Agent-Based Model (ABM) Validation Considerations

Philip Cooley, Eric Solano
RTI International
Research Triangle Park, NC, USA
e-mail: solano@rti.org

***Abstract*—This paper describes the use of validation methods in model building. We address issues associated with the increasing complexity of models that is in part a response to the growing popularity of Agent-Based Models (ABM), commonly used to study cognitive, natural, and social phenomena. The first section of this manuscript discusses model categories and attributes. The second section discusses the stages of validating a simulation model: verification, validity, and sensitivity analysis. The third section presents specific validation approaches, with an emphasis on six specific tests that are described in detail. The final section summarizes the goals of model validation and modeling.**

Keywords-Agent-Based Models, Validation, Verification, Infectious Disease Models.

I. INTRODUCTION

A number of global events point out the need for effective modeling. These include the H1N1 pandemic of 2009 and most recently, the Chilean earthquake tragedy, in which observers used modeling to issue tsunami warnings to Hawaii. The tsunami warnings overestimated the effect of the waves that would ultimately reach Hawaii, and "scientists will pore over reams of data" [1] as they work to understand what happened. However, some scientists say that "there should be a rigorous examination of long-standing assumptions within computer-generated models that are used to estimate the strength and impact of tsunamis," and that the "main problem right now is that we have unsubstantiated assumptions built into our warning system and we really have to check those [1]."

Due to significant reductions in the cost of computational resources and the increasing power of those resources, the nature and type of computer models used in a number of areas including disease transmission processes are changing. In particular, Agent-Based Models (ABM) are a relatively new technology growing in use. One reason is that ABM are an important method for representing and describing interacting heterogeneous agents. Recently, they have been applied to H1N1 infectious disease applications [2-7]. The heterogeneous property of agents enables ABM to describe more sophisticated and complex environments. Many researchers believe that human systems are complex processes that are poorly described by existing/alternative equation-based models (EBM) and it is easier to incorporate existing knowledge about human interactions and decisions into an ABM than into a model described by analytical equations [8]. The downside of this enhanced flexibility is that validating ABM may be more complicated because the processes they describe are more complicated; consequently, rigor is more difficult to achieve because of the complex environment.

A. Validation Definitions

Various definitions of validation appear in the literature. Schlesinger et al. [9] define validation as "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model." Midgley et al. [10] define validation as demonstrating that the "correct" equations have been solved by referencing an external and independent test. Macal [11] defines validation as the process of determining the extent to which a model or simulation accurately represents the "real" world from the perspective of its intended use. The final definition of validity presented here is from Ziegler [12], who distinguishes three types of validity:

- replicative validity—the model matches externally available data that has been generated by the modeled system (retrodiction).
- predictive validity—the model matches data that can be acquired from the modeled system, and
- structural validity—the model reflects observed behavior and matches the process inherent to the process to produce the behavior.

B. Model Characteristics

The type of model used to describe the phenomena of interest depends on the nature of the phenomena, the available supporting information about the phenomena, and the purpose of the model. A major issue that affects the type and quality of the validation method that can be applied is the degree of heterogeneity required to describe model elements. In many cases, the level of detail that is incorporated into the model architecture is dictated by the model's purpose. For example, if intervention strategies to prevent disease spread depend on individual agent characteristics, those characteristics have to be included in

 ISBN: 978-1-61208-169-4

the description of the agents. A review of the important categories of models and their characteristics is presented below. These categories are not mutually exclusive.

1) Agent-Based Models (ABM)

ABM have been used to describe phenomena such as social systems and immune systems, which are distributed collections of interacting entities (agents) that function without a leader. Simple agents interact locally according to simple rules of behavior, responding in appropriate ways to environmental cues and not necessarily striving to achieve an overall goal. An ABM consists of a set of agents that encapsulate the behaviors of the individuals that make up the system, and model execution consists of emulating these behaviors [13].

2) Equation-based Models (EBM)

EBM describe the modeled phenomena using a set of equations that interconnect the behavior of individuals or groups of individuals to the environment they inhabit. Manipulating the model's interconnections allows assessing control scenarios through evaluation of the equations. Historically, an important category of EBM is system dynamics, an approach based on describing simulation processes using ordinary differential equations (ODE) [15].

3) Social Network Models

The structure and dynamics of social networks are critically important to many social phenomena. There are a number of important questions in social networks research, but a lack of data does not allow them to be answered. For example, one of these questions is how social networks change over time.

Social network models are built around two basic entities of a directed graph: the node and the edge. Networks are a form of relational data and arise in many fields, and graphs are a natural method for representing the structure of these relationships. In these applications, nodes usually represent people or agents, and edges represent a specified relationship between them. This framework has many applications, such as assessing the influence of the structure of social networks on the spread of epidemics, assessing the interconnectedness of the World Wide Web, and examining long-distance telephone calling patterns.

4) Deterministic Models

A deterministic model is a mathematical model that employs parameters and variables that are not subject to random fluctuations. Therefore, the system is at any time entirely defined by the initial conditions, in that the assumptions and equations the user selects "determine" the results. The only way the outputs change is if the user changes an assumption (or an equation).

5) Stochastic Models

In many real-life situations, observations are influenced by random effects throughout an entire interval of time or sequence of times. A stochastic model includes elements of randomness that can be introduced at one or many points of the model. Thus, every time the model is applied, a different result is produced even if the parameters and logic are unaltered. Running the model many times provides a measure of the variability in the process that can be captured by the model. In many cases, stochastic models are used to simulate deterministic systems that include smaller-scale phenomena that cannot be accurately observed. The stochastic nature of these types of models is caused by at least three sources: noise in the parameter realization; the representation of a truly random process, and/or a deterministic process that is measured with imprecise tools. The last scenario, though not truly random, produces random-type behavior. In complex systems such as hybrid ABM/EBM, all three sources of randomness could be present. Thus, comparing individual trajectories/outcomes is not straightforward because an infinite number of outcomes are possible. Therefore, a comparison of two stochastic processes should be based on trajectory/outcome generalizations.

6) Monte Carlo Simulation Methods

Monte Carlo models are a class of computational approaches that rely on repeated random sampling to compute results [16]. Monte Carlo methods are often used in simulating physical and mathematical systems. Because of their reliance on repeated computation of random numbers, these methods tend to be used when it is unfeasible or impossible to compute an exact result with a deterministic model. These methods are useful in studying systems with a large number of coupled degrees of freedom and for modeling phenomena with uncertainty in inputs. It is a successful method in risk analysis when compared with alternative methods or human intuition.

II. VALIDATION STAGES

There are three steps in the validation process: (A) verification, which assesses the accuracy of the programmed model; (B) validation, which assesses the accuracy of the phenomena (as described by the model assumptions) against external criteria such as data or other factual information; and (C) sensitivity analysis, which determines the robustness of model estimates with respect to changes in model assumptions.

A. Model Verification

With a complicated computer program, programming errors can result in output that is the result of a mistake rather than a surprising consequence of the model. Verification is

 ISBN: 978-1-61208-169-4

the process of checking that a program does what it was planned to do. In the case of simulation, the difficulties of verification are complicated by many simulations being based on a stream of random numbers—meaning every run is different—and it is only the distribution of results that can be anticipated by the theory. Therefore, it is essential to debug the simulation using a set of test cases, perhaps of extreme situations in which the outcomes are easily predicted. Setting up a suite of such test cases and re-running the simulation against them—each time a major change is made—can help ensure that more errors have not been introduced. This process can be made easier by using a version control system that automatically records and tracks model results from each version of the simulation program.

B. Model Validation

Validation processes attempt to demonstrate whether the simulation is a good model of the target phenomena. A model that can be relied on to reflect the behavior of the phenomena is valid. One way to ascertain its validity is by comparing the model's output to data collected from the target. However, a few caveats are warranted:

- Both the model and the target processes are likely to be stochastic, so exact correspondence would not be expected on every occasion. Whether the difference is large enough to cast doubt on the model depends in part on the expected statistical distribution of the output measures. Unfortunately, with simulations, these distributions are rarely known and are not easy to estimate.
- Some simulations are path-dependent and early random number choices can greatly influence outcomes. Outcomes may also depend on the initial conditions chosen, which will affect the paths taken by the simulation.
- Even if the results obtained from the simulation match those from the target, there may be some aspects of the target that the model cannot reproduce.
- A model may be correct but the target data available for validation is either incorrect or not known.
- Data accuracy issues also arise when a model is intentionally highly abstract. Relating the conclusions drawn from the model to specific data from the target may be difficult. In highly abstract models, it is unclear what data could be used for direct validation. This issue arises with models that employ synthetic populations, in which the population is either intentionally remote from the simulation or does not exist at all. For these models, questions of validity are difficult to assess.

C. Model Sensitivity Analysis

Sensitivity analysis investigates how projected performance varies along with changes in the key assumptions on which the projections are based. Once a model appears to be valid, at least for the initial conditions and parameter values for which a simulation has been run, a modeler is likely to consider a sensitivity analysis to answer questions about the extent to which the behavior of the simulation is sensitive to assumptions that have been made. Sensitivity analysis is also used to investigate the robustness of a model [10, 14]. If the behavior is very sensitive to small differences in the value of one or more parameters, a modeler should be concerned about getting accurate estimates for those sensitive parameters.

The principle behind sensitivity analysis is to vary the initial conditions and parameters of the model by a small amount, re-run the simulation, and observe differences in the outcome. This is done repeatedly while systematically changing the parameters. Unfortunately, even a small set of parameters can quickly result in a very large number of combinations of variations in parameter values, and the resources required to perform a thorough analysis can be prohibitive.

Randomization of parameters to obtain a sample of conditions is one of several uses of random numbers in simulation. Random numbers can also be used to: vary exogenous factors (all the external and environmental processes that are not being modeled); model the effects of agents' innate attributes; and address simulation techniques that yield different results, depending on the order in which the actions of agents in the model are simulated.

Results from the simulation will need to be presented as distributions, or as means with confidence intervals. Once a random element is included, the simulation must be analyzed using the same statistical methods that have been developed for experimental research: analysis of variance to assess qualitative changes (e.g., whether clusters have or have not formed), and regression to assess quantitative changes.

III. METHODS

A. The General Process

A model is usually developed to examine a specific set of issues; therefore, model validity should be examined with respect to them. For example, if a disease transmission model is focused on a single epidemic period, and if the pathogen generating the epidemic confers immunity, having the model discriminate between agents that are susceptible to disease and agents that have contracted disease is important. However, if the focus of the study is to determine effective intervention strategies, the outcome of persons contracting disease is unimportant.

Model validation is difficult to make into a structured task. As a model develops, modelers should conduct formal theory predictions (analytical validity) and empirical data

 ISBN: 978-1-61208-169-4

comparisons (historical data validity). These tests can be done with varying levels of sophistication. In some cases, looking for simple equivalence is possible. In other cases, running the model hundreds of times is necessary to ensure that the results are robust across a variety of parameter settings.

After designing the model, researchers should spend a substantial amount of time testing model performance under a variety of conditions. Model components can be validated with historical data. Subject area experts can examine the face validity of the predictions to confirm the similarity of model output to their perceptions of how the modeled events should have developed and progressed. Modelers should examine their results to test the implications of the core model assumptions. If possible, they should use real data from external sources and compare model results with the external data.

Modelers should also conduct a set of experiments to set model parameters to their extreme values. Model results using extreme parameter settings should have obvious outcomes.

Once the logical boundaries of the parameter settings are determined, a sensitivity analysis can be performed on all model parameters. In this analysis, model results are generated across a wide range of theoretically feasible parameter settings. This allows the effect of each model parameter on the dependent (outcome) variables to be quantified by generating a numerical estimate of the partial derivative of the outcome variables with respect to changes in the parameter variables.

Simulation models based on ABM use more details to represent a specific model than do those based on EBM. This introduces greater opportunities for validation. Also, using the partial derivative sensitivity estimates as a criterion identifies those parameters that require accurate estimates. Validation of simulation models based on ABM in general should be judged by fidelity, realism, and resolution. These models should be validated on empirical data, as is commonly done for empirical models. Validation is possible through prediction and retrodiction. The quality of the data should be an important criterion for determining the weight of individual validation components. Sensitivity analysis is also necessary for simulations in which parameters are imperfectly measured. Finally, sensitivity analysis should be performed not only on model parameters but also on rules used by the simulation to specify the agent's interaction mechanisms.

B. General Validation Approaches

Many validation approaches have been described in the literature. In general, we will follow the procedures reported in [17]. Schreiber describes four sets of validation tests. We have added sensitivity analysis as a fifth test to assess model robustness. The five tests are defined as follows:

1. Theory-Model Tests determine whether the model describes the conceptions in the minds of the modelers.
2. Model-Model Tests connect the developed model to other pertinent models that describe the same or similar phenomena.
3. Model-Phenomena Tests connect the programmed model to the phenomena that are observed via available data.
4. Theory-Model-Phenomena Tests simultaneously examine the model in the context of both theory and phenomena. Because models, theories, and phenomena often overlap, these categories are more constructed conveniences than concrete truth.
5. Global Sensitivity Tests assess model parameter sensitivity.

C. Validation Tests

A number of validation tests are derived from the general approaches cited above. Note that these validation tests begin after the model has been verified, but in many instances they supplement the model verification processes. Examples of these tests are described below.

1) Calibration

Calibration is the process of tuning a model to fit detailed real data. This is a multi-step, often iterative process in which the model's processes are altered so that the model's predictions come to fit, with reasonable tolerance, a set of detailed real data. This approach is generally used for establishing the feasibility of the computational model; it shows that the model can generate results to match the real data. Calibrating a model may require the researcher to both set and reset parameters and to alter the fundamental programming, procedures, algorithms, or rules in the computational model. To an extent, calibrating establishes the validity of the internal workings of the model and its results.

2) Theory-Model Tests

In Theory-Model tests, the central problem is whether the model matches the theory. As programmed thought experiments, models can have a transparency (assuming the code is written clearly and assumptions are described clearly) that raw theories may lack. Theory-Model tests are also called Cross-Model Validity tests, which emphasize the connectedness of the epistemological framework.

Docking Validity Tests are standard tests of Theory-Model validity. Docking tests use a second model (developed independently) to investigate whether the index model and the second model proceed in like manner or yield similar results. Analytical Validity Tests are similar to Docking Validity Tests, except they compare results from

 ISBN: 978-1-61208-169-4

the index model with results from published accounts about the second process and/or the inputs and outputs that are connected to this process [18].

The Face Validity Tests uses the broad knowledge and experience of substantive experts as the source of the data. A model is presented to persons who are knowledgeable about the source problem, and they are asked whether this model is reasonably compatible with their knowledge and experience. The Narrative Validity Test is similar to Face Validity, but it relies on published accounts about the process usually presented by observers of the phenomena. The Narrative Validity Test is amenable to consensus from a team of scholars. Within the context of a group discussion, the group will more likely disagree about whether a model fits their experience than whether it fits a narrative description.

The Turing Test, named for mathematician Alan Turing, examines whether a group of experts can tell the difference between data generated by a model and data generated by the real world. Extreme Point Tests are useful Theory-Model approaches from two perspectives. First, they are an important debugging tool in that they frequently identify subtle code problems. Second, these tests can be used to check model behavior on extreme scenarios.

3) Model-Model Tests

Model-Model tests have a number of variations. In general, these tests involve comparing the index model with other similar models or with theoretical models. In this scenario, a commonly used test is the Cross-Model validity test [19], which validates computational models by investigating whether several models can produce the same results after changing an element/variable in the agent architecture.

Comparing two models allows modelers to recognize significant differences between model results. Identifying the assumptions that caused the differences is an important outcome because it often defines a difference in model assumption or a parameter that is imperfectly known.

4) Model-Phenomena Tests

This category of tests compares the occurrence of specific events represented in the model with the occurrence of the event as represented by real-world data. Comparing model-time series results with the results of previously collected data is one example. Some models forecast results of specific events that follow other events, or alternatively forecast the duration of a specific event. Results from these models can be compared with the actual occurrence of the sequence of the phenomena in the data.

5) Theory-Model-Phenomena Tests

These tests examine the model and the phenomena simultaneously and compare the occurrence of particular events in the model with the occurrence of the events in the source data. Historical Data tests compare model results against previously collected data of some part of the simulated scenario.

6) Global Sensitivity Analysis

Global Sensitivity Analysis tests adjust the parameter settings of the model to determine how sensitive the model predictions are when small changes are made in model parameters. If particular results, such as control strategy predictions, change as a consequence of slightly altered parameter values, then modelers should exercise caution when making claims about model outcomes. Running a comprehensive set of sensitivity analysis tests is not a trivial issue. For example, scientists are confronted with a huge parameter space and very little notion of reasonable parameter values. This requires running many simulations to determine feasible model outcomes. Given a large parameter space, enumerating every possible combination of parameters may be out of the question. This suggests a need for an adaptive process that can steer a search of the parameter space toward more useful/realistic model outcomes.

D. Component Validation Issues

So far, we have discussed tests designed to examine the entire model as a single entity. Testing individual components can also be useful, especially if the social network and agent state change driving force (e.g. disease transmission) components are disjoint entities. In this situation, validating model components allows examining the performance of the model's individual components; degenerate tests may interrupt some elements of the model and examine the impact on overall results, and trace testing examines individual agents as they work through the modeling environment. Animation methods can support this test to compare the visually displayed qualities of the model with the qualities observed in source data. Trace testing combines our theoretical expectations of the model and our observations about the model and real-world phenomena.

ABM have been criticized because of the large number of assumptions used to implement them. This increases the number of components requiring authentication in the model. However, proponents of ABM might argue that even though detailed models increase the number of component assumptions that have to be reconciled, the assumptions are presented explicitly. Most of us generally understand explicit assumptions and can therefore attempt to validate them. Consequently, they form the basis for judging the validity of one component of a model. However, implicit assumptions are often buried in the logic of EBM and are therefore hidden. In some instances, when implicit assumptions are identified, they are recognized as crude and a necessary evil, with the basic assumptions behind them unchangeable as a part of the fabric of the approach.

 ISBN: 978-1-61208-169-4

ABM and EBM use distinct approaches to describe the same process. They both make a judgment about an identical set of assumptions. ABM represent the assumptions explicitly, while EBM represent assumptions about the same set of processes implicitly, hidden within the fabric of the methodology.

Overall, representing assumptions explicitly allows ABM to expose the weaknesses of the assumptions and define new knowledge requirements for improving model performance.

IV. SUMMARY

Overall, model validation is a common problem in computational modeling of cognitive, natural, or social phenomena: Determining whether the model is the right one and if it captures the essential mechanisms behind the modeled empirical phenomenon is important. As we have seen above, model predictions can be compared with the empirical data to draw conclusions about the plausibility of the model's assumptions. However, this approach does not measure the model's accuracy with respect to unseen data or alternative models designed to explain the same phenomenon. As noted above, there are other methods of validation that can help the modeler, including drawing on the knowledge and experience of subject matter experts.

A related issue is model selection and determining whether a particular model most accurately explains the target phenomenon. Comparing several models and reporting their relative predictions is one way, but this approach often attributes superior performance to inherent model complexity or ad hoc assumptions included in the model.

The goal of modeling is to increase understanding of the underlying mechanisms of the phenomenon; a model that fits the data perfectly does not necessarily capture the essential mechanisms behind the modeled phenomenon. Instead, the model may simply be flexible enough (i.e., over parameterized) to account for the random noise introduced into the model by various means [20].

REFERENCES

[1] Sample H.A. Scientists say tsunami models should be tested. Boston.com. 2010. March 2; News/Science/Articles (col 1). September 12, 2011 <http://www.boston.com/news/science/articles/2010/03/02/scientists_say_tsunami_models_should_be_tested/>.

[2] Longini I. Jr., Nizam A., Xu S., et al. Containing pandemic influenza at the source. Science, 2005; (309):1083-1087.

[3] Ferguson N.M., Cummings D., Fraser C., Wheaton W.D., Cooley, P.C., & Burke, D.S. 2006. Strategies for mitigating an influenza pandemic. 2006. Nature. Jul 27;442(7101):448-52.

[4] Lee B.Y., Brown S.T., Cooley P.C., Zimmerman R.K., Wheaton W.D., Zimmer S.M., Grefenstette J.J., Potter M.A., Assi T., Furphy T., Wagener D.K., Burke D.S. A computer simulation of employee vaccination to mitigate an influenza epidemic. 2010. Am J Prev Med. 38(3):247-257.

[5] Lee B.Y., Brown S.T., Cooley P.C., Potter M.A., Wheaton W.D., Voorhees R.E., Lando J., Stebbins S., Grefenstette J.J., Zimmer Cooley, P., Zimmerman R.K., Assi T., Bailey R.R., Wagener D.K., Burke D.S. Simulating school closure strategies to mitigate an influenza epidemic. 2009 Dec. J Public Health Manag Pract. [Epub ahead of print].

[6] Cooley P.C., Lee B.Y., Brown S., Cajka J., Chasteen B., Ganapathi L., Stark J.H., Wheaton W.D., Wagener D.K., Burke D.S. Protecting health care workers: a pandemic simulation based on Allegheny County. 2010 Feb. Influenza and Other Viruses. 4(2), 61–72

[7] Halloran E.M., Eubank S., Ferguson M.N., Longini, M.I., Barrett C., Beckman R., Burke S.D., Cummings A.D., Fraser C., Germann C.T., Kadau, K., Lewis, B., Macken A.C., Vullikanti A., Wagener K.D., & Cooley P.C. Modeling targeted layered containment of an influenza pandemic in the USA. 2008 Mar. PNAS;105(12): 4639-4644.

[8] Van Dyke Parunak, H., Savit, R., Riolo R.L. Agent-based modeling vs. equation-based modeling: A case study and users' guide.

[9] Schlesinger, S., Crosbie, R.E., Gagne R.E., Innis, G.S., Lalwani, C.S., Loch J., Sylvester R.J., Wright R.D., Kheir N., and Bartos D. Terminology for model credibility. Simulation. 1979. 34(3):103-104

[10] Midgley D., Marks R., Kunchamwar D. The building and assurance of agent-based models: an example and challenge to the field. Journal of Business Research. 2007. Aug;60(8): 84-893. Complexities in Markets Special Issue. doi:10.1016/j.jbusres.2007.02.004.

[11] Macal C. Model Verification and Validation. The University of Chicago and Argonne National Laboratory. Workshop on Threat Anticipation: Social Science Methods and Models. 2005. April 7-9, Chicago, IL.

[12] Ziegler B.P. Theory of Modeling and Simulation. Krieger: Malabar; 1985.

[13] Boero R., Squazzoni F. Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science. Journal of Artificial Societies and Social Simulation. 2005. 8(4) 6. September 12, 2011<http://jasss.soc.surrey.ac.uk/8/4/6.html>.

[14] Rahmandad H., Sternman J. Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equation models. MIT Sloan School of Management, System Dynamics Group. May 2005. September 12, 2011 <http://web.mit.edu/~jsterman/www/Rahmandad-Sterman070906.pdf>.

[15] Suter K. The Club of Rome Revisited. ABC Science. 1999. September 12, 2011 <http://www.abc.net.au/science/slab/rome/default.htm>.

[16] Hammersley J.M. Handscomb DC. Monte Carlo Methods. London: Methuen; 1975. ISBN 0416523404.

[17] Schreiber D. Validating agent-based models: From metaphysics to applications. Midwestern Political Science Association's Annual Conference in Chicago. Apr 2002.

[18] Gilbert N., Troitzsh, K.G. Simulation for the social scientist, second edition. Berkshire, UK: Open University Press; 2005.

[19] Sargent R.G. Verification and validation of simulation models. Proceedings of the Winter Simulation Conference; 1998 December 13-16; pp. 121-130.

[20] Laine, T. Methodology for comparing agent-based models of land-use decisions. Indiana University Computer Science Dep. and the Cognitive Science Program, Bloomington. Proc. of the Sixth Annual Int. Conf. on Cognitive Modeling. 2004; 410-411; Mahwah, New Jersey; Lawrence Earlbaum.

 ISBN: 978-1-61208-169-4