Published in final edited form as:

Bull Math Biol. 2019 January ; 81(1): 1–6. doi:10.1007/s11538-018-0496-1.

Issues in Reproducible Simulation Research

B. G Fitzpatrick

Tempest Technologies and Department of Mathematics, Loyola Marymount University

Abstract

In recent years, serious concerns have arisen about reproducibility in science. Estimates of the cost of irreproducible preclinical studies range from 28 billion USD per year in the US alone (Freedman et al, 2015) to over 200 billion USD per year worldwide (Chalmers and Glasziou, 2009). The situation in the social sciences is not very different: reproducibility in psychological research, for example, has been estimated to be below 50% as well (Open Science Collaboration, 2015). Less well-studied is the issue of reproducibility of simulation research. A few replication studies of agent-based models, however, suggest the problem for computational modeling may be more severe than for lab experiments (Willensky and Rand, 2007; Donkin et al, 2017; Bajracharya and Dubox, 2013). In this perspective, we discuss problems of reproducibility in agent-based simulations of life and social science problems, drawing on best practices research in computer science and in wet-lab experiment design and execution to suggest some ways to improve simulation research practice.

Introduction

In recent years, serious concerns have arisen about reproducibility in science. Sensational reports from Amgen (Begley and Ellis, 2012) and Bayer (Prinz et al, 2011) found that 47 out of 53 and 52 out of 67 preclinical studies published in high-profile journals were not reproducible. Even the more conservative estimates of problematic research in biomedicine place the rate of reproducibility at less than 50% (Freedman et al, 2015). Moreover, estimates of the cost of irreproducible preclinical studies range from 28 billion USD per year in the US alone (Freedman et al, 2015) to over 200 billion USD per year worldwide (Chalmers and Glasziou, 2009). The situation in the social sciences is not very different: reproducibility in psychological research, for example, has been estimated to be below 50% as well (Open Science Collaboration, 2015).

Less well-studied is the issue of reproducibility of simulation research. As computational models become integrated into biological research, and as techniques such as machine learning are adopted for drug discovery, the reliability of computational results must be investigated.

Attributed to Max Gunzberger (Smith, 2017), the aerodynamics research axiom that

Everyone trusts the experiment but the wind tunnel expert; no one trusts the simulation but the computational fluids expert!

seems to suggest an even worse problem with computational reproducibility. Willensky and Rand (2007) draw close parallels between computational model building and the process of

experimental science as they detail the challenges of replicating an agent-based model (ABM) from published literature. Their replication success required extensive personal interaction with the previous model's authors. Since that effort, the "Overview, Design concept, and Details" (ODD) protocol (Grimm et al, 2006; Grimm et al, 2010; Railsback and Grimm, 2012) has provided a clear and consistent framework for model reporting. As Donkin et al (2017) and Bajracharya and Dubox (2013) discovered, a soild ODD protocol may still not be sufficient for simulation replication. Each of their studies implemented a single ABM in distinct computational environments. In each case, serious problems with reproducibility were found, even when a single team built the same conceptual model into different software implementation.

Differences across the Sciences and Engineering

In this perspective, we are focusing on the problem of reproducible stochastic simulations – and especially ABM simulations – in the context of life and social science applications. We should note important similarities and differences with physical sciences and engineering. Between-subject variability is perhaps the greatest distinction between physical and life sciences. Within-subject variation is also fundamentally different: the multiscale nature of organisms (and societies) creates variation whose stochastic characterization is more challenging than that of engineering components. Engineering in the presence of turbulence in fluid flow, optical propagation, and combustion are physical and chemical problems that approach the variability challenges of life and social sciences. A key distinguishing factor between physical sciences and engineering on the one hand and life and social sciences on the other is the between-subjects variability that complexifies living organisms at nearly every scale of interest. Admittedly with some exceptions, physical systems can often be reduced to components that operate with a high degree of certainty, determinism, and/or uniformity. Living systems are very difficult to reduce to such constituents, a problem that leads many scientists to prefer simulation technologies like ABMs (An et al, 2009; Railsback and Grimm, 2012).

The Language of Simulation Reproducibility

The term "reproducibility" itself requires some refinement of definition within contexts of experimental and simulation research. For simulations, Axtell et al (1996) suggested three levels of replication standard: *numerical identity*, in which simulation comparisons produce numerically identical outcomes; *distributional equivalence*, in which comparisons demonstrate statistical similarity in repeated simulation outcomes, and *relational alignment*, in which the results show qualitatively similar relationships between inputs/parameters and outcomes. Generally speaking, for ABMs that use stochastic simulation, numerical identity is too much to expect, and distributional equivalence and relational alignment are the replication standards of primary interest.

From setting criteria for attaining reproducibility, we must also clarify our language about the ABM itself. Following Willensky and Rand (2007), we say that an ABM is a dynamic simulation of a population of heterogeneous agents that obey specific rules. A *conceptual model* is a textual, mathematical, diagrammatic (or combination) description of the agent

characterization and processes of rule-based interaction of an ABM. An *implementation* or operationalization is a formalization of a conceptual modeling into an executable computational format in which numerical output can be derived. Typically, implementation occurs in software.

With notions of reproducibility and model system in place, we turn to the challenges.

Where the Problems Lie

Many sources of variability may confound replication studies in experiment and simulation. Willensky and Rand (2007) delineate the following challenges for implementation.

1. Time: A model reconstructed or even rerun at a different time;
2. Hardware: The computational hardware on which an ABM is implemented;
3. Languages: the software environment or programming language used to construct an implementation;
4. Toolkits: programming libraries used in conjunction with language to construct an implementation;
5. Algorithms: Underlying mathematical processes used in conceptual and implementation; and
6. Authors: Individuals building ABMs, conceptual models, and implementations.

To this list, we add

7. Translation: moving from conceptual model to implementation.

Time and hardware, by themselves, are less likely to be culprits in failures to reproduce simulation results. The problems found by Donkin et al (2017) and Bajracharya and Dubox (2013) appear to center on languages and toolkits (and perhaps algorithms on which toolkits are built). The Donkin study clearly points out the challenges of comparing NetLogo, with its high level structures, with Repast, which involves lower level programming. Willensky and Rand (2007) discuss in detail author and translation issues in their replication effort. All of these studies cite the utility of source code availability for replicators to create as high fidelity a facsimile as is possible.

Moving Forward

Two recommendations already well described in the literature involve the publication of both an ODD protocol and original source code. These are well documented in the references and their own citations. Beyond these two recommendations, we see practices in the software development and preclinical experimental research communities that may also have positive impacts on simulation reproducibility.

View the simulation as an experimental system.

An et al (2009) and An et al (2017) advocate that an ABM is more productively studied as a system in and of itself, rather than as a model of a system. In much the way model

organisms are used to investigate problems of human health, the ABM is a “middleware” object existing between traditional mathematical models and the real system of interest. Validation of agent-based models involves a number of steps. Statistical experimental design can inform thinking about simulation development, analysis, and validation (Santer, Williams, and Notz, 2003).

Validate in multiple stages.

Validation of agent-based models involves a number of steps. North and Macal (2007) suggest the following stages of an agent model validation process:

1. Requirements Validation: Have the model requirements been properly specified for the problem at hand?
2. Data Validation: Have the data used to calibrate the model been properly collected and verified?
3. Face Validation: Do the model assumptions and outputs appear reasonable?
4. Process Validation: Do the steps in model execution, agent decision, and computational flow correspond to real-world processes?
5. Theory Validation: Does the model make valid use of the theory on which it is based?
6. Agent Validation: Do agent behaviors correspond to real individual behaviors?
7. Output Validation: Do the model outputs compare to observed data?

The first six steps here connect closely to specifications required in the ODD protocol. The level of detail applied in these steps will certainly impact the ability of future researchers to use (and reproduce!) the model. The seventh step requires careful consideration in terms of (a) the goals of the model building exercise, (b) the replication standards that would arise in reproduction efforts, and (c) the software development process.

Don't document the code: code the documentation.

Since the computational simulations in which we are most interested involve software, we feel that clean code is a crucial step in reproducibility. Bob Martin in *Clean Code* (2008) notes the importance of code readability, of the structure of functions, methods, or modules, of meaningful naming. A clean code would ideally be the readable software implementation of a well-constructed ODD protocol.

Write code to the tests.

The coding philosophy of Test-Driven Development (TDD) translates a modeling component into a coding requirement and from there into specific test cases that must be passed by a code module (Beck, 2003; Madeyski, 2010; Mäkinen and Munch, 2014). Code is written to pass the tests that model the requirements.

Code written in the TDD paradigm tends to be very modular and well-aligned with the clean code design principles. As such, it can reinforce the North and Macal validation process.

Individual code module tests can be constructed to validate execution steps, agent decisions and behaviors, and model outputs. However, the focus of TDD is more on verification (are you building the thing right?), as opposed to validation (are you building the right thing?) of the model.

Use a version control repository.

Development of a complex simulation, even if a single developer is responsible for all code, requires oversight and eventual dissemination of the code, testing suite, and relevant documents. GitHub (<https://github.com>) is perhaps the best known platform for code sharing, version controlling, and developer collaboration. An example of such a repository is <https://github.com/kdahlquist/GRNmap>, a gene regulatory network modeling project.

Keep a computational laboratory notebook.

Long a key step in bench and field experimentation, the lab notebook contains a record of procedures and results. Detailing the building and running of simulations provides similar benefits to the computational experimentalist. Many institutions have policies concerning lab notebooks, but few (if any) do for computation. The article “Ten Simple Rules for a Computational Biologist’s Laboratory Notebook,” Schnell (2015) offers a solid set of record-keeping principles.

Complaints about recommendations such as these typically center on their time consuming nature. For “one run and done” computational projects, our experience is that these four suggestions add considerable time and effort to project completion. However, over a multi-year project with multiple contributors (students, post-docs, research associates, etc), having readable code with coded tests in a structured repository and a record of computational development and experimentation accelerates scientific progress dramatically. Moreover, a simulation model built on such a foundation can have much broader impact beyond the developing laboratory.

Acknowledgments

This work was partially supported by National Institute on Drug Abuse grant 1R43DA041760–01.

References

- An G, Mi Q, Dutta-Moscato J, and Vodovotz Y (2009). Agent-based models in translational systems biology, Wiley Interdisciplinary Rev Sys Bio Med 1(2), 159–171.
- An G, Fitzpatrick B, Christley S, Federico P, Kanarek A, Neilan R, Miller, Oremland M, Salinas R, Lenhart S, and Laubenbacher R (2017). Optimization and control of agent-based models in biology: A perspective,” Bulletin of Mathematical Biology 79(1), 63–87, 2017. [PubMed: 27826879]
- Axtell R, Axelrod R, Epstein JM, Cohen MD (1996). Aligning simulation models: A case study and results. Computational and Mathematical Organization Theory 1,123–141.
- Bajracharya K, and Duboz R (2013). Comparison of three agent-based platforms on the basis of a simple epidemiological model (WIP). In: Proceedings of the Symposium on Theory of Modeling & Simulation - DEVS Integrative M&S Symposium, 6–11.
- Beck K (2003). Test-Driven Development: By Example, Pearson, Boston.
- Begley CG, and Ellis L (2012). “Drug development: raise standards for preclinical research,” Nature, 483, 531–533. [PubMed: 22460880]

- Chalmers I, and Glasziou P (2009). "Avoidable waste in the production and reporting of research evidence," *Lancet* 374, 86–89. [PubMed: 19525005]
- Epstein J and Axtell R (1996). *Growing Artificial Societies*, Brookings Institution Press, Boston.
- Freedman LP, Cockburn IM, and Simcoe TS (2015). "The economics of reproducibility," *PLoS Biol* 13,6, e1002165 Available at <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165> [PubMed: 26057340]
- Grimm V, Revilla E, Berger U, Jeltsch F, Mooij WM, Railsback SF, Thulka H-H, Weiner J, Wiegand T, DeAngelis DL (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from Ecology, *Science*, 310, 987–991. [PubMed: 16284171]
- Grimm V, Berger U, DeAngelis DL, Polhill JG, Giske J, Railsback SF (2010). The ODD protocol: A review and first update, *Ecological Modelling* 221, 2760–2768.
- Madeyski L (2010). *Test-Driven Development: An Empirical Evaluation of Agile Practice*, Springer, Heidelberg.
- Mäkinen S, and Munch J (2014). Effects of test-driven development: A comparative analysis of empirical studies, in *Software Quality: Model-Based Approaches for Advanced Software and Systems Engineering*, Winkler D, Biffl S, and Bergsmann J eds, Springer, Cham.
- Martin RC (2008). *Clean Code: A Handbook of Agile Software Craftmanship*, Pearson, Boston.
- North M, and Macal C (2007). *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation*, Oxford University Press, Oxford.
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, 349, 6251.
- Prinz F, Schlange T, and Asadullah K (2011). "Believe it or not: how much can we rely on published data on potential drug targets?" *Nature Reviews Drug Discovery* 10, 712–3.
- Railsback S, and Grimm V (2012). *Agent-Based and Individual-Based Modeling: A Practical Introduction*, Princeton University Press, Princeton.
- Santer TJ, Williams BJ, and Notz WI (2003). *The Design and Analysis of Computer Experiments*, Springer, New York.
- Simmons JP, Nelson LD, Simonsohn U (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366 [PubMed: 22006061]
- Smith R (2017). Personal communication.
- Wilensky U, and Rand W (2007). Making models match: replicating an agent-based model. *JASSS* 10(4) 2.