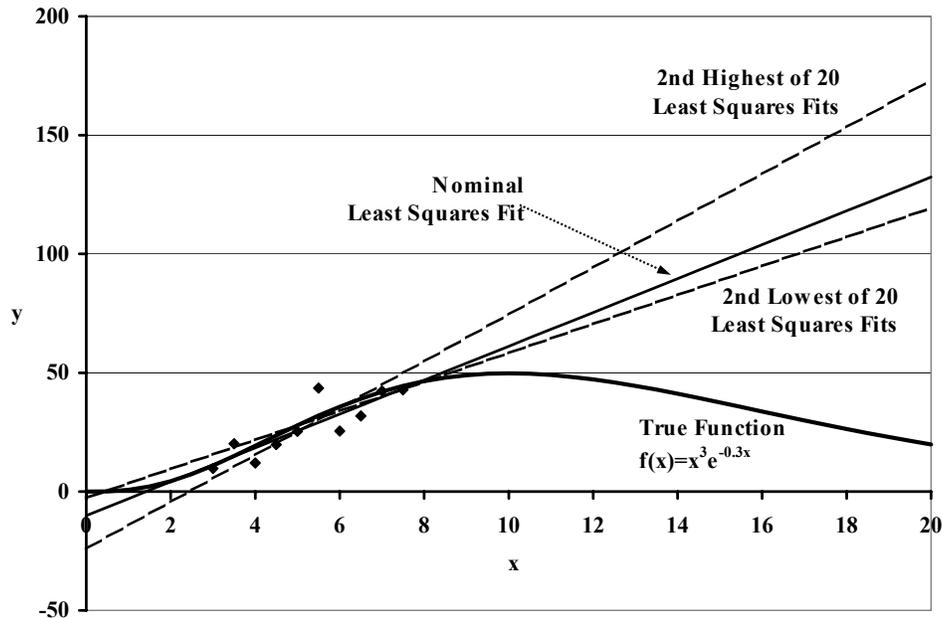


Figure E.1: Calibration and Structural Uncertainty

The problem, of course, is that we are extrapolating beyond the range of the data. The data cannot reliably guide us in this task; something more is needed. We suggest that the “something more” is a good story, an explanation for how the data came to be. This explanation can then be used to suggest what hypothetical data might look like in ranges not covered by the actual data. For example, the physical meaning of the variable y might preclude its being negative, which

would imply that a linear function with a negative intercept is a poor fit. Similarly, one might reason that y cannot become arbitrarily large. A more powerful story might describe the process that generates y from x as involving two factors. One causes y to increase with increasing x , while the other has the opposite effect. The former factor might be expected to dominate when x is small, the latter when x is large. In the present example, of course, the first factor is modeled as x^3 , the second as $e^{-0.3x}$, but other functional forms could be proposed. To be less abstract, if x represents time, x^3 might represent the volume of an expanding cloud, e.g., from a rocket, and e^{-bx} might represent the decay of some chemical process.

A good story may be needed for interpolation as well as for extrapolation. Indeed, when we use a linear or quadratic function to interpolate between points in a dataset, we are implicitly assuming that the function is smooth. When fitting a sample from a waveform with sine and cosine function, we acknowledge that the result will match the actual waveform (as opposed to the data sample) only if the waveform contains no components with wavelengths much shorter than the distance between successive data points. “The function is smooth at this scale” isn’t a very complicated story, but it’s a story nonetheless.

This is not to say that “the function is smooth” is the only story one can use for interpolation. Nor, of course, is it necessarily the right story. In particular, the sparser your data, the more one might question it.

Appendix F: Motivated Metamodels

Davis & Bigelow (forthcoming) describe an experiment to test approaches to building metamodels. For an object model we used EXHALT-CF (Davis et al, 2002), a model which treats the halt phase of a military operation. In its simplest version, the halt phase is a mere race. An attacking force (Red) is advancing on an objective while the defenders (Blue) interdict its armored vehicles with long-range fires. Red will halt when he reaches his objective (a Red win) or when Blue has killed a specified number of vehicles (a Blue win), whichever comes first. EXHALT-CF, however, adds many embellishments relevant to current strategic concerns about real-world military operations, especially in the Persian Gulf.

First, the model must represent Blue deployments. Some number of shooters may be stationed in theater in peacetime. Depending on strategic warning, diplomatic relations, Red's deceptiveness, and Red's ability to threaten bases in theater (e.g., with weapons of mass destruction), Blue may or may not be able to augment this number before Red begins his advance. Once Red's advance begins, Blue will deploy more shooters into the theater, up to a theater capacity, which reflects logistical shortcomings.

The effectiveness of Blue shooters is measured by kills per shooter-day. Early in the campaign, Blue may be unable or unwilling to attack the Red column because of Red air defenses. After a period of air-defense suppression, Blue's attacks will start. Even then, however, sortie rates may be reduced because of a continued threat of attack with mass-destruction weapons, which would force Blue personnel to work in protective gear or would force Blue to operate from more distant, and more poorly prepared bases.

The weapons and strategy Blue selects will also influence Blue shooter effectiveness. Blue may select an area weapon, capable of killing several Red armored vehicles per shot. To counter this, Red may space his vehicles more widely. Or Blue may select a point weapon, which kills no more than one vehicle per shot, and is unaffected by Red's vehicle spacing. Also, Blue will likely have limited supplies of his best weapons, and revert to lesser weapons when his best are exhausted. Blue may attack the entire Red column in depth (the "In Depth" strategy) or focus his attack on the leading edge (the "Leading Edge" strategy). If Blue does the latter, his attack may slow Red, but each sortie may be less effective due to deconfliction problems.

We set out to build a metamodel that would estimate Red's halt distance as a simple function of EXHALT-CF's inputs. We generated a dataset of a thousand cases. First we used a purely statistical approach, blindly regressing the outcome (the Red halt distance) against everything in sight. Even adding cross product terms as independent variables did little to improve matters. Of course, had we been fitting a response surface in a small neighborhood of a base case, this approach might have worked well. Assuming the model is smooth (twice differentiable when seen as a function), Taylor's Approximation from elementary calculus guarantees success if the

neighborhood is small enough. But we were looking for a metamodel that performed well over a much larger region of the input parameters.

We constructed a series of four metamodels. For the first we (almost) blindly regressed Red’s penetration distance on the 25 input variables that we had varied to create the analysis dataset. For the second, we introduced some composite variables that actually appear as intermediate variables in the object model. For example, three of the variables in the analysis dataset are the number of Red divisions, the armored vehicles per division, and the fraction of Red vehicles that must be killed to effect a halt. These variables influence Red’s penetration distance exclusively through their product, which we duly included in the second metamodel as a replacement for the original three.

We based the third and fourth metamodels on an explicit story, hardly more complicated than this. Red will start moving at a time T_{delay} and thereafter move at a velocity V until he reaches his objecting or Blue has killed enough vehicles, whichever comes first. Blue will begin shooting at a time T_s and thereafter kill vehicles at a rate proportional to the number of shooters and the effectiveness of each shooter. The time to kill enough vehicles is the ratio of required kills to the kills per day. This story is very similar to the one described in Appendix B, Equation (B.1). In metamodel 3 we used a simple estimate for the average number of shooters in the theater. In metamodel 4 we used a more complicated expression for average shooters that accounted for the fact that more shooters would deploy if the campaign were longer. As shown in Table F.1, the more “theory” we added, the simpler the metamodels became and the better they fit the data.

Table F.1: A Dose of Theory Improves the Metamodel

Metamodel #	# Calibration Coefficients	# Aggregate Variables	RMS Error (km)
1	15	14	140
2	11	10	84
3	9	5	30
4	9	5	8

REFERENCES

- Abrahamse, Allan F., James H. Bigelow, R.J. Gladstone, Bruce F. Goeller, Thomas F. Kirkwood, Robert L. Petrueschell (1977), *Protecting an Estuary from Floods - A Policy Analysis for the Oosterschelde. Vol. II, Assessment of Security from Flooding*, RAND, R-2121/2-NETH
- Bigelow, James H., Joseph G. Bolten, James C. DeHaven (1977), *Protecting an Estuary from Floods - A Policy Analysis for the Oosterschelde. Vol. IV, Assessment of Algae Blooms, A Potential Ecological Disturbance*, RAND, R-2121/4-NETH
- Bigelow, James H., F. J. Los, Nico M. de Rooij, J. G. C. Smits (1982), *Policy Analysis of Water Management for the Netherlands: Vol. VI, Design of Eutrophication Control Strategies*, RAND, N-1500/5-NETH
- Bigelow, James H. (1984), *Managing Recoverable Aircraft Components in the PPB and Related Processes: Executive Summary*, RAND, R-3093-MIL
- Bigelow, James H., Paul K. Davis, Jimmie McEver (2001), *Case History of Using Entity-Level Simulation as Imperfect Data for Informing and Calibrating Simpler Analytic Models for Interdiction*, in Proceedings of the 2000 Winter Simulation Conference, Joines JA, Barton RR, Kang K, and Fishwick PA, eds., pp. 316-325.
- Catlett, Louis, Richard Stanton, Orhan Yildiz (1979), *Controlling the Oosterschelde Storm-Surge Barrier - A Policy Analysis of Alternative Strategies. Vol. IV, Basin Response to North Sea Water Levels: The BARCON SIMPLIC Model*, RAND, R-2444/4-NETH
- Crawford, Gordon B. (1988), *Variability in the Demands for Aircraft Spare Parts: Its Magnitude and Implications*, RAND, R-3318-AF.
- Davis, Paul K. (2002), *Analytic Architecture for Capabilities-Based Planning, Mission-System Analysis, and Transformation*, RAND, MR-1513-OSD.
- Davis, Paul K. (1995), *Aggregation, Disaggregation, and the 3:1 Rule in Ground Combat*, RAND, MR-638-AF/A/OSD.
- Davis, Paul K. (1994) (ed.), *New Challenges in Defense Planning: Rethinking How Much Is Enough*, RAND, Santa Monica, CA.
- Davis, Paul K., James H. Bigelow (1998), *Experiments in Multiresolution Modeling (MRM)*, RAND, MR-1004-DARPA
- Davis, Paul K., David C. Gompert, Richard Hillestad, Stuart Johnson (1998), *Transforming the Force: Suggestions for DoD Strategy*, RAND, IP-155.
- Davis, Paul K., James H. Bigelow, Jimmie McEver (1991), *Exploratory Analysis and a Case History of Multiresolution, Multiperspective Modeling*, RAND, RP-925.
- Davis, Paul K., James H. Bigelow, Jimmie McEver (1999), *Analytical Methods for Studies and Experiments on "Transforming the Force,"* RAND, DB-278-OSD
- Davis, Paul K., James H. Bigelow, Jimmie McEver (2000a), *Informing and Calibrating a Multiresolution Exploratory Analysis Model with High Resolution Simulation: The Interdiction Problem as a Case History*, in Proceedings of the 2000 Winter Simulation Conference, pp. 316-325

- Davis, Paul K., James H. Bigelow, Jimmie McEver (2000b), *Effects of Terrain, Maneuver Tactics, and C4ISR on the Effectiveness of Long-Range Precision Fires: A Stochastic Multiresolution Model (PEM) Calibrated to High-Resolution Simulation*, RAND, MR-1138-OSD.
- Davis, Paul K., Jimmie McEver, Barry Wilson (2002), *Measuring Interdiction Capabilities in the Presence of Anti-Access Strategies: Exploratory Analysis to Inform Adaptive Strategy for the Persian Gulf*, RAND, MR-1471-AF.
- Davis, Paul K., James H. Bigelow (forthcoming), *Motivated Metamodels: Synthesis of Cause-Effect Reasoning and Statistical Metamodeling*, RAND, MR-1570-AF
- Defense Science Board (1998), *Defense Science Board 1998 Summer Study Task Force on Joint Operations Superiority in the 21st Century: Integrating Capabilities Underwriting Joint Vision 2010 and Beyond*, Office of the Under Secretary of Defense for Acquisition and Technology.
- Dewar, James A., James J. Gillogly, Mario L. Juncosa (1991), *Non-Monotonicity, Chaos, and Combat Models*, RAND, R-3995-RC
- Gritton, Eugene C., Paul K. Davis, Randall Steeb, and John Matsumura (2000), *Ground Forces for a Rapidly Employable Joint Task Force*, RAND, MR-1152-OSD.
- Hodges James S. (1980), *Onward Through the Fog: Uncertainty and Management Adaptation in Systems Analysis and Design*, RAND, R-3760-AF/A/OSD
- Larsen, Ralph I., C. E. Zimmer (1965), *Calculating Air Quality and Its Control*, APCA Journal 15:12;565-572
- Law, Averill M., W. David Kelton (1991), *Simulation Modeling and Analysis*, 2nd ed., McGraw-Hill
- Los, F. J. (1991), *Mathematical Simulation of Algae Blooms by the Model BLOOM II, Version 2*. Delft Hydraulics Laboratory
- Matsumura, John, Randall Steeb, Tom Herbert, M. Lees, Scott Eisenhard, Angela Stich (1997), *Analytic Support to the Defense Science Board: Tactics and Technology for 21st Century Military Superiority*, RAND, DB-198-A
- Matsumura, John, Randall Steeb, Ernest Isensee, Tom Herbert, Scott Eisenhard, John Gordon IV (1999), *Joint Operations Superiority in the 21st Century: Analytic Support to the 1998 Defense Science Board*, RAND, DB-260-A/OSD
- McEver, Jimmie, Paul K. Davis, James H. Bigelow (2000), *EXHALT: An Interdiction Model for Exploring Halt Capabilities on a Large Scenario Space*, RAND, MR-1137-OSD.
- National Research Council (1997), *Modeling and Simulation*, Vol. 9 of Technology for the United States Navy and Marine Corps, 2000-2035, National Academy Press, Washington, D.C.
- Rosenau, James N. (1997), *Many Damn Things Simultaneously: Complexity Theory and World Affairs*, in *Complexity, Global Politics, and National Security*, David S. Alberts and Thomas J. Czerwinski, eds, National Defense University, Washington, D. C.
- Saltelli, Andrea, Karen Chan, E. Marian Scott, eds. (2000), *Sensitivity Analysis*, John Wiley & Sons, Ltd.
- Simon, Herbert A (1982a), *Models of Bounded Rationality: Vol. 1. Economic Analysis and Public Policy*, MIT Press, Cambridge Mass

Simon, Herbert A (1982b), *Models of Bounded Rationality: Vol. 2. Behavioral Economics and Business Organization*, MIT Press, Cambridge Mass

Smith, Giles K., Gordon F. Acker, James H. Bigelow, David J. Dreyfuss, S. V. La Forge, Richard Y. Pei, Susan A. Resetar, Robert L. Petruschell (1988), *Design, Performance, and Cost of lternative LHX Configurations*, RAND, R-3625-A