Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

SIMULATION IN CONTEXT: USING DATA FARMING FOR DECISION SUPPORT

Philip Barry
Matthew Koehler

The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102, U.S.A.

ABSTRACT

Data Farming leverages high performance computing to run simple models many times. This process allows for the exploration of massive parameter spaces relatively quickly. This paper explores a methodology to use Data Farming as a decision support tool. Data Farming can be a highly effective in this role because it allows one to present to a decision-maker not only what may be the most likely outcome but what are possible outcomes, especially outliers that might have far reaching impact. The terrorist attacks of September 2001 are a good example of an outlier with very high impact. A case study is presented using a simple terrorist attack simulation and decision-maker utility model.

1 INTRODUCTION

Project Albert (Brandstein and Horne 1998, Fry and Forsyth 2002) is driven by the idea that simple models run millions of times can provide insights into outliers, nonlinearities, intangibles, and adaptation in ourselves and others that might otherwise slip past our decision-makers. For this to be done properly it needs to be question driven, therefore, the decision-maker or other subject matter expert should be involved. This paper will discuss the methodology of Project Albert, Data Farming, in the context of decision making in the face of very rare events.

Data Farming is a broad term that encompasses many separate processes that form a cohesive whole. The heart of Data Farming is a complex multivariate question that does not lend itself to a closed form analytic solution. Once a question is formulated, its essential aspects are captured in one or more models by a collaborative team (made up of modelers and subject matter experts). The utility of Data Farming lies in the ability to create the models quickly, run them many times, and easily analyze and interpret the results. This approach requires relatively abstract models to be tractable. The use of abstract models increases the importance of the subject matter expert because they must be able to distill the situation, as defined by the question, to its essence. Once this "distillation" is created it is put into the Data Farming environment and run many times.

There are two aspects to how the model is run within the Data Farming environment. First, large parameter space is explored in either a full factorial experimental design or a sampling approach such as a Near-Orthogonal Latin Hypercube (Lucas et al. 2002). Second, due to the sensitivity of the models to slight perturbations in the initial layout or the random number stream during the run, they are run many times with the same parameter combinations but with different random seeds. Alternatively, the parameter combinations may not be set initially, but may be created using one of a number of different evolutionary or natural algorithms to find near-optimal parameter combinations based upon a user defined fitness function.

Once the runs of the model are completed the output data is analyzed to determine if the model was created correctly and if it adequately captured the essence of the question. If there is some problem either with how the model was created or how it captured the question the model can be changed very quickly and the experiments run again. Once the modeler and subject matter expert are satisfied that the model represents the question at hand, the analysis enters the Operational Synthesis (Horne 2001) cycle. The insights generated from the Data Farming can be used to inform other aspects of the analytic processes, be they legacy models, traditional decision support, or even a war game.

"Insights" in the previous sentence is not used gently in this context, these models are abstractions and no one would argue that they can answer a question fully in and of themselves. Data Farming should be part of a larger analytic process. The remainder of this paper introduces Data Farming and then discusses how to use the results in context with respect to the decision maker. An utility approach is used to integrate the Data Farming methodology into a decision support environment.

2 EXPLORING VAST PARAMETER SPACES

As discussed above, Data Farming is frequently used for problems without closed form solutions. Therefore, to date, all of the models within the Data Farming environment are agent-based models (ABM) designed to be very

sensitive to initial conditions. However, there is nothing inherent about the Data Farming environment that requires the use of ABM's. ABM's have been utilized thus far because they are well suited to exploring the aforementioned intangibles and nonlinearities that plague many or our decision-maker's questions. Specifically, MANA, Pythagoras, Socrates, PAX, and NetLogo are incorporated into the Data Farming environment. MANA, Pythagoras, and Socrates are designed to be, but not limited to, combat simulations. PAX is used to model peacekeeping and peace support operations. Finally, NetLogo is an open-ended ABM environment in which the user codes the model in an endogenous scripting language.

Furthermore, even these simple models can have very large input parameter space. Moreover, many of the parameter combinations in this space may be significant to the decision-maker. Worse yet, the importance of the parameter combinations may be difficult or even impossible to accurately state *a priori*. This necessitates a methodology such as Data Farming, one which allows for searching through a great number of these parameter combinations to find regions of interest to explore more fully on a second, more focused, iteration.

These models combined with the Data Farming environment creates a unique capability to explore low base-rate phenomena (LBP). LBP can be of particular interest to a decision-maker because they may represent an extremely rare but potentially catastrophic outcome of a particular system. The crisis at Three Mile Island and the September 2001 terrorist attacks each represent extremely rare events with far reaching consequences. Unfortunately, these events can be extremely difficult to study because they represent not only a specific set of parameter combinations but also the right initial conditions. The case study examined in this paper deals with LBP.

Even assuming that we can model the initial conditions and use Data Farming techniques to generate LBP, there still is the open question as to how to identify the phenomena of interest and how to best analyze it. There are a number of established and emerging tools well suited to this task. There are two visualization tools designed around the Data Farming environment that are geared to examining high dimensional data with such things as landscape plots, parallel coordinate plots, whisker diagrams, violin plots, and many types of scatter plots (Meyer and Johnson 2001). Also, nonparametric statistics are well suited to this data, as it may be highly skewed and far from normal, and you may have little understanding of the underlying distribution from which the data came. Furthermore, Bayesian Networks and other forms of statistical learning are proving quite useful. Data Mining is also providing many techniques relevant to the analysis of Data Farming results. Cluster Analysis and Decision Trees can provide useful results to a decision-maker. Perhaps of most immediate utility is the technique of generating Classification Rules. This technique can provide a decision-maker with an outcome and what parameter value(s) may lead to that outcome, including associated distributions which can be extremely helpful in understanding not only what is most likely but also what is possible.

3 DATA FARMING IN CONTEXT

The objective for using Data Farming techniques is to provide insight into decisions. The fact that a low base-rate occurrence may happen only has value with respect to the potential consequences in the larger scheme of things. For example, a system failure is only as important as what functionality fails in a given context. A safety critical function in an airplane has potentially catastrophic consequences; however, a failure in the aircraft's entertainment system, while inconvenient, is not nearly as serious. Appropriately, the amount of resources potentially allocated to identify and track down safety critical functions would be significantly greater than the amount allocated to find issues with the entertainment system.

3.1 A Single Attribute Framework

In the utility analysis literature one finds a framework that readily lends itself to examine the low base-rate problem. In essence, one can model the preferences of the decision-maker and use this to balance the costs and benefits for the decision. Perhaps most importantly, the utility analysis framework exposes the belief structure of the decision-maker and allows for tuning during the analyses.

Consider the simplest case, a single attribute that one seeks to optimize. We can write this simple expected value equation as:

$$E(\mathbf{X}) = \sum_0^n \sum_0^j p_n x_j$$

where p is the probability of a given n event in N**,** where N is the space of possible options. X is the relevant impact of each j factor mapped to a single variable. For example, if X is the number of users that leave an internet service provider, x_1 could be the number of users in Virginia and x_2 could be the number of users in Maryland.

So, in this simple case, we could use Data Farming to determine the probabilities p_n for each n. Furthermore, we could also then use Data Farming to identify the appropriate distributions for x_j. By plugging the values into the equation, we then have a decision relevant context by which to evaluate the choices.

Continuing with the example, suppose two strategies are being evaluated for setting new rates for an internet service provider. After Data Farming, it is determined that it is very unlikely that anyone will leave given the proposed changes to the rates. However, there is a 0.1 chance that Option 1 will result in 7500 subscribers leaving from Virginia and 2500 subscribers being added in Maryland. Option 2 is shown to

have only a 0.08 chance of going wrong. Data Farming has indicated that when Option 2 does go wrong 5000 subscribers leave Virginia but only 1000 subscribers are added in Maryland. Plugging into the equation yields $e(x_1) = 0.1$ (-7500 + 2500) or 500 persons expected to leave. For option 2, this results in $e(x_2) = 0.8$ (-5000 + 1000) or 320 persons expected to leave. If the goal of the decision is to minimize the number of losses, then strategy 2 is the best option.

3.2 Modeling the Decision Maker

The preceding discussion assumed a single criteria that was linearly important to the decision maker, or at least any non-linear preferences were not specified. However, frequently the decision maker has non-linear preferences. Take for example, the terrorist scenario discussed in detail in section four. If one terrorist makes it through, it can be viewed as a significant loss. If two terrorists get through it is certainly worse, but not necessarily twice as bad. This is important because Data Farming will provide a wide spectrum of possible events and their associated probability distributions. However, we need to model at some level the value structure of the decision maker to provide a context for the insights from the simulation.

The risk literature provides a good starting point for this. At an abstract level, we can view a decision maker as risk averse, risk neutral or risk seeking. Informally, a risk averse individual prefers a known outcome to a lottery or an event with chance, a risk neutral individual is indifferent to the sure bet and the lottery, and finally, the risk seeking individual prefers the lottery to the sure thing.

We can write compact mathematical descriptions of each of these individuals. We may view the decision maker as an individual assessing the goodness or utility u of a given decision x in the space of possible decisions X. A risk averse individual can be modeled with the expected utility of a decision as follows,

$$u(x) = a + b(-e^{-cx}),$$

where a and b are scaling coefficients and c is a parameter that indicates the risk averseness of the individual. Similarly, we can model the risk neutral individual by using

$$u(x) = a + b(cx)$$

and the risk seeking individual as

$$u(x) = a + b(e^{cx}).$$

We can obtain the expected value of x by Data Farming, then use these equations to understand their meaning in context.

Consider our example discussed in detail later in the paper which looks at a squad of ten terrorists. We can use the risk averse model to represent a decision maker who believes that unless almost all of the terrorists are caught, there is limited utility in taking a given action. This can be readily represented by

$$u(x) = e^{(x-10)},$$

where u(x) represents the utility of catching x terrorists. The risk neutral models can be used to represent a situation where each additional terrorist that is caught represents a linearly increasing utility. This can easily be represented as

$$u(x) = \frac{x}{10}.$$

Lastly, the risk seeking model can be used to represent the situation where there is an initial large utility in catching the first few terrorist but the net benefit decreases each time more terrorists are caught. Represented as

$$u(x) = 1 - e^{-x},$$

this could represent a situation where no additional information is obtained even though additional terrorists are caught. Figure 1 illustrates these preference curves.

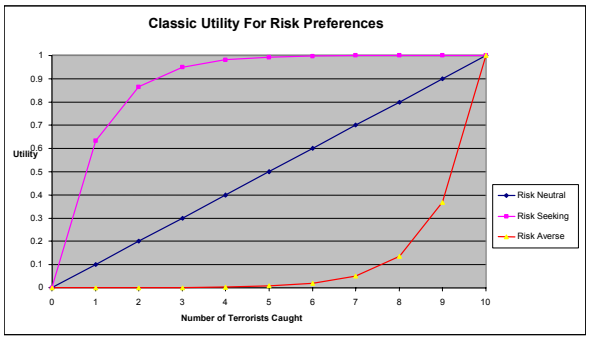

Figure 1: Preference Curves for Single Attribute

This domain understanding is key for using Data Farming. The model of decision maker utility allow for the contextualization of the insights gained through Data Farming. Basic utility analysis provides the first step in understanding where this plays in. The next step is to use multi-attribute utility analysis to balance competing goals, such as minimizing costs and preventing the maximum number of hostile acts.

3.3 Multi-Attribute Approaches

Multi-attribute utility analysis is a well founded approach to balancing several objectives. In any real decision it is most likely that it is desired to derive the "best" solution balanced against a number of factors. In this section, we will illustrate some of these concepts within the Data

Farming framework. The interested reader is referred to Keeney (1992) for a more detailed treatment.

The basic idea for multi-attribute utility analysis is to break down the assessment of the utility of a given decision into component pieces, assess these pieces and integrate them. The preference models of the decision maker must be taken into account in a quantified manner, such as described in the previous sections. The equations make good use of independence conditions of the variables. There are four major types of independence assumptions: preferential, weak difference, additive and utility. Preferential independence states that the relative preferences of two sets of variables is independent of other variables. Weak difference independence states that the preference between two levels for a given variable does not depend upon other variables. Additive independences describes the situation where the preference for a given outcome of two variables is independent of the combination of the two variables but relies only on maximizing (or minimizing) each variable individually. Utility independence states that the preference of an outcome involving one variable is independent of the levels of other variables.

Each of these variables has well described canonical forms, but for brevity we will only examine utility independence. From (Keeney 1972) the utility function for a number of variables which are utility independent is:

$$u(x_1, \ldots\ldots, x_N) = \sum_{i=1}^{N} k_i u_i(x_i) + \sum_{i=1}^{N} \sum_{j>1}^{N} k_{ij} u_i(x_i) u_j(x_j) + \sum_{i=1}^{N} \sum_{j>i}^{N} \sum_{h>j}^{N} k_{ijh} u_i(x_i) u_j(x_j)\ u_h(x_h) + \ldots.. + k_{1\ldots N} u_i(x_i) u_j(x_j) \ldots.\ u_N(x_N)$$

where the k's are the weighting factors less than 1 that describe the relative importance of the each variable x.

For the two variable case, this can be written as:

$$u(x_1, x_2) = k_1 u_1(x_1) + k_2 u_2(x_2) + (1 - k_1 - k_2) u_1(x_1) u_2(x_2)$$

which simplifies to

$$u(x_1, x_2) = k_1 u_1(x_1) + k_2 u_2(x_2) \text{ if } k_1 + k_2 = 1$$

4 EXAMPLE

The example used here is of a very abstract, simple terrorist attack. In the model a group of ten red agents start at the right edge of the battle space and move to a single target at the left edge of the battle space (see Figure 2). In between where they start and where they are trying to go is a group of fifty blue agents that try to stop them, as well as a large number of small obstacles that block movement, sensors, and weapons. There were three different defensive postures used by the blue forces. These included: evenly dispersed throughout the battlespace, tightly clustered around the target, and finally, evenly dispersed throughout the battlespace but with increased capabilities for tracking and neutralizing terrorist if they are detected. Using the modeling methodology described above, we identified two factors in

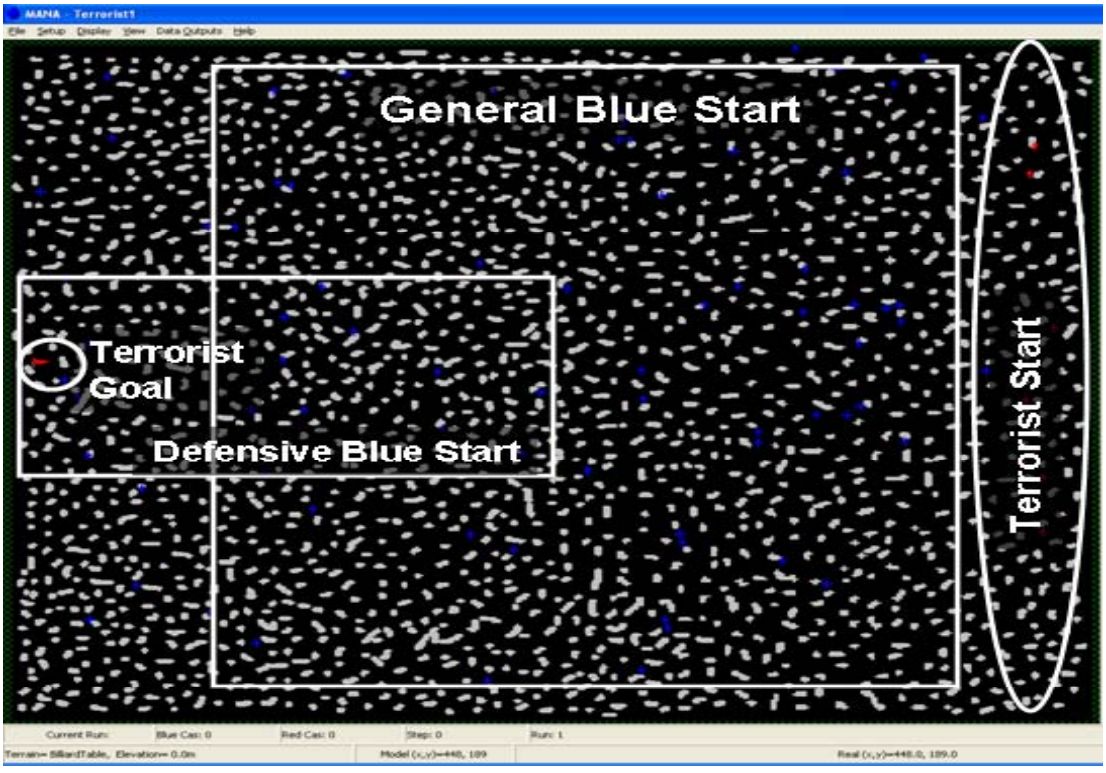

Figure 2: The Example Terrorist Scenario

which we are interested. The first is the number of blue casualties, modeled as

$$u_1(x_1) = e^{\frac{-x}{10}},$$

where x_1 represents the number of blue casualties with the utility decreasing exponentially with increasing blue casualties. We similarly model the utility for red casualties as before, with increasing utility for more red casualties as

$$e^{(-1(10-x_2))},$$

with x_2 as the number of red casualties. For this decision we have decided that eliminating terrorists is four times more important than preserving blue forces, so $k_2 = 0.8$ and $k_1 = 0.2$. We examine this in the three scenarios that have been identified as possible courses of action. As described above: Option 1, is the base case where blue is evenly distributed throughout the battlespace and is represented by X_1. In Option 2 (X_2) a trigger is set such that when blue defensive forces detect red forces the blue forces are better able to track, follow, and neutralize red forces. In Option 3 (X_3) blue forces are in a more tightly concentrated defensive posture around the target.

The simulation was run twenty-five thousand times (varying the random seed) for each of eight red force parameter combinations (for a total of one million model runs), varying such things as how aggressively red tries to avoid blue, how large of a numeric advantage red requires before it will advance on the target, and others. For X_2 and X_3, the utility is clearly better than X_1, with X_2 being the better choice (see Table 1). Several observations are worth noting here. X_1 is very unstable as is evidenced by the large standard deviations for the Blue and Red dead. However, since the net utility is so much less (0.15), it is clearly not worth pursuing. X_2 , while dominating, only has a net utility of 0.589. While this is the best score out of the three scenarios, there is clearly room for improvement which may warrant additional research into Blue tactics.

Table 1: Model Output and Utility Equations

	Expected Blue Dead	Std Dev Blue Dead	Expected Red Dead	Std Dev Red Dead	$k_1 = 0.2\ u_1(X) * k_1$	$k_2 = 0.8\ u_2(x) * k_2$	Net Utility
X_1	2.80	3.58	3.99	4.15	0.15	0.002	0.153
X_2	1.02	1.04	9.33	1.03	0.18	0.409	0.589
X_3	0.65	.799	9.11	1.15	0.19	0.328	0.516
Outlier for X_2	7	n/a	7	n/a	0.09	0.039	0.139

Perhaps more interesting is the outlier discovered for scenario two (X_2). This outlier only occurred twice in the total number of runs. However, in this outlier there are seven red and seven blue casualties. By the value model described for the utility equation, this represents a utility of 0.139 which is noticeably less than the base case. While unlikely, this scenario is sufficiently poor to warrant additional investigation into the causes.

5 CURRENT RESEARCH DIRECTIONS

There is a great deal of ongoing research in the Data Farming community. The 1st Marine Expeditionary Force is utilizing Data Farming to explore how to protect convoys from attack by both ambushes and improvised explosive devises using a variety of sensor and weapons platforms. Furthermore, they are using this methodology to explore how civilian populations are affected by military action within a stability and support operation. The US Marine Corps Combat Development Command is currently using Data Farming to understand the full implications of more fully distributed operations. Finally, the Commandant of the US Marine Corps had tapped into Data Farming to provide another perspective on the ship that will replace the USMC's current aircraft carrier, and on the proper mix of smart and "dumb" munitions necessary to carry out a successful ship to objective maneuver, including the use of loitering munitions. The US Army TRAC-Monterey has utilized this methodology to focus the use of large scale legacy model.

There are many international groups that use the Data Farming methodology, as well. The German military is using this methodology to model Peacekeeping operations. NATO commenced an effort including the exploration of how to exploit the utility of network-centric warfare doctrine using the tools and techniques of Data Farming. The Swedish Defense University is looking at command and control concepts with this methodology, also. The Australian and Singaporean militaries are exploring many issues including sensor use. New Zealand has used Data Farming for many studies including tactics used in Peacekeeping missions in East Timor.

We are actively engaged in transitioning this technology to the field. Utility analysis has shown promise in framing this work so the results from Data Farming can be related to the decision makers value models. This work will be integrated into the aforementioned topics with the aim of transitioning Data Farming from an investigative technique to a more formal decision support tool.

6 REFERENCES

Brandstein, Alfred, and Gary Horne, Data Farming: A Meta-Technique for Research in the 21st Centery. In Maneuver Warfare Science 1998, eds. F. G. Hoffman and Gary Horne. USMC, Washington, DC, 1998.

Fry, Ashley and Adam Forsyth, The Australian Army and Project Albert: Pursuing the Leading Edge of Military Thinking and Technological Development. In Maneuver Warfare Science 2002, eds. Gary Horne and Sarah Johnson. USMC Project Albert, Quantico, VA, 2002.

Horne, Gary E., Beyond Point Estimates: Operational Synthesis and Data Farming. In Maneuver Warfare Science 2001, eds. Gary Horne and Mary Leonardi. USMC Combat Development Command, Quantico, VA, 2001.

Keeney, R.L., Utility Functions for Multiattributed Consequences. Management Science, 18, 276-287, 1972.

Keeney, R.L., Value Focused Thinking, Harvard University Press, Cambridge, MA, 1992.

Lucas, Tom, Susan M. Sanchez, Maj. Lloyd Brown, and Maj. William Vinyard, Better Designs for High-Dimensional Explorations of Distillations. In Maneuver Warfare Science 2002, eds. Gary Horne and Sarah Johnson. USMC Project Albert, Quantico, VA, 2002.

Meyer, Theodore and Sarah Johnson, Visualization for Data Farming: A Survey of Methods. In Maneuver Warfare Science 2001, eds. Gary Horne and Mary Leonardi. USMC Combat Development Command, Quantico, VA, 2001.

AUTHOR BIOGRAPHIES

MATTHEW KOEHLER is an Artificial Intelligence Engineer at The MITRE Corporation. He is currently an analyst with Project Albert working to transition the methodology and capabilities from an experimental sector to operational forces. He has graduate degrees from the George Washington University Law School and the Indiana University School of Public and Environmental Affairs, and an undergraduate degree in Anthropology from Kenyon College. His email address is <mkoehler@mitre.org>.

PHILIP BARRY, Ph.D. is a Senior Principal Engineer at the MITRE Corporation. He has been working in the modeling and simulation arena for a number of years, most recently with Project Albert. He has graduate degrees in Information Technology and Systems Engineering from George Mason University and an undergraduate degree in Aerospace Engineering from the University of Virginia. His email address is <pbarry@mitre.org>.