Power Analysis for Agent-Based Modeling Determining the Appropriate Number of Runs

Davide Secchi[1] and **Raffaello Seri**[2]

Abstract.

Agent-based modeling (ABM) is emerging as a powerful tool for the analysis of human behavior in several situations in which direct observation is impossible or unfeasible [4]. This computational simulation technique is still at its early stages in the social sciences [24], with sociology starting before other disciplines [7] and management reluctantly catching up [19]. The enthusiasm that many have shown with ABM is witnessed by the increasing number of scientific outlets that publish such simulations—among many are the journal of *Computational and Mathematical Organization Theory* and the *Journal of Artificial Societies and Social Simulation*.

It has been pointed out that the diffusion of ABM is to be found, among other aspects, in an increase of available computational power [6], a set of simplified tools for modeling [19], and the ability to create extremely complex and descriptive simulations [5]. As a result, these simulations are usually employed to study configurations, structures, and outcomes that are not directly intelligible by a simple review of the initial conditions of a system. In other words, ABM are a tool to study emergent properties of complex systems among other characteristics.

In spite of the very significant use of these simulations, there is still uncertainty on the conditions under which a simulated system produces relevant or irrelevant results—in relation to a given research question. The issue is particularly relevant for the study of emergent properties in that some of them may result from misinterpreting the data of the simulated model. These issues can be divided into two, separate but (probably) interconnected. We write 'probably' because, to our knowledge, there are no studies robustly and soundly connecting these two issues together. The first issue relates to the question: when is it the right time to stop the simulation in a single run? This is a problem concerning the length of a run—some software may call these 'steps'—, given a certain configuration of the simulation's parameters. It can be rephrased as to when it is enough time for emergent properties to emerge. Length may depend on the research question, on the type of simulation, or it can be calculated [20]. To partially address these issues, some propose steady state and stationarity analyses [8]. The second issue relates to the number of runs per configuration of parameters: how many times should one run the simulation for each configuration of parameters? This is a problem of establishing an appropriate number such that the data is not biased or that effects are discarded when they should not. Some have suggested power analysis as a viable technique to tackle with this problem [16, 21, 22]. Given the relevance of this latter issue, this abstract is dedicated to it, in an attempt to link emergent properties of the ABM to the appropriate number of runs to be performed.

[1] University of Southern Denmark, Denmark, email: secchi.davi@sdu.dk
[2] Insubria University, Italy, email: raffaello.seri@uninsubria.it

The use of ABM in these situations is generally accomplished by identifying some parameters whose impact on the results of the simulations is deemed relevant. For each one of these parameters, a certain number of values are chosen to represent the whole range of variability of the parameters. The researcher then identifies a certain number of configurations of values. For each configuration of the parameters, the model is run several times, say n, and the values of one (or more) outcomes for each run are collected. The impact of the values taken by the configurations of parameters on the expected outcomes can then be compared using an ANOVA test. The null hypothesis H_0 of the test posits that, for example, the different parameter configurations have no impact on the average values of the outcome variable. Failing to reject the null hypothesis when it is true is very important and it is done to avoid what can be framed as a *false positive*, i.e. accepting a result when it should be rejected. In statistics this event is called Type-I error and, although it is considered good practice to test for statistical significance as it relates to Type-I error avoidance, this does not exclude the possibility that Type-II error manifests itself. This corresponds to failing to reject the null hypothesis when it is false, an error that can be referred to as a *false negative*. In order to develop testing methodologies that reduce this mistake, statisticians have since long identified a quantity of interest, called *statistical power* [2, 14] and corresponding to the probability that the null hypothesis is rejected when it is false, i.e. to one minus the Type-II error rate. High power gives the researcher the confidence that Type-II error is under control and that potentially meaningful results are not discarded on the basis of a false null hypothesis being wrongly accepted. As statistical power increases with the number of observations on which the test is performed, this concept allows the researcher to answer to a different question, i.e. how many times a simulation should run.

Therefore, we explore the issue of how many times a simulation should run. This is an often neglected issue [16, 21] that, sooner or later, all modelers dealing with simulations of complex systems encounter. The literature takes an agnostic stance on how many runs n—per configuration of parameters—a simulation is to be performed. In fact, the focus has mostly been on defining the 'steps,' the time, or the interactions within each run through sensitivity and convergence analysis [15, 17, 23]. Our central assumption is that the number of runs to perform in a simulation is crucial for results to bear some meaning. Of course, this is not true for all simulations and it depends on scope, nature of the simulated phenomenon, purpose, and level of abstraction. For social simulations with a strong stochastic component where emergence and complexity [1] make results differ even within the same configuration of parameters, knowing how many times are enough for differences to emerge (or not) becomes an extremely relevant information.

We discuss the use of statistical power to decide the number of runs of an agent-based model. We first try to indicate—very broadly—to what type of simulations this approach may apply. Then, mediating from research on sample size determination for the behavioral sciences, we introduce some considerations on statistical power analysis and testing theory. We provide an overview of how statistical power analysis can be applied to agent-based models and simulations. The objective is accomplished explaining why modelers should employ measures of power. We highlight some of the important positives of using this tool in ABM, specifically (a) determining the number of runs, (b) avoiding models that can be better analyzed with simpler tools, and (c) learning how to manipulate parameter values.

We also stress the dangers of underpower, i.e. the situation in which power is not sufficient, and overpower, i.e. the situation in which power is too high. As far as underpowered simulations are concerned, tests may fail to reject null hypotheses that are, in fact, false. Instead, in overpowered studies, the risk is that they may lead modelers to notice effects so small that are not worth considering. All in all, overpowered simulations end up being less reliable than appropriately powered simulations. But it is clear that underpower is generally more dangerous than overpower. As a consequence of this, we discuss the appropriate levels of Type-I and Type-II error rates to be used in computational simulations. We strongly advocate that researchers choose a significance level α equal to 0.01 (while the standard in social sciences is 0.05) and strive for a power π of at least 0.95 (while the standard in social sciences is 0.80). We justify the strengthening of the usual values with the fact that the conditions under which the ABM experiment is performed are at least partially set by the researcher, so that imposing stricter requirements comes quite naturally as a tool to prevent personal biases.

Once the values for α and π have been set, the value of n can be recovered if only a further quantity is specified, the so-called *effect size*, a measure of how far the true data are from the null hypothesis. Several techniques are available to infer a plausible value for the effect size. A first technique is to compute the effect size on the basis of the data just collected, in order to assess whether the level of power is sufficient. However, for some technical reasons among which the high variability, the use of post-experiment power calculations as a data analytic tool is considered inappropriate and should be avoided as statistical malpractice [9]. A second technique is to use "canned" effect sizes [2, 3], i.e. some average values obtained by a review of the literature undertaken by Cohen. The use of these quantities has been criticized in statistics [13], in spite of the fact that they are generally used in social sciences and they pose less problems than the first technique detailed above. A third technique is to use similar studies that have been performed in the past [13, 11]. While this is generally recognized as good practice when the literature on which the effect size is computed is substantial, it has been criticized because it provides sample sizes with large variability [10] when the number of studies is small. Therefore, we suggest that researchers use an eclectic approach in which previous studies, pilot runs and canned effect sizes are used to obtain a guess of the effect size that goes under the name of *smallest effect size of interest* or SESOI (see [12] for the definition in a different context).

As an illustration, we take an agent-based model (ABM) with a strong stochastic component and provide two examples that show how crucial the issue is and, at the same time, offer a practical guide on how to conduct the computation. Implications and concluding remarks follow.

REFERENCES

[1] P.W. Anderson, 'More Is Different', *Science*, **177**(4047), 393–396, (August 1972).

[2] J. Cohen, *Statistical power analysis for the behavioral sciences*, Hillsdale, NJ: LEA, 2nd edition edn., 1988.

[3] J. Cohen, 'Quantitative methods in psychology: A power primer', *Psychological Bulletin*, **112**(1), 155–159, (1992).

[4] *Simulating Social Complexity. A Handbook*, eds., Bruce Edmonds and Ruth Meyer, Heidelberg: Springer, 2013.

[5] Bruce Edmonds and Scott Moss, 'From KISS to KIDS — an 'anti-simplistic' modelling approach', in *Multi Agent Based Simulation*, ed., P. Davidson, volume 3415 of *Lecture Notes in Artificial Intelligence*, 130–144, New York: Springer, (2005).

[6] Guido Fioretti, 'Agent-based simulation models in organization science', *Organizational Research Methods*, **16**(2), 227–242, (2013).

[7] N. Gilbert and P. Terna, 'How to build and use agent-based models in social science', *Mind and Society*, **1**, 57–72, (2000).

[8] Jakob Grazzini, 'Analysis of the emergent properties: Stationarity and ergodicity', *Journal of Artificial Societies and Social Simulation*, **15**(2), 7, (2012).

[9] J.M. Hoenig and D.M. Heisey, 'The abuse of power: The pervasive fallacy of power calculations for data analysis', *American Statistician*, **55**(1), 19–24, (2001).

[10] E.L. Korn, 'Projecting power from a previous study: Maximum likelihood estimation', *The American Statistician*, **44**(4), 290–292, (1990).

[11] D. Lakens, 'Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas', *Frontiers in Psychology*, **4**, 863, (2013).

[12] D. Lakens, 'Performing high-powered studies efficiently with sequential analyses', *European Journal of Social Psychology*, **44**(7), 701–710, (December 2014).

[13] R.V. Lenth, 'Some practical guidelines for effective sample size determination', *American Statistician*, **55**(3), 187–193, (2001).

[14] X.S. Liu, *Statistical Power Analysis for the Social and Behavioral Sciences*, New York: Routledge, 2014.

[15] D. Mungovan, E. Howley, and J. Duggan, 'The influence of random interactions and decision heuristics on norm evolution in social networks', *Computational and Mathematical Organization Theory*, **17**(2), 152–178, (2011).

[16] F.E. Ritter, M.J. Schoelles, K.S. Quigley, and L. Cousino-Klein, 'Determining the numbers of simulation runs: Treating simulations as theories by not sampling their behavior', in *Human-in-the-loop simulations: Methods and practice*, eds., L. Rothrock and S. Narayanan, 97–116, London: Springer, (2011).

[17] S. Robinson, *Simulation. The Practice of Model Development and Use*, New York: Palgrave, second edn., 2014.

[18] D. Secchi and R. Seri, 'Controlling for false negatives in agent-based models: a review of power analysisin organizational research', *Computational and Mathematical Organization Theory*, **23**(1), 94–121, (2017).

[19] Davide Secchi, 'A case for agent-based model in organizational behavior and team research', *Team Performance Management*, **21**(1/2), 37–50, (2015).

[20] Davide Secchi and Raffaello Seri, "How many times should my simulation run?' Power analysis for agent-based modeling', in *European Academy of Management Annual Conference*. Valencia, Spain, (2014).

[21] Davide Secchi and Raffaello Seri, 'Controlling for 'false negatives' in agent-based models: A review of power analysis in organizational research', *Computational and Mathematical Organization Theory*, **23**(1), 94–121, (2017).

[22] Raffaello Seri and Davide Secchi, 'The problem of determining the number of runs in your simulation', in *Simulating Social Complexity. A Handbook*, eds., Bruce Edmonds and Ruth Meyer, in press, Heidelberg: Springer, 2nd edn., (2017).

[23] J. Shimazoe and R.M. Burton, 'Justification shift and uncertainty: why are low-probability near misses underrated against organizational routines?', *Computational and Mathematical Organization Theory*, **19**(1), 78–100, (2013).

[24] Klaus G. Troitzsch, 'Historical introduction', in *Simulating Social Complexity. A Handbook*, eds., Bruce Edmonds and Ruth Meyer, 13–21, Heidelberg: Springer, (2013).