

Reliable Knowledge and Error in Simulation Models.

Mark Boyland
March 9, 2002

University of British Columbia
Vancouver, Canada

ABSTRACT

ATLAS and SIMFOR both simplify and abstract reality. Their inputs are not literally represented within the model, nor can their outputs be literally transferred by decision makers into management plans. Understanding the simplifications and abstractions in creating datasets and algorithms is crucial to understanding what types of knowledge can be gained from the outputs. What aspects of a spatial harvest schedule can suffer generalization to a statement of how the landscape will actually appear in 100 years? When is accuracy error (mistakes in measurement) less important than abstraction error (mistakes in understanding)? Both abstraction and accuracy error are present in all model outputs; the challenge for modellers is to recognize how reliable knowledge can be found in spite of error. Based on the relevant sections of “Decision-support systems: it’s the question not the model” (Bunnell and Boyland, 2002), this paper explores in simple terms a framework for understanding the different types of error within models.

This extension report was commissioned by the ATLAS/SIMFOR Project. Funding was provided from the Collaborative Research fund, a special FRBC fund, administered by the Science Council of British Columbia on behalf of FRBC.



INTRODUCTION

As Magritte pointed out, that is not a pipe. You cannot pick it up, light it, or smoke it. That is not a pipe, that is a painting of a pipe. It is a representation of what we understand a pipe to be, expressed through a new medium. Magritte has captured much of the essential “pipeness” in the painting though, and we can relate strongly to a real pipe through the pipe painting.

Models in forest management are similar to Magritte’s pipe. They are representations of what we understand a forest to be, expressed in a new medium – a model. The model is not the forest, but it can capture some of the essential “forestness” in the model, enough so that we can successfully relate to a real forest through a modelled forest.

Entrants to the modelling field often directly believe in models. They think the model is a literal translation of the forest, and that the outputs of the model are literal predictions. Quickly they learn that this cannot be the case, and a crossroads appears: either the model is literally wrong, and therefore useless, or the model needs to be interpreted, and therefore needs to be understood within its own terms. ATLAS (Nelson 1998) schedules harvests on a landscape, producing a harvest schedule. Is that schedule meant to be followed literally? If not, is there some way of extracting valuable information from it? Effective use of models stems from an understanding of the simplifications and abstractions that take place during model creation. Understanding what types of errors are introduced in this process, and how model outputs can be generalized is the key to gaining reliable knowledge from models.

The initial sections of this paper will explore the two major types of error in models. I will then discuss how these errors become embedded in model components, and give an example of model building using deer habitat relationships. The final sections of the paper will examine how models can be reliably interpreted. This paper is an extension note drawing in part from a talk given in Davos, Switzerland (Bunnell and Boyland 2002).

ERROR – ACCURACY & TRANSLATIONAL

Before learning about how to understand and interpret a model, some familiarity with ways to evaluate models is useful. The best way to do this is to concentrate of

where models can go wrong – errors. Mathematically, errors are quantified by many types of indices and error terms, from simple r^2 to complicated formulas. Here we are more interested in the nature of errors, rather than their enumeration. Conceptually, error is easily defined as a deviation from truth. Truth is somewhat more difficult to define, however the two definitions below conveniently divide error into two categories. These two categories of accuracy error and translation error create a simple framework for understanding the effect of error within model output.

Where there is some rigorously predefined system, truth can be determined by fiat. Error is then easily measured, though still managed only with difficulty if the system is complex. Suppose we define the Diameter at Breast Height (DBH) of a tree as the diameter of a tree, external to bark, taken at a point 1.3 meters above the highest ground position adjacent to the bole, not influenced by litter accumulation or butt swell. Then a measure following those specifications is a true measurement of DBH. The definition of DBH defines the truth of DBH. Errors are mistakes in measurement that are outside the truth of DBH, which we term *accuracy errors*. They are accidents of measurement, and artefacts of data resolution or statistical uncertainty. They are mistakes made within a structure, after the structural framework has been determined.

The error resident in the choice of structure, we term *translation errors*. Translation error occurs in the creation of the system that structures data or a relationship. If we wish to know the volume of a tree, some abstract notion of how a tree's volume is distributed must first be formed. Representative parameters of that distribution have to be determined before a technician approaches the tree with a tape measure. The physical fact of the tree is translated into an abstract notion of how to represent the tree. The choice of how to represent the tree frames our understanding (defines the truth) of the tree in a model. Translation error results from choosing an abstraction that either misrepresents or misses elements that are relevant to interpreting the outputs of a model. The model and real world do not correspond. Translation error can influence output, regardless of how well a tape measure is wielded, by directing measurement to things inappropriate to the purpose.

MODELS STRUCTURE

Models represent systems with two basic building blocks: data and algorithms. Data contain information on a system's initial state; in a forested setting this sometimes comprises of a GIS database, with current information on tree species, age, and site index. Algorithms capture processes that change data over time. These processes can be natural, or industrial. A simple algorithm is aging: a stand's age is increased by one for every year that passes in simulation. A more complex algorithm is the choice of which stands to harvest each year.

Data

Both accuracy and translational error are found in data and algorithms. Accuracy error in data is widely recognized, though less commonly explicitly incorporated into modelling than it should be. Translational error in data is only rarely discussed, though is

at least as important. Perhaps this is because we acquire data from such concrete things (trees, deer, shrubs), that we forget the data's context. But, moving from the trees themselves to a number describing the trees involves a theory about how the number can be related to the tree, and why it is relevant. For example, spatial data in a GIS locates attributes onto a map by circling them within polygons. A set of classification values governs where to draw polygon lines to determine distinct areas on a map comprised of gradients. A small change in attributes (fine resolution) produces many small polygons, while large changes in attributes (coarse resolution) lead to fewer, larger polygons. Many distinct maps can be derived for the same land base using different data structures. As a data input to an optimizing harvest scheduler, a map with more polygons will provide more opportunities for efficiencies than a map with larger, fewer polygons. The size of the polygons impacts the simulated level of timber harvest. Simply splitting a map with larger polygons into smaller units does not provide the same efficiencies, because the classification method (data structure) itself is responsible. The underlying methodology for creating the GIS database causes translational imprecision, influencing the projected harvest level.

Algorithms

Many algorithms are associated with terms such as 'hypothesis' and 'theory', that invite sceptical thought about their validity. Algorithms expressed statistically are continually evaluated, perhaps using r^2 and SE. When the algorithms are not statistical expressions, but simulation equations or other rules for change, they lack convenient measures of error and error is more difficult to assess. Greater confusion is created by transplanting algorithms from one system to another, or changing the conditions from those under which data were collected – two practices almost inevitable within models. Translational error in algorithms is easily shown in an example with ATLAS. ATLAS is a spatial harvest scheduler, and often uses the Oldest First algorithm, which chooses for harvest the oldest harvest unit within the planning area, subject to environmental and sustainability constraints. The algorithm is roughly consistent with a forest engineer's activities when converting old-growth forests: the oldest stands are converted into vigorously-growing, second growth stands, and no second-growth stands are harvested until the older, more profitable first-growth stands are harvested. Age represents a forest engineer's choice well (truth by correspondence), and will always select first growth over second-growth stands. The situation differs once a forest is comprised wholly of second- or third-growth. The choice between a 82 year-old stand and a 90 year-old stand is not likely to be made on age, but on a combination of species, volume, and/or distance to nearest road. Using the Oldest First algorithm now introduces a large component of translation error, by using wrong, or incomplete, indicators of choice between two harvestable stands.

Most importantly, data and algorithms lose translational accuracy when their use deviates from the conditions that were designed for. Models that are accurate with one interpretation, data set, or algorithm, can be very misleading in another situation. No model can be so isomorphic with the real world that its outputs correspond to the real world under all conditions. Each model, and model use, should be examined within the context of the questions put to it.

MODELLING A RELATIONSHIP

Creating a model is a process of abstraction and simplification. First, by simplifying a system's state and processes enough to be manageable, and then by abstracting them into numbers and equations. Elements of the complex, real world are translated (abstracted) into data, algorithms, and tautologies. Only parts of the complex, real world are chosen (simplification), and these are restricted in their range and properties to those that are measurable and relevant. We shift our focus from the very broad, real world to a very specific model. For this shift to be useful, the very specific must encompass those elements that successfully represent relevant features of the real world.



Figure 1a.



Figure 1b.

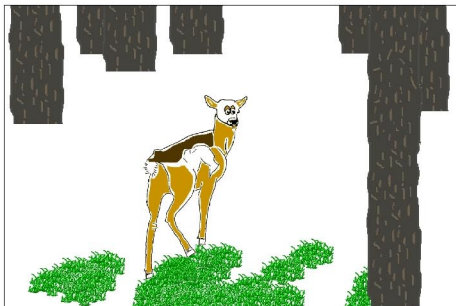


Figure 1c.

$$\text{Birth Rate} = (2 - 0.019)1/5$$

$$\text{HSI}(\text{forage}) = \max\{S1, S2, S3, S4\}$$

$$\text{Forage}(\text{winter}) = (\text{cover} * \text{snow})0.23k$$

Figure 1d.

Figure 1. Creating a model. Simplifying the real world (1a) to a few elements to model (1b). Simplifying further by reducing the elements to modellable characteristics. Abstracting those reduced elements into numerical equations (1d).

The sequence shown in figure 1 illustrates this process, moving from the real world to habitat algorithms. The deer and deer habitat in figure 1a is simplified to the major habitat elements in figure 1b. No model incorporates even a small fraction of the number of components in the real world because the model would be too complex to build. Then, these elements are simplified, because not every aspect of this reduced set of elements can be modelled. Perhaps only net productivity for shrubs and snow interception for the canopy. This is represented by the simplification of the elements into cartoons (figure 1c). These few elements, simplified into small aspects of their complexity, are then abstracted into numerical equations (figure 1d). From the concrete real world, to abstract

equations in three easy steps. Again, for this to be useful, these specific, simplified, abstracted equations must successfully represent the real world. Magritte's brush strokes captured some of the pipe, and our habitat equations are designed to capture some of the deer's needs for survival. But at each step in the process there are very real opportunities for translational error, by choosing the wrong elements to model, or by modelling them inappropriately.

Users of models must begin with the very specific model and generalize back to the very broad, real world. The same route of abstraction and simplification used in creating the model must be re-traversed to find reliable knowledge. The outputs of a model are the interactions of data and algorithms. When interpreting model output, the major issue is whether the abstractions that the data and algorithms represent are sufficiently reliable to suffer the treatment of generalization from model output to practice in the real world.

INTERPRETING OUTPUT

Interpreting model output requires an understanding of how error propagates. Using the two types of error from above – translational and accuracy error – figure 2 illustrates two separate processes of error transmission. Model output is the product of algorithms interacting with data, and error in output is the product of algorithm error and data error interacting.

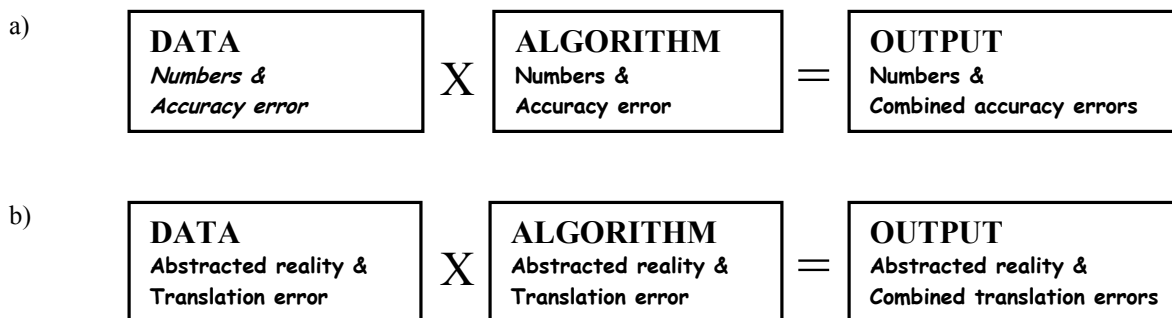


Figure 2. Two separate tracks of error propagation: accuracy error (1a) and translational error (1b). Both error types should be considered separately.

The rule of significant digits controls the precision of output – no digits are allowed beyond those within the least precise input. It is a useful rule because it links the input precision with the output precision: the inputs control the outputs. The concept is useful for error as well. We could view it as *significant output*, defined as that output whose interpretation is constrained within the translational validity of the data and algorithm. Reducing the complete output set to just the significant output can be done in two ways: by restricting or changing the interpretation of the output to maintain its significance, or by changing the output to be significant within the interpretation.

An example will best illustrate this process. The output of ATLAS can be very complex. It produces a spatially explicit harvest schedule, perhaps for over 300 years of management. These harvest units, while precisely located in space and time, are unlikely

to be viable management options. The model output is beyond the data and algorithm translation validity; it is not significant output yet. But by changing the output format, significant output can be found. The precise timing and spatial locations of the harvest schedule are almost certainly invalid, but some aspects of these predictions are more likely to be valid. Translating the output into summaries concerning the size class of the harvest units, the average rate of harvest, or the average growth rate of the stands will result in significant output. By ignoring the precise predictions of where and when to harvest, and restricting interpretation to the significant output, managers can gain useful insight into the generic spatial pattern of harvest, or the average annual harvest from the landbase. Another way of finding significant output is to change the interpretation of the spatial harvest. Instead of regarding it as an ironclad prediction of what will happen (which is invalid), consider it as one possible route among many million others (which is valid).

In all but the most simple models, the actual model outputs are not directly valid, but must be translated or reduced in some manner. Just as many digits are included through many calculations and then reduced to significant digits at the end to increase accuracy, models often use more complex data and algorithms than are translationally valid, and must be reduced in the output to find significant output.

CONCLUSION

Einstein commented that the process of science is like the process of using a cloak room ticket. A cloakroom ticket does not in any way resemble a coat, but it does correspond to your coat – using the ticket will produce your coat. This is the same with Magritte's pipe picture. The paint on the canvas is not a pipe, but has enough "pipeness" to be useful in contemplating pipes. That is the goal in modelling, to use numerical abstractions that capture enough elements of forests and habitat to be able to successfully relate the model to the real world.

Results from models can never be taken literally. A computer printout is most certainly neither forest nor deer, nor does it represent exactly what the forest is or will be. In most cases, what the output provides is not the reliable knowledge that the user desires. A sophisticated user should not be troubled by this. An important use of the model is the interpretation of output, separating the unreliable from the reliable and finding significant output. Discovering unreliable output through one interpretation does not discount the whole model. One can find unreliable elements in any model. Sophisticated users continue to examine ways of using the model that will allow for interpretations of output that do translate into reliable knowledge. Most models can inform to some degree. Some models employ abstractions that allow simplistic translation of output, and relatively straightforward interpretation. Others expose translational errors in all but a few interpretations. Learning about the translations is as important as learning about the resource – the translations are how we understand the resource.

LITERATURE CITED

Bunnell, F.L. and Boyland, M.N. 2002. Decision-support systems: it's the question not the model. *J. Nature Cons.* In Press.

Nelson, J. 1998. Forest Planning Studio (fps) – ATLAS Program. University of British Columbia, Vancouver Canada. Or -- <http://www.forestry.ubc.ca/atlas-simfor/>