The Practice of Archiving Model Code of Agent-Based Models

Marco A. Janssen

Arizona State University, United States

Journal of Artificial Societies and Social Simulation **20** (1) 2
<http://jasss.soc.surrey.ac.uk/20/1/2.html>
DOI: 10.18564/jasss.3317

Received: 24-Oct-2016 Accepted: 23-Nov-2016 Published: 31-Jan-2017

To evaluate the concern over the reproducibility of computational science, we reviewed 2367 journal articles on agent-based models published between 1990 and 2014 and documented the public availability of source code. The percentage of publications that make the model code available is about 10%. The percentages are similar for publications that are reportedly dependent on public funding. There are big differences among journals in the public availability of model code and software used. This suggests that the varying social norms and practical convenience around sharing code may explain some of the differences among different sectors of the scientific community.

Keywords: Bibliometrics, Replication, Open Science, Computational Science

Introduction

There is increasing concern over the repeatability and reproducibility of computational science (Barnes 2010; Joppa et al. 2013; Morin et al. 2012; Peng 2011; Easterbrook 2014). If computational scientific enterprises want to be accumulative, more transparency is required, including the archiving of computer code in public repositories. A recent study reported that around 50% of findings published in the Association for Computing Machinery (ACM) conference proceedings and journal articles could not be compiled into valid executables by computer science students, even after authors were requested to provide source code and build instructions (Collberg & Proebsting 2014). Various code repositories have been created (Stodden et al. 2012, 2015; Rollins et al. 2014; 2015; McLennan et al. 2010; DeRoure et al. 2009), but their use is limited.

In this paper, we document the practice of archiving model code for agent-based models, an increasingly popular methodology in the social and life sciences. Recent years have seen the emergence of standard platforms such as Cormas (Bousquet et al. 1998), Netlogo (Wilensky 1999), Repast (Collier 2003), and Mason (Luke et al. 2005), but also text books (Railsback and Grimm 2012; Wilensky and Rand 2015), conferences and summer schools. As such, the use of agent-based modeling has become a recognized method in the life and social sciences.

Since the use of what we now call “agent-based modeling” did not originate from a particular

discipline or application, we may expect that the applications will spread widely across various disciplines. Part of this exercise is to map the use of the method in different fields and document whether there are different practices in sharing model code and model documentation.

In the rest of this paper, we first describe the methodology used to derive a sample of 2367 publications presenting the results of agent-based models and the protocol we used to collect metadata on the availability of model code, the software used, and the way models are documented. We, then, report the descriptive statistics of the data and perform a network analysis of the publications citing each other. We conclude with a discussion on the implications of our findings.

Methodology

In order to derive a sample of relevant publications, we used the search term "agent-based model*" on the ISI Web of Science database in the spring of 2015 for publications up to 2014. The term "agent-based model*" could be used in the title, abstract or keywords. This resulted in 2855 publications. All publications were evaluated in order to verify that it was about an agent-based model. Reviews, conference abstract or presented conceptual models were discarded. This resulted in 2367 publications that report a model and results of model simulations.

For each publication, we checked whether the model code was made available through a provided URL to a website or as an appendix. We also checked whether the URL was still available. Hence, our criterion on public availability of the model code depends on the valid information provided in the article. We recognize that the model code could be published online but not mentioned in the article or could be provided by authors if we had requested this. As such our estimate of the public availability of model code is an underrepresentation of what might be available with more investigation.

Furthermore, we listed which programing platform was used and which sponsors funded the research. Finally, we recorded how the model was described in the articles and appendices. Based on Müller et al. (2014), we distinguished the following items:

- **Narrative**. How was the model description organized? Did it use a standard protocol called Overview-Design-Details (ODD) (Grimm et al. 2006), or did it use a non-prescriptive narrative.
- **Visualized Relationships**. How were the relationships visualized? Did it include flow charts, a Unified Modelling Language (UML) diagram or provide an explicit depiction of an ontology that describes entities and their structural interrelationships.
- **Code and formal description**. How were the algorithmic procedures documented? Did the authors provide the source code? Did they describe the model in pseudocode or use mathematical equations to describe (parts) of the model?

The downloaded information from ISI Web of Science included references for each article. This information was entered into a database and unique identifiers were provided for the publications in order to perform a network analysis. The resulting database can be found at: https://osf.io/8n663/.

Results

Out of the 2367 articles 236 articles contained information (often via a link to an online database) on the availability of the source code, which is 10.0%. Excluded from the count were 69 articles which provided a link to online databases, but either the website did not exist anymore or the link was password protected. Although authors may be able to provide the code if one requests it, as sometimes stated in the publication, we only consider a model code publicly available if the actual code is made publicly available. In some cases, code might have been made available without mentioning it in the publication. But this would be unknown to us since we only rely on the information in the publication.

Figure 1 describes the number of publications on agent-based models over time. Each publication is a new or updated agent-based model for which computer code is used to generate the published results. We see an exponential increase of the number of publications. Figure 2 shows that the percentage of the publications that makes the model code publicly available is below 10% until 2012 and increases to about 15% in 2014. With the rapid increase of the absolute number of publications, this means a very sharp increase of the amount of model code made publicly available. Nevertheless, for 90% of the publications the model code is not publicly available, which will hinder replication of the results and the accumulation of knowledge.

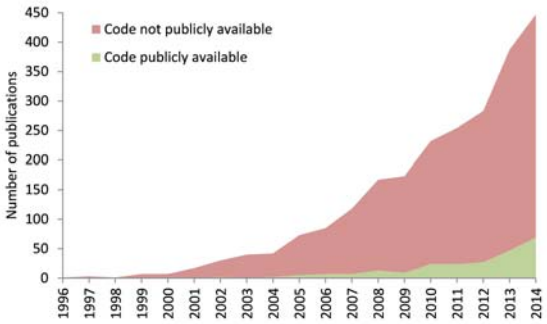

Figure 1. Number of publications over time.

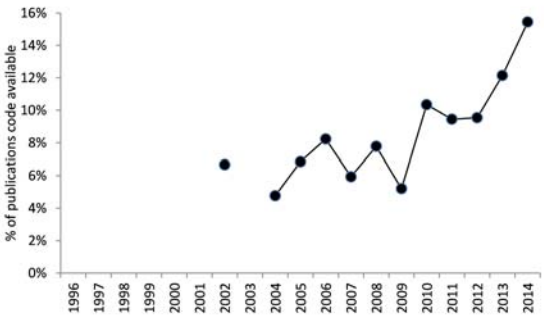

Figure 2. Percentage of publications for which model code is publicly available.

What facilitated the increase of archiving model code? To investigate this we traced where the code was made available (Table 1). The most common option is to have the code available on the journal publisher's website. The next most common option is to have the code available on the

author's personal website or that of the researcher or research group. In some cases, authors made their code available via a Dropbox link or ResearchGate post. There are various public archives for computer code such as Github, SourceForge, CCPForge, Bitbucket, Dataverse and GoogleCode, but the most commonly used archive is the specialized Computational Model Archive at OpenABM.org with code of 55 publications from our data set. Finally, we consider platform specific repositories such as Netlogo and Cormas.

Table 1: The locations where source code was stored, as referred to in the journal articles

Location name	Description	Number of publications
Journal	As supplementary information	72
Personal	Websites of researchers or research groups	71
OpenABM	Computational Model archive at https://www.openabm.org/	55
SourceForge	https://sourceforge.net/	9
Github	https://github.com/	8
Netlogo	http://modelingcommons.org/ or https://ccl.northwestern.edu/netlogo/models/community	7
Cormas	http://cormas.cirad.fr/	6
CCPForge	https://ccpforge.cse.rl.ac.uk/	3
BitBucket	https://bitbucket.org/	1
Dataverse	https://dataverse.harvard.edu/	1
Dropbox	Dropbox.com	1
GoogleCode	https://code.google.com/	1
ResearchGate	https://www.researchgate.net	1
Invalid	URLs did not work or password protected	69

Figure 3 shows the use of different locations where code is archived over time. This demonstrates the increase of the use of open source archives, especially OpenABM. Figure 3 also demonstrates that model code that was available for publications about 10 years ago are often not accessible anymore. This demonstrates the importance of storing model code and documentation in public archives to preserve the scientific output for future generations.

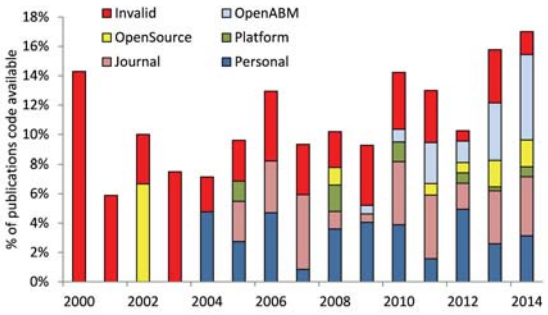

Figure 3. The percentage of model publications split up in different categories where the source code of the model is available.

The 2367 publications appeared in 722 different journals which demonstrate the spread and scope of the use of agent-based models. The 10 most popular journals are listed in Table 2. This table shows the wide diversity of standards and practices of the journals. The popular journal JASSS has a high percentage of publications that make the model code available. They also indicate in their guidelines: "Authors are strongly encouraged to include sufficient information to enable readers to replicate reported simulation experiments." Although it is not a requirement, the journal encourages authors to share model code.

On the other hand, a popular journal such as Physica A has no articles for which model code is made available. The short articles in this journal typically describe models mathematically and present results of computer simulations.

Table 2: Model code availability of the 10 most popular journals in the database

Journal	Number of publications	% of code publicly made available
JASSS – The Journal of Artificial Societies and Social Simulation	135	42.2%
Physica A	103	0%
PLoS ONE	87	9.2%
Ecological Modelling	61	27.9%
Journal of Theoretical Biology	60	11.7%
Environmental Modelling and Software	50	34%
Advances in Complex Systems	50	10%
Computational and Mathematical Organization Theory	31	19.4%
Computers, Environment and Urban Systems	30	6.7%
Environment and Planning B	30	0%

Since most research is sponsored by tax money, sponsors often explicitly require that the data, including software code, be made publicly available. About 55% of the publications list the sponsors of their research. In some cases, these are multiple sponsors. In Table 4 we list the 10 most common sponsors mentioned and provide the percentage in which model code is made publicly available. This table clearly shows that publicly funded research does not produce a higher percentage of publications with publicly available model code. The numbers suggest that there is no enforcement of public data availability required by the sponsors.

Table 4: Model code availability for the 10 most common sponsors

	Number	Percentage that made code publicly available
NSF (USA)	258	14.7%
NIH (USA)	170	14.7%
European Commission	110	5.5%
National natural Science Foundation of China	74	5.4%
United Kingdom Engineering and Physical Sciences Research Council	31	16.1%
Netherlands Organization for Scientific Research	22	13.6%
Netherlands Organization for Scientific Research	74	5.4%
Natural Sciences and Engineering Research Council of Canada	20	5%
Australian Research Council	18	5.5%
German Research Foundation	18	5.5%
United States Army	15	6.7%

Which software platforms were used? Not every manuscript provides information on which software is used. In fact, 1223 of the 2367 publications (52%) do not provide information on the software implementation. Of those who provide information, we find more than 100 different types of platforms and computer languages. Some publications use combinations of platforms and languages. In Table 5, we list the 10 most commonly used platforms and languages as mentioned in the publications. Netlogo and Repast are the most common, and they are agent-based modeling specific platforms.

Table 5: Model code availability for the most common platforms or programming languages

Platform	Number	Percentage publicly available code
Netlogo	312	32.4%
Repast	147	18.4%

C(++)	137	17.5%
Java	95	18.9%
Matlab	85	14.1%
AnyLogic	44	4.5%
Swarm	37	13.5%
Python	33	24.2%
CORMAS	31	29.0%
R	28	25.0%

How are models described in journal publications? Table 6 reports the various ways in which models are documented. A verbal narrative is the most frequent description. A more precise narrative is the ODD protocol (Grimm et al. 2006) which provides a structured description of the different components and mechanisms of the model. A mathematical description is also commonly used, but note that this does not mean that in all those publications a complete mathematical description is provided. In many cases, some key equations are provided which are essential to understand the model together with the verbal narrative.

Table 6: Relative frequencies in which models are described in the publication

Types of documentation	**Percentage**
Verbal Narrative	93.3%
Mathematical description	53.5%
Flowcharts	34.2%
Source Code	10.0%
Pseudo Code	9.7%
ODD Protocol	6.7%
UML	3.2%

Do model publications build on each other? Among the 2367 publications, there are 2704 citations, which is an average of 2.3 connections of each paper. We map the network of connections between the articles in Figure 4 using the ForceAtlas 2 algorithm in the network visualization tool Gephi. We focus here on the largest number of connected papers in the network. Based on the evaluation of the paper topics in the various clusters of the network, we indicate different topic areas. The most dense topic area of interactions (meaning citations) is land use change modeling. This is an application area of agent-based modeling that has many users. Figure 4 also demonstrates that the lack of archiving model code is widespread among all research domains. Figure 5 depicts the software that is used, as mentioned in the publication. We see here also that the 10 most commonly used languages and platforms are used among all topic areas.

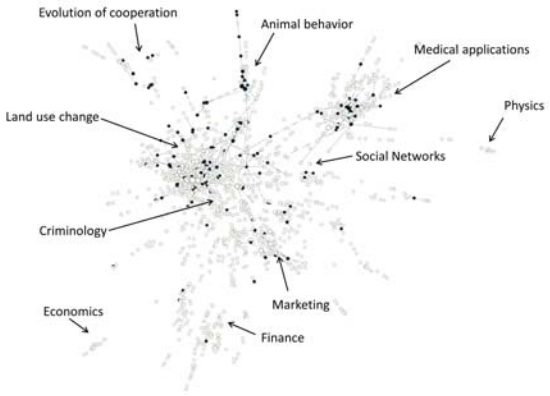

Figure 4. Network of model publications connected with other model publications among the 2367 publications in the dataset. Green nodes define whether the model code is publicly available. Red nodes define whether model code is not publicly available. Note that only publications are depicted that have a connection with another publication..

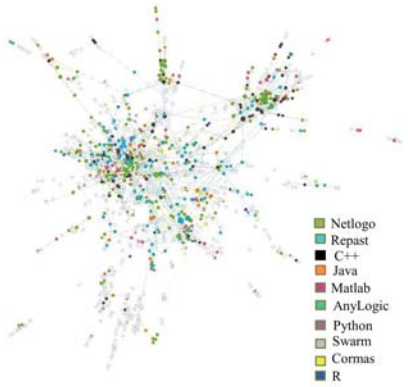

Figure 5. The network of publications with connections to other model publications colored according to the known use of the computer language or platform. White nodes indicate that the software used is unknown or is a less frequently used platform.

Conclusions

In this article, we provided a brief report on the practice of making agent-based model code publicly available. We relied on information in the publications. We found that about 10.0% of the publications provide model code, and that this percentage is increasing. We noticed major differences between journals and platforms. The increasing use of some common easy to use platforms like Netlogo and R, makes it more convenient to share model code, but journals need to facilitate this. Most journals do not provide any information on requirements for computational studies in their journals. Only recently, some high profile journals have started to encourage transparency of scientific research by improving the standards of reproducibility (McNutt 2014). So far, the focus is on biomedical and behavioral research, but computational research is expected to follow (Alberts et al. 2015).

Why should we care about the level of model code sharing? Given that science is an accumulative process of knowledge production, the lack of information about what other scholars have done might slow down the process. Furthermore, reinventing the wheel because colleagues do not

share their codes is a waste of resource funds. Sharing code would increase the pace and quality of knowledge production. However, there is also substantial cost to individuals to make their code available. They will have to spend additional time to document their work well, and clean up their code. However, this would also improve their own ability to update their work many years later. Another challenge is that not all information can be shared. If sensitive data is used, the actual data files might not be provided, but some placeholders to demonstrate the model could be used. In some cases, the sponsors of the research may restrict the dissemination of the model code. It would be up to the journal to determine whether such a publication can still be considered for a scientific journal.

What is clear is that there is need for improving incentives for researchers to dedicate the time and effort required to write detailed model descriptions, including providing source code and associated metadata and ensuring the accessibility of the necessary runtime environment. It would be more effective if these incentives could be embedded within the actual system of incentives that are already established in the academic world: public recognition by peers through citations and recognition by employers and scholarly organizations as evidence of valuable research activity. Related incentives include requirements by funding agencies and journals to sufficiently document and disseminate model-based research (Morin et al. 2012; Peng 2011).

These are only initial results of a broader project to map the field of computational modeling. Since most publications have been only recently published, due to the exponential increase of agent-based model publications, the impact of model availability on citations cannot yet be evaluated in a reliable way. A further extension of the database will include a broader range of agent-based simulation models (including those that use different terms like multi-agent simulation, agent-based simulation and agent-based computational economics), as well as updating the database with more recent publications. The resulting database will enable us to derive a better understanding of the practices in the rather fragmented scholarly landscape of computational modeling.

In conclusion, sharing the model code of agent-based models is still rare but the practice is now slowly improving. The technical facilities are available to archive model code. However, to increase the actual sharing of model code and enhance knowledge accumulation, journals are called to improve their standards and research sponsors must enforce their policies in such a direction.

Acknowledgements

The author would like to acknowledge the contributions of Allen Lee, Calvin Pritchard and Tejas Patel in developing the software for the metadata and Michael Barton for his feedback on the analysis. The author also would like to thank Mady Tyson, Rachael Gokool, Yee-Yang Hsieh, and Juan Rodriguez for entering and analyzing metadata. We acknowledge financial support for this work from the National Science Foundation, grant numbers 0909394 and 1210856.

References

ALBERTS, B., Cicerone, R.J., Fienberg, S.E., Kamb, A., McNutt, M., Nerem, R.M., Schekman, R. Shiffrin, R. Schekman, R. Shiffrin, R., Stodden, V., Suresh, S., Zuber, M.T., Pope B.K. and Jamieson, K.H. (2015). Self-correction in science at work. *Science*, 348: 1420-1422.

BARNES, N. (2010). Publish your computer code: it is good enough. *Nature*, 467: 753.

BOUSQUET, F., Bakam, I., Proton, H. and Le Page, C. (1998). Cormas: common-pool resources and multiagent systems. *Lecture Notes in Artificial Intelligence*, 1416: 826–837.

COLLBERG, C. and Proebsting, R.A. (2016). Repeatability in computer systems research. *Communications of the ACM*, 59(3): 62-69.

COLLIER, N. (2003). REPAST: an extensible framework for agent simulation: http://repast.sourceforge.net.

DE ROURE, D., Goble, C. and Stevens, R. (2009). *The design and realisation of the myExperiment virtual research environment for social sharing of workflows, future generation computer systems*, vol. 25. pp. 561– 567: http://eprints.soton.ac.uk/id/eprint/265709.

EASTERBROOK, S.M. (2014). Open code for open science. *Nature Geoscience*, 7: 779-781.

GRIMM, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Müller, B., Pe'er, G., Piou, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R.A., Vabø, R., Visser, U. and DeAngelis, D.L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modeling*, 198, 115-126.

JOPPA, L.N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D. and Emmott, S. (2013). Troubling trends in scientific software use. *Science*, 340, 814–815.

LUKE, S., Cioffi-Revilla, C., Panait, L., Sullivan, K. and Balan, G. (2005). MASON: a multiagent simulation environment. *Simulation*, 81:517–527.

MCLENNAN, M. and R. Kennell (2010). HUBzero: a platform for dissemination and collaboration in computational science and engineering. *Computing in Science & Engineering*, 12(2): 48–52.

MCNUTT, M. (2014). Journals unite for reproducibility. *Science*, 346: 679.

MORIN, A., Urban, J., Adams, P.D., Foster, I., Sali, A., Baker, D. and Sliz P. (2012). Shining light into black boxes. *Science*, 336: 159–160.

MŰLLER, B., Balbi, S., Buchmann, C.M., de Sousa, L., Dressler, G., Groeneveld, J., Klassert, C.J., Bao Le, Q., Millington, J.D.A., Nolzen, H., Parker, D.C., Polhill, J.G., Schlüter, M., Schulze, J., Schwarz, N., Sun, Z., Taillandier, P. and Weise, H. (2014). Standardised and transparent model descriptions for agent-based models: current status and prospects. *Environmental Modeling and Software*, 55: 156-163.

PENG, R.D. (2011). Reproducible research in computational science. *Science*, 334: 1226 –1227.

RAILSBACK, S.F. and V. Grimm (2012). *Agent-based and individual-based modeling: a practical introduction*. Princeton University Press, Princeton.

ROLLINS, N.D., Barton, C.M., Bergin, S., Janssen, M.A. and Lee, A. (2014). A computational model library for publishing model documentation and code. *Environmental Modeling and Software*, 61: 59-64.

STODDEN, V., Hurlin, C. and Perignon, C. (2012). RunMyCode.org: a novel dissemination and

collaboration platform for executing published computational results. Proc. IEEE 8th Int'l Conf. E-Science: http://dx.doi.org/10.2139/ssrn.2147710.

STODDEN, V., Miguez, S. and Seiler, J. (2015). ResearchCompendia.org: cyberinfrastructure for reproducibility and collaboration in computational science, computing in science and engineering. Jan/Feb: 12-19.

WILENSKY, U. (1999). NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (1999) http://ccl.northwestern.edu/netlogo.

WILENSKY U. and Rand, W. (2015). *An introduction to agent-based modeling: modeling natural, social and engineered complex systems with NetLogo*. MIT Press, Cambridge.