STATISTICAL INFERENCE AND ARCHAEOLOGICAL SIMULATIONS

Enrico R. Crema

Enrico R. Crema is a lecturer in the Department of Archaeology, University of Cambridge.

The last decade saw an exceptional increase in the number of archaeological simulations covering a range of topics as diverse as settlement dynamics, spread of farming, origins of inequality, and cultural evolution. The wider accessibility of dedicated programming languages (e.g., NetLogo for Agent-Based Modelling), and the flexibility of general-purpose data science languages (e.g., R and Python) are enabling a new generation of scholars and students to dive into the field of archaeological simulations with less effort than ever. Retrospective papers are continuously being published, and it is becoming increasingly common to see research projects including a "modeller" postdoctoral position. Simulations are certainly not new in archaeology (see Lake 2014 for a historical review), and future archaeologists will know whether we are finally approaching the "plateau of productivity" of the notorious hype cycle or if we are still ascending the "peak of inflated expectation" (and about to face a "trough of disillusionment").

The question of where we are now and where we are headed becomes harder (if not pointless) to answer once we consider the many flavours of archaeological simulations that have been proposed in the last 40 years. Several classifications have been suggested to make sense of this rich variation, using criteria ranging from thematic content of the model to its degree of realism or abstraction. One such classification, originally devised by Mithen and discussed in detail by Lake (2014), focuses on the ultimate objective of the simulation model. Simulations can thus be used to support *theory building* by providing a heuristic device to explore the implications of one or more behavioural assumptions; be part of *method development*, by generating artificial datasets to test the efficiency and the limitations of an analytical technique; or be based on a particular historical-geographic context, with outputs that are directly comparable to observed data, and hence be employed for *hypothesis testing*. Pure *theory-building* models have a long history across different fields of studies, and although occasionally criticised for their high levels of abstraction (cf. Dronamraju 2011: Ernst Mayr's critique of "beanbag genetics" in population genetics), they reached substantial maturity in subfields such as cultural evolutionary studies. The success of *method development* simulations are harder to evaluate, partly because of the comparatively small number of archaeologists engaged in the development of new statistical techniques. Nevertheless, these simulations are successfully being employed to assess the reliability of existing techniques (within the field or "borrowed" from other disciplines), often to ascertain whether they are robust to different forms of archaeological biases such as spatially uneven sampling strategies, time-dependent taphonomic loss, and time-averaging (e.g., Crema et al. 2017; Premo 2014).

The third category in this classification—*hypothesis-testing* models—is the main focus of this article. As pointed out by Lake (2014), the distinction between *hypothesis-testing* models and *theory building* is not always clear-cut, and very often the two objectives coexist informally in the same simulation. This is particularly the case for empirically grounded models designed to emulate a specific historical window and from this to explore more general aspects of human behaviour such as the emergence of hierarchical societies (Kohler et al. 2012). While there have been discussions on the effectiveness of such a "realist-particularist" approach (Costopoulos 2015; Kohler 2015), it is undeniable that a substantive number of archaeological simulations are designed to emulate, whether for hypothesis testing or external validation, some aspects of reality in a predefined window of time and space. Yet, formal comparisons between simulation outputs and the empirical record are not as frequently carried out as one might expect, and in many cases attempts have been limited to visual or qualitative inspections. Moreover, the dominant focus of many works has been model building and description (and to a lesser extent parameter exploration), with far less attention given to the fit (or the lack thereof) between simulation outputs and observed data as well as the broader implications of the whole exercise.

Two, Four, and Six

In 1960, P.C. Wason published the results of a psychological experiment that aimed to explore a particular form of inferential bias. Participants were presented with a numerical sequence—2, 4, and 6—and were asked to identify the underlying "rule" generating the numbers. To aid the process, they were allowed to propose as many "test" sequences of three numbers as they wished and were informed whether the proposed triplet could also be generated by the rule. Most participants proposed sequences designed to confirm their initial clue (e.g., that the rule was "increasing intervals of two", suggesting for example 10, 12, and 14). When informed that their proposed triplet could also be generated by the algorithm, the participants would stop the testing procedure (or continue further tests with the same hypothesised rule), ultimately concluding that their algorithm was the correct one. The right answer, however, was "increasing numbers," a simple rule that can yield a wide range of sequences (e.g., 1,2, and 3; 5, 25, and 125; 10, 98, and 99; etc.), matching outputs from a variety of alternative and more complex rules (e.g., "increasing intervals of two," "increasing multiple of the first number," etc.). Because the majority of participants were aiming to reproduce the observed pattern and hence seeking to "prove" their hypothesis, they failed to identify the correct answer. In contrast, an approach designed to *disprove* an initial clue (e.g., by testing 4, 5, and 6) would have avoided such a mistake.

Wason's study highlights our natural tendency to seek confirmation (rather than rejection) of our theories and hypotheses and, more importantly, how this can lead to erroneous conclusions when dealing with patterns that could have been generated from more than one possible generative process—i.e., when we are dealing with *equifinality*. The theoretical implications of this problem and the related issue of *multifinality* (same process, multiple possible patterns) have been discussed in the literature (see, for example, Premo 2010), and it is known to have even contributed to the abandonment of the whole enterprise by early adopters such as Ian Hodder (Lake 2014).

It would follow that if we are pursuing *hypothesis testing*, and wish to avoid Wason's inferential pitfall we should be designing simulation models to *disprove* our theories rather than seeking their confirmation. This, however, introduces a paradoxical situation where the worst outcome in the external validation of a computer simulation is a perfect fit to data. A complete lack of fit can help dismiss parameter ranges, or question the validity of key assumptions, while a partial fit can generate new ideas on "what is missing," with the simulation model acting as a comparative template (Kohler et al. 2012). A perfect fit, which arguably would be rarely achieved (especially with highly realistic models), would be less informative—there would be nothing left to explain; alternative explanations are not considered and hence cannot be dismissed a priori and, because of equifinality, we are not able to conclusively state that the proposed model is the "correct" one.

A "Generative" Statistical Inference

It is surprising that the issues of *equifinality* and *multifinality*, which are at the core of this problem, have not been discussed in relation to inferential statistics where the comparison of model and data is the disciplinary bread and butter. At its foundation, statistics is based on probability distributions, which capture the expected variation in the observed data given a parameter value of a statistical model (e.g., what are the probabilities of getting 0, 1, 2, 3, and 4 heads given 4 tosses of a coin with a probability of heads equal to 0.5?), and *likelihood functions,* which capture the variation of the most likely parameter values given the observed data (e.g., what are the odds of getting 3 heads out of 4 tosses, using a coin with a probability of heads equal to 0.1, 0.2 ,0.3, 0.4, and so on). Although within the realm of a single model, the former is a depiction of multifinality (a parameter value generating different outcomes) and the latter of equifinality (multiple parameter values generating the same outcome), and in both cases variations (of the outcome or the parameters) are formally quantified in probabilistic terms. How does inferential statistics then deal with equifinality and multifinality? Either by aiming at the rejection of a particular model (i.e., the frequentist null-hypothesis testing approach) or by comparing multiple models using information criteria. The latter approach in particular can be used to directly test competing hypotheses against each other, providing the possibility of formally comparing alternative explanations for an observed pattern (Rubio-Campillo et al. 2017), potentially drawn from distinct bodies of archaeological theories (Eve and Crema 2014). If the objective of our modelling enterprise involves some comparison with the empirical record, why are we not adopting these, arguably better, inferential tools[1.]?

There are at least three sets of reasons. First, pure *hypothesis-testing* models in archaeological simulations are not common. As mentioned earlier, the great majority of empirically ground, realistic simulations are simultaneously also *theory-building* devices. Models are constructed on the basis of a given historical-geographic context, but a substantial effort is spent on exploring the parameter space to evaluate the consequences of the embedded assumptions. One reason why we do not observe pure *hypothesis-testing* models is that ideal null hypotheses or established alternative explanations that are readily formalised in the literature are rarely available. In other words, model-based archaeology is still in its early stages, whereas *theory building* is still central and there are no "off-the shelf" models ready to be tested against data. It is no coincidence that the few notable exceptions where a simulation-based, *generative* statistical inference has been used are those with a well- and long-established body of formal models already available. For example, Crema and colleagues (2016) have recently reexamined the Neolithic pottery assemblage from Merzbach Valley in Germany, comparing outputs of a simulation model with different modes of social learning (unbiased, conformist, and anti-conformist) derived from cultural evolutionary theory, whilst Por i and Nikoli (2016) studied the demographic changes at the Mesolithic site of Lepenski Vir in Serbia, using long-established population growth models.

The small number of parameters and the comparatively high levels of abstraction in these and other examples illustrate the second reason why the adoption of a statistical inference for the analysis of computer simulations is difficult. The great majority of these models cannot be analytically "solved," so expected outputs of a given parameter value can be obtained only through a simulation run. To obtain the approximate equivalent of a *probability distribution*, we would thus need to rerun a model with the same parameter settings many times, and, crucially, to obtain something comparable to a *likelihood function*, we would need to do this for every possible combination of model parameters and record how often, and under which circumstances, the output perfectly matches the observed data. The number of simulations required to achieve such a task becomes almost immediately intractable with the increasing number of parameters. However, an approximate solution based on some measure of distance to the observation (rather than a perfect match) can drastically reduce the computational requirements, making the combination of statistical inference and simulation modelling feasible. One of the most promising approaches in this direction is *approximate Bayesian computation* (ABC), a computational method that enables probabilistic estimates of parameter values as well as comparisons of different models against the same observed data. Still, such an approach is possible only by using modern computer technology, as the number of required simulations is in the order of magnitude of millions—well above the typical number of runs observed in archaeological simulations.

Third, the choice of what exactly we are trying to "fit" can severely limit model design and even bias the inferential process. *Summary statistics* that numerically describe complex phenomena (e.g., diversity indices) are very often *insufficient*, i.e., they entail a loss of information compared to the full dataset and can introduce further levels of equifinality. This is worsened by the fact that observed archaeological data are also profoundly affected by postdepositional events, sampling strategies, and loss of crucial information (e.g., via time-averaging) that are rarely reproduced within simulation models despite potentially shaping a large component of the observed pattern.

A Future for Hypothesis-Testing Models?

Is there a future for *pure* hypothesis-testing models in archaeology? Increased computational resources and a wider development of formal models within archaeology can certainly benefit the use of approaches similar to ABC. This seems to be the case for fields with a longer tradition and a greater role of formal models such as population genetics. Similarly, in ecology, attempts have been made to formalise the comparison between the output agent-based simulations and empirical data using ABC (van der Vaart et al. 2016), or to devise alternative model selection criteria (Piou et al. 2009). The ABC approach itself is also benefiting from continuous methodological development and refinement by the statistical community, showcasing how the combination of a consolidated inferential paradigm with the flexibility of formal simulation-based modelling is both an attractive and promising cross-disciplinary research agenda. Within archaeology, different bodies of theory will inevitably have different stances towards this approach, and the temptation to exclusively rely on borrowed models from adjacent fields or to limit the inferential exercise to tractable problems, data, and hypotheses will be the greatest limit of its wider application. The lack of a unified body of theory in the social sciences will on the one hand impede the spread of reusable models, but at the same time will offer a unique opportunity to contrast a wider range of alternative explanations directly against data. Whether the latter will be achieved, or even sought, remains an open question.

Acknowledgments

This article benefited from discussions with various colleagues, in particular Anne Kandler, Mark Lake, and Xavier Rubio-Campillo, to whom I am extremely grateful.

References

Costopoulos, André
2015 How Did Sugarscape Become a Whole Society Model? In *Agent-based Modeling and Simulation in Archaeology (Advances in*

Geographic Information Science), edited by Gabriel Wurzer, Kerstin Kowarik, and Hans Reschreiter, pp. 259–269. Springer International Publishing, Cham, Switzerland.

Crema, Enrico R., Anne Kandler, and Stephen Shennan
2016 Revealing Patterns of Cultural Transmission from Frequency Data: Equilibrium and Non-equilibrium Assumptions. *Scientific Reports* 6:39122. https://doi.org/10.1038/srep39122

Crema, Enrico R., Andrew Bevan, and Stephen Shennan
2017 Spatio-temporal Approaches to Archaeological Radiocarbon dates. *Journal of Archaeological Science* 87:1–9.

Dronamraju, Krishna
2011 *Haldane, Mayr, and Beanbag Genetics*. Oxford University Press, New York.

Eve, Stuart J., and Enrico R. Crema
2014 A House with a View? Multi-model Inference, Visibility Fields, and Point Process Analysis of a Bronze Age Settlement on Leskernick Hill (Cornwall, UK). *Journal of Archaeological Science* 43:267–277.

Kohler, Timothy A., R. Kyle Bocinsky, Denton Cockburn, Stefani A. Crabtree, Mark D. Varien, Kenneth E. Kolm, Schaun Smith, Scott G. Ortman, and Ziad Kobti
2012 Modelling Prehispanic Pueblo Societies in Their Ecosystems. *Ecological Modelling* 241:30–41.

Kohler, Timothy A.
2015 Review of *Agent-based Modeling and Simulation in Archaeology (Advances in Geographic Information Science)*. Electronic document, http://jasss.soc.surrey.ac.uk/18/2/reviews/2.html, accessed November 12, 2017.

Lake, M. W.
2014 Trends in Archaeological Simulation. *Journal of Archaeological Method and Theory* 21:258–287. https://doi.org/10.1007/s10816-013-9188-1

Piou, Cyril, Uta Berger, and Volker Grimm
2009 Proposing an Information Criterion for Individual-Based Models Developed in a Pattern-Oriented Modelling Framework. *Ecological Modelling* 220:1957–1967.

Por i , Marko, and Mladen Nikoli
2016 The Approximate Bayesian Computation Approach to Reconstructing Population Dynamics and Size from Settlement Data: Demography of the Mesolithic-Neolithic Transition at Lepenski Vir. *Archaeological and Anthropological Sciences* 8:169–186. https://doi.org/10.1007/s12520-014-0223-2

Premo, L. S.
2010 Equifinality and Explanation: The Role of Agent-Based Modeling in Postpositivist Archaeology. In *Simulating Change: Archaeology into the Twenty-First Century*, edited by Andre Costopoulos and Mark W. Lake, pp. 28–37. University of Utah Press, Salt Lake City.
2014 Cultural Transmission and Diversity in Time-Averaged Assemblages. *Current Anthropology* 55:105–114.

Rubio-Campillo, Xavier, María Coto-Sarmiento, Jordi Pérez-Gonzalez, and José Remesal Rodríguez
2017 Bayesian Analysis and Free Market Trade within the Roman Empire. *Antiquity* 91: 1241–1252.

Van der Vaart, Elske, Alice S. A.. Johnston, and Richard M. Sibly
2016 Predicting How Many Animals Will Be Where: How to Build, Calibrate and Evaluate Individual-Based Models. *Ecological Modelling* 326:113–123. https://doi.org/10.1016/j.ecolmodel.2015.08.012

Wason, P. C.
1960 On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology* 12:129–140. https://doi.org/10.1080/17470216008416717

Notes

[1] Although it can be argued that seeking to identify a mismatch between simulation and data shares some similarity to the null-hypothesis testing approach.