Synthese (2013) 190:203–218
DOI 10.1007/s11229-012-0142-7

The perils of tweaking: how to use macrodata to set parameters in complex simulation models

Brian Epstein · Patrick Forber

Received: 27 February 2012 / Accepted: 27 February 2012 / Published online: 4 July 2012

Abstract When can macroscopic data about a system be used to set parameters in a microfoundational simulation? We examine the epistemic viability of tweaking parameter values to generate a better fit between the outcome of a simulation and the available observational data. We restrict our focus to microfoundational simulations—those simulations that attempt to replicate the macrobehavior of a target system by modeling interactions between microentities. We argue that tweaking can be effective but that there are two central risks. First, tweaking risks overfitting the simulation to the data and thus compromising predictive accuracy; and second, it risks compromising the microfoundationality of the simulation. We evaluate standard responses to tweaking and propose strategies to guard against these risks.

Keywords Microfoundations · Overfitting · Tuning · Calibration · Simulation · Estimation

Many computer simulations are intended by their designers to be "bottom-up" models of macroscopic phenomena. Examples are found across the sciences:

- The flow of plasmas, modeled by simulating the interactions among large numbers of fast-moving particles,
- the folding of proteins, by simulating the interactions among amino acids,
- the dynamics of ecosystems, by simulating the interactions of individual predators and prey,

B. Epstein and P. Forber have contributed equally to this work.

B. Epstein (✉) · P. Forber
Philosophy Department, Tufts University, Miner Hall, 14 Upper Campus Rd.,
Medford, MA 02155, USA
e-mail: brian.epstein@tufts.edu

- climate change, by simulating interacting volumes of atmosphere, ocean, and land,
- the fluctuations of stock markets, by simulating the interactions among traders,
- traffic jams, by simulating the movements of cars over road-segments,
- and so on.

In general, such simulations start with models of interacting individual entities that in the aggregate produce the macroscopic phenomenon of interest. (Call these the *microfoundational* entities.) Then they calculate the results of large numbers of interactions over time, observing as macroscopic regularities develop.

In many of these simulations, much is already known about the macrophenomena of interest. A traffic modeler may have detailed data about the conditions under which traffic jams occur, and the protein folder knows the structure of a great number of actual proteins. Thus simulations—like any other models—are built iteratively. The modeler does not just write a simulation once, set it running, and read off the results. Anyone who has written a simulation knows that the initial runs are invariably hopeless. Even if we start with a good understanding of the relevant microentities, it takes repeated tweaks and refinements for the results of the simulation to begin to approximate the macrodata.

This, however, raises an immediate problem. The aim of a bottom-up model is to generate the macrophenomena from models of interacting microentities. "Tuning" a model, or tweaking the microparameters whenever we get results we do not like, can amount to slapping an ad hoc bandaid on a broken model, insulating the model from any empirical risk. If we hope to simulate the macrophenoma using micro-interactions in an empirically informative way, we put our thumbs on the scale if we smuggle in the macrodata whenever we get results we do not like.

Is it acceptable to use macrodata in setting microparameters, and if so, when? This is part of a larger and under-theorized issue: the process of simulation model improvement. Here we focus on tweaking in particular, just one type of improvement for bottom-up simulation models.

In this paper, we argue that some of the most prevalent guidelines are overstated, and that iterative tweaking of models is allowable in many circumstances. But we also raise other problems for tweaking that have largely escaped notice. There are, we argue, two distinct risks involved in the use of macrodata to set microparameters.

One risk is that tweaking compromises the simulation's *predictive accuracy*. This is tied to traditional problems of model selection and validation, problems that become particularly thorny when applied to computer simulations. Simulators often worry about "tuning" models to fit the data, and often discuss these worries in connection with the procedures for "calibrating" and "validating" simulations. Looking at tweaking from the lens of statistical inference, it can be understood as a coarse-grained strategy for maximum likelihood estimates for parameter values. As explored in the model selection literature, complex models with a large number of adjustable parameters run a high risk of *overfitting* the model to the data. So tweaking may produce an ad hoc simulation that merely accommodates the data. But it need not always fall into this trap. This is true whether we use existing data to fit our models or seek new data to do this.

A second risk is not commonly discussed. This is risk to the simulation's *microfoundationality*; that is, to whether the simulation succeeds at being bottom-up. Even when the use of macrodata in setting microparameters improves predictive accuracy, the risk is that this victory may be achieved only by covertly smuggling the macrodata into the simulation. If part of the aim of certain simulations is to have the macrophenomena emerge bottom-up from their microfoundations, this risk provides a different reason for suspicion of tweaking.

We argue, however, that addressing these two risks does not mean foreswearing the use of macrodata in tweaking models. We propose strategies for addressing each of the risks. Unfortunately, there are tradeoffs among these strategies. For instance, overfitting risks may be addressed in part by restricting the number of parameters by, for instance, making individual microentities homogeneous. But, as we will see, homogeneity among the elements of a simulation increases the risk of smuggling.

1 Good versus bad tweaking

It is easy to describe examples in which the use of macrodata to set microparameters is illicit. Consider, as a simple case, an agent-based simulation of the behavior of a large school of fish.

1.1 Tuning the herring

A school of herring may consist of tens of millions of individual fish, each reacting to its environment and interacting with its neighbors. As a whole, schools exhibit large-scale patterns. For instance, on encountering a predator such as a killer whale, a school may split in two, leaving a wide berth for the predator, and rejoining once the whale has passed through the school. These macroproperties of entire schools, measured using multibeam sonar and echosounders (Axelsen et al. 2001; Nottestad and Axelsen 1999), present high-quality data of the behaviors of herring schools on encountering predators. They show that schools have a number of anti-predation maneuvers, including splitting in two, turning sharply in one direction or another, creating a vacuole around the predator, and so on.

Suppose we build a simple agent-based model of herring behavior, intending to simulate these macroscopic anti-predation maneuvers. In the first iteration of the simulation, we begin with some data we possess about the behaviors of individual herring. We observe that individual herring can perceive a threat in two ways: they can perceive it directly, or they can perceive a disturbance in their environment, such as a threat reaction by other herring in their neighborhood. Our information about individual herring also indicates that in response to a threat, an individual herring will reverse direction and swim away as quickly as possible. So the basic elements of our simulation are herring-agents with these behaviors, along with predator-agents, who swim in straight lines.

When we run the simulation, we find a certain kind of behavior predicted in the face of a threat. As a predator approaches, each of the herring-agents reverses direction, a reversal which cascades through the school. Soon, the school is swimming in the opposite direction it had been going.

This first iteration of a model is plausible enough. Unfortunately, it does not match the macroscopic data we have of herring behavior. On encountering a threat, the actual data shows that the school does not reverse direction, but splits in two. Assigning these simple characteristics to the agents, the simulation predicts the wrong macrobehavior for the school. So we revise the model.

In the second iteration, we use the macrodata that the school splits in two on encountering a threat to tweak the properties of individual fish. To capture the observed macrobehavior, we assign the agents two different personalities. We make the 5 million fish in the left half of the school "lefties," and the 5 million in the right half "righties." The lefties are disposed, on perceiving a threat, to take a sharp left turn for a while, before returning to their original direction. The righties are disposed to turn right.

When the simulation is run, this tweak works. On encountering the predator the school splits in half, just as the macrodata predicted. But this tweak is an illicit one. All we really did, by assigning the fish different personalities in the right and left halves of the school, was to bake the macro-outcome into the micro-assumptions. Despite the fact that the simulation produces the correct macrophenomenon, it cannot be regarded as a good "bottom-up" simulation.

While this example is an exaggerated one, it is worrisome, because the "tweaking" method will be familiar to anyone who has ever written a simulation. Simulations invariably begin as highly simplified idealizations, with the researcher hypothesizing which microproperties are the crucial ones for generating the macrophenomena, and what the appropriate parameter settings should be for modeling those microproperties. Only through a process of iterative testing and revision—sometimes going through generations of researchers—do the results of simulations ever manage to approach the macrophenomena.

1.2 Overreacting

Given the perils of tweaking, some theorists adopt strict rules for the use of macrodata, or prohibit its use altogether. Quite generally, theorists divide the use of macrodata into two categories: the use of macrodata in model evaluation is put into the umbrella of model "validation," and the use of macrodata in setting parameters is "calibration."

Randall and Wielicki (1997), for instance, take a hard line, insisting that parameters can and should "be set once and for all before a model is run" (Randall and Wielicki 1997, p. 405):

> "Good empiricism" is…applied before the model is run. The empirical parameters of a model should be measured and then set, on the basis of these measurements, before the model is used to make a prediction. The parameters should not be adjusted a posteriori to improve the agreement between the model predictions and other data. Tuning consists of … adjusting parameters after a model is run to improve the agreement between the model results and data. Tuning is bad empiricism. Calibration is bad empiricism with a bag over its head (p. 404).

While they admit that in practice scientists are often forced to resort to tuning, they claim that this is an inferior strategy; once sufficient understanding of the underlying processes is achieved "then there is no excuse for continued tuning" (Randall and Wielicki 1997, p. 404).

The "hard line" reaction finds its way into the actual practice of simulation design. The following, for instance, is a fairly typical description of the methods employed for constructing a simulation. Here the model is a cellular-automaton simulation of vascularization:

> In total, the model drew upon published independent experimental data obtained from in vitro and in vivo experimental studies to govern 48 free parameters which represented different aspects of the remodeling process, including cell proliferation rates and cell migration rates… All of the rules, equations, and parameters that governed the simulated tissue environments and the cellular behaviors were derived a priori and unequivocally from the literature or from independent experimental observation. They were not altered by the CA simulation or any results generated by the simulation (Pierce et al. 2004).

The experimenters take pains to insist that their microparameters were set "once and for all" before the simulation was run.[1]

Similar reasoning underlies a more pessimistic reaction to validation. Oreskes et al. (1994) suggest that calibration generates illusory confidence in a simulation, and we must set our sights lower. Likewise, Kleindorfer et al. (1998) suggest that validation be replaced with establishing social credibility. Rykiel (1996) argues that validation criteria vary across contexts, and that primary use of validation is not confirmation, but a kind of pragmatic establishment of model credibility.

But these are overreactions. Not only do modelers in fact develop models iteratively, using macrodata to tweak microparameters, but it is as easy to come up with cases where such tweaking is clearly fruitful.

Consider, for example, the van der Waals equation for gases.[2] The van der Waals equation, $(\mathrm{P} + \mathrm{n}^2\mathrm{a}/\mathrm{V}^2)(\mathrm{V} - \mathrm{nb}) = \mathrm{nRT}$, introduces two parameters: the volume excluded by a mole of particles, and the attraction between the particles. Near the critical point, this equation is a substantial improvement over Boyle's equation.

Van der Waals introduced his revision to the ideal gas law in part in response to macro-measurements of departures of the behavior of certain gases from Boyle's law under various conditions (van der Waals 1910). Though it is a macroscopic equation of state, however, his equation also provides a straightforward route for estimating the properties of individual particles. In the nineteenth century, it was of course impossible to measure the excluded volumes and interaction energies of particles directly. Calculations of these parameters for different gases were thus made on the basis of experimental macro-measurements. Over time, other ways of estimating these parameters were developed. For instance, the mechanical measurements were triangulated with other macroproperties, such as polarizability and molar refractability. Still later, techniques were developed for estimating the parameters more directly, such as using X-ray crystallography to measure atomic spacing. All these later measurements together, of

[1] Notice also the massive complexity of the simulation, with 48 free parameters per cell. As we will discuss below, such complexity introduces substantial risks.

[2] Our thanks to an anonymous reviewer of an earlier paper for suggesting this example.

course, refined the values of the radii and interaction energies beyond those that could be inferred from the mechanical properties of gases alone.[3]

This successful estimation of radii and interaction energies can be seen as an instance of iterative microparameter-tweaking. Suppose we began with a crude guess at the interaction energies and radii for a particular gas, and ran a simulation on that basis. From the mismatch of the output of that simulation with the empirical macrodata, we changed the parameters slightly, and re-ran the simulation. And so on, until we arrived at a good match. This numerical method, performed properly, would arrive at the same result as the analytic method of solving the equations for their parameters, given empirical measurements of P, V, and T.

1.3 The moderate response: also an overreaction

In response to this sort of case, and to the actual practice of modelers, a more moderate stance is often taken. The proposal is that a sharp separation be made between the data used for "calibrating" a model, and the data used for testing the model. Müller and von Storch (2004), for instance, regard calibration and validation as important steps in the construction of models, but argue that a Chinese Wall be erected between the data used for calibrating a model and that used for validating a model. They also insist that calibration be involved only in tweaking parameters, and should not be used for any other purpose, such as modifying functional forms or changing underlying mechanisms.

This proposal, however, also misses the mark. Neither of the two prohibitions is justified. First, if the use of macrodata is acceptable for tweaking parameters, it may also be acceptable for modifying functional forms or mechanisms. Consider, for instance, the argument given by Bearman et al. (2004) for a hypothesis about relationship taboos among American adolescents. They observe that the network of romantic interactions in a particular high school closely resembles a chainlike spanning tree, with nearly a complete absence of short cycles. Using an agent-based model, they build evidence for hypotheses about the rules governing the interactions which generate that structure. In other words, they infer and modify the mechanisms of their simulation to match the observed macrostructure.

Second, the "Chinese Wall" restriction is also too strong. Bearman et al., for instance, violate it, using macrodata both for the formulation and validation of their models. In more general contexts, a number of people have argued in recent years against the once-pervasive view that scientific hypotheses must be tested using only "novel evidence" (e.g., Glymour 1980; Worrall 2002; Hitchcock and Sober 2004; Mayo 2008). In connection with simulation in particular, we discuss this point in more detail in the next section.

In the next two sections, we address predictive accuracy and microfoundationality, and in the final section we apply these issues to difficult cases. We argue that both are risks, and the use of macrodata in tweaking microparameters involves taking such risks. Moreover, we argue that there can be tradeoffs between the risk of compromising

[3] Cf. Bondi (1964).

predictive accuracy and the risk of compromising microfoundationality. While macrodata should therefore be used judiciously, its proper use can enhance bottom-up models.

2 From the perspective of statistical inference

Comparing simulation tweaking to other kinds of inferences from data can help determine whether and when tweaking is good or bad. The consensus view is that tweaking is (almost) always a bad thing. One source of this skepticism is the intuition that accommodation is epistemically inferior to novel prediction, and that tweaking is an extreme form of accommodation. On this view, tweaking involves a kind of guided manipulation of simulation parameters to generate better fit between the simulation outcomes and existing macrodata, and this manipulation compromises the accuracy of the simulation. A superior method of testing simulations involves somehow estimating or inferring parameter values independently, and then assessing fit between the simulation outcomes and *new* macrodata. While there is something to this intuition worth rescuing, it is often false (Hitchcock and Sober 2004). Investigating when and why it is false is informative, and is best done by looking at simulation tweaking through the lens of statistical inference.

Maximum likelihood estimation (MLE) provides a one way to estimate the value of parameters that generate the best fit between model and data. MLE is common, effective, and epistemically viable. Of course, successful MLE requires a number of formal assumptions. In particular, basic MLE assumes that the data are the product of a random variable that obeys a known probability model but has an unknown exact distribution. The exact distribution is specified by the parameter values, which are being estimated from the data. Consider a simple example: tossing a coin with an unknown bias. We assume that coin tossing obeys a binomial model, and begin tossing to estimate a single parameter, the probability of coming up heads (P). Say we get heads 63 out of 100 tosses. The maximum likelihood estimate for the parameter is $P = 0.63$. Based on the sample size we can estimate a confidence interval, and such intervals narrow as sample size is increased. Notice that the background probability model plays a crucial role in constraining the inference problem. Also notice that MLE uses *all* the data, and more data increases the precision of the estimate.[4]

The inferential problem of estimating parameters for simulations of herring, traffic, or economies introduces a degree of complexity that makes the formal assumptions of statistical MLE problematic. The simulations used to model the behavior of these systems are far from the well-behaved probability models assumed by MLE, and often there is no clear way to get traction on what the family of distributions should be for estimating the numerous simulation parameters from the data. Estimating the bias of a coin assuming a binomial model is a far cry from tweaking herring evasion parameters to fit the macrolevel schooling patterns we actually see. Yet, following a suggestion made by Dawkins et al. (2001, pp. 3663–3664), there is a clear analogy

[4] The formal justification for MLE is well established and we gloss those details here; see, e.g., Sokal and Rohlf (1994) for biological applications, Royall (1997) on likelihood functions, or Burnham and Anderson (2002) for the MLE and model selection.

between tweaking and statistical techniques for estimation, especially MLE—both procedures involve manipulating parameter values to increase the fit between what counts as the model (the simulation or the probability model) and the observations. The difference is that MLE, due to the strong formal assumptions, can use formal methods to infer the maximum likelihood estimate from the data whereas tweaking a simulation involves a step-by-step process of altering parameter values and evaluating fit between simulation outcomes and macrodata. This step-by-step process is necessary because of the increased complexity of the simulations. Thus, we can fruitfully view tweaking (tuning, calibrating) as the coarse-grained strategy for MLE with very complex models.

2.1 The risk of overfitting

This analogy casts doubt on the more extreme prohibitions against tweaking. Why think tweaking is such a bad thing? If it counts as coarse-grained MLE then tweaking represents an innovative way to estimate parameter values for very complex models, not an undue epistemic risk. In fact, the analogy does show that a general prohibition against parameter tweaking is misguided. Fitting a model to data is an excellent way to produce an accurate model of some system. However, the general prohibition responds to a genuine epistemic worry, a worry also brought about by the tweaking-MLE analogy: *the risk of overfitting the model to the data*.

The risk of overfitting faces any MLE problem for models with multiple parameters. It is commonly discussed in terms of the curve-fitting problem. Given enough parameters, we can generate a curve that will fit all the data points in a scatter plot exactly. Such a curve may have the best fit, but it provides useless predictions. If our model has too many parameters then MLE will overfit the model, compromising *predictive accuracy*. Simulations are massively complex models with many parameters, and therefore face a severely increased risk of overfitting.

Statisticians have developed tools to cope with the risk of overfitting for MLE: model selection statistics.[5] The formal details are complex but the core idea is simple. Model selection statistics balance fit to data against the complexity of a model in order to rigorously compare the overall predictive accuracy of models of differing complexity (Burnham and Anderson 2002). The number of adjustable parameters determines the complexity of the model. Increasing the number of parameters, and thus increasing the complexity of the fitted model, incurs a penalty. Statistical model selection identifies the model that provides the best tradeoff between fit and complexity in order to maximize expected predictive accuracy.

Statistical model selection cannot be applied without modification to assessing the predictive accuracy of complicated simulations, for simulation parameter tweaking is not the same as fitting a polynomial to a set of data. Yet, continuing the analogy between tweaking and MLE, model selection provides an important insight. Suppose we want to fit a microfoundational simulation to some macrodata. Consider one

[5] In the philosophical literature AIC has received the most attention. See, for example, the argument by Forster and Sober (1994) that this model selection framework shows how simplicity matters in science.

simulation with a large number of parameters. We can make this into a model selection problem by sorting the parameters into two groups: adjustable and fixed parameters. Start by treating all parameters as fixed. Then begin by tweaking one, then two, and so on. Any of the parameters are fair game for tweaking, but we decide how many parameters should count as adjustable by deploying model selection considerations. The increased fit generated by tweaking another parameter must counterbalance the increase in complexity by allowing more parameters to be tweaked. Any tweaking of simulation parameters risks overfitting the simulation to the data, a risk that increases with the number of parameter we allow to be adjusted. Tweaking *all* the parameters of a very complex simulation to increase fit will almost certainly compromise the predictive accuracy of the simulation. But tweaking *some* parameters to increase fit is simply good scientific practice, for lack of fit to data compromises predictive accuracy just as well. So tweaking can be legitimate insofar as the risk of overfitting can be mitigated.

2.2 When is tweaking viable?

In sharp contrast to the statistician, the consensus among scientists who work with very complex simulations is that tweaking is an epistemically inferior way to determine parameter values (see, e.g., Randall and Wielicki 1997). While these scientists often see tweaking in just the way described here as analogous to MLE, they claim that it should be avoided because it compromises a scientist's ability to test their hypotheses about how the simulations represent the target systems. Tweaking insulates the simulation from disconfirmation.[6] This is an endorsement of the tenacious intuition that accommodating a theory to existing data is problematic, and that novel predictions or tests provide a better source of epistemic support.

Hitchcock and Sober (2004) use the model selection framework to explore accommodation versus prediction. Let us briefly discuss their example. They contrast two scientists, Penny the predictor and Annie the accommodator, who are trying to fit a polynomial function C to the data D. Penny uses a subset of D to fit her curve C_p, then uses it to accurately predict the rest of the data in D, whereas Annie fits her curve C_a to the entire data set D. Is C_p better than C_a? There is no general answer because it depends on how C_p and C_a are inferred; in some cases C_p will be better, in some cases C_a will be better, and in other cases the choice will be irrelevant (Hitchcock and Sober 2004, pp. 15–21). One case is particularly relevant to the views in the simulation literature. If both Penny and Annie guard against overfitting, so both use model selection to select the curve that maximizes the tradeoff between fit and complexity, then *Annie's* C_a provides the best curve, for she uses all the data. Since Penny uses only a subset of the data then C_p can't have a higher predictive accuracy than C_a (Hitchcock and Sober 2004, p. 17). Just as there is no general reason to prefer curves that make novel predictions to curves fitted using the entire data set, we should expect no general solution to the question of whether tweaking is legitimate.

[6] "The problem with tuning is that it artificially prevents a model from producing a bad result" (Randall and Wielicki 1997, p. 404). Recall that they define tuning as "the practice of adjusting parameters after a model is run, to improve the agreement between the model results and data" (Randall and Wielicki 1997, p. 405).

This moral reveals the flaw in the moderate solution to tweaking. Recall that the moderate solution involves partitioning the data into two sets, one used strictly for calibrating the simulation (i.e., tweaking the parameters to fit the simulation model to the data), and the other set used to test the fitted simulation model. While this heuristic does help guard against overfitting, it does so at the cost of underutilizing the data. Other strategies, such as model selection statistics, can guard against overfitting while utilizing all the data, as Hitchcock and Sober make clear. The cost of the "Chinese Wall" heuristic for calibration and testing decreases as the total amount of data increases. In machine learning a common technique uses algorithms to train models on a subset of data (called the training set), then measure the success of models applied to new data sets, but such techniques are only effective with very large data sets (Bishop 2006). Thus, it can be effective if we have sufficient data, but in all other cases it provides only a crude response to the problem of overfitting that makes inefficient use of available data.

The same considerations also answer the misconceived skepticism towards calibration found in Oreskes et al. (1994). They argue that such procedures often involve further tweaking of the simulation when testing it against the second data set, and so do not provide any test or support of the simulation model—the result is that "the so-called verification is a failure" (Oreskes et al. 1994, p. 643). Yet so long as we take sufficient steps to guard against overfitting, using all the data to tweak or fit a simulation does not result in failure, but is a step towards providing a predictively accurate simulation.

Tweaking, like accommodation, is legitimate provided we guard sufficiently against the risk of overfitting. And tweaking utilizes the data to maximum effect. But how do we know we are not guilty of overfitting? The answer will depend crucially on the details. It is often observed that independent information about the underlying processes is beneficial for simulation construction. Our analysis clarifies one of the benefits: using independent information to constrain parameter values will always help mitigate the risk of overfitting. In our herring simulation, having independent information on the behavioral repertoire of individual herring will provide constraints to tweaking that guard against overfitting our simulation to the macrodata on predator evasion. Integrating background information into the simulation is crucial for this reason.

Notice that the discussion in this section has focused on predictive accuracy, a feature of the simulation model that compares the *output* of the model to the data. Tweaking should be avoided if it overfits the model to the data. But we may have independent concerns about the effects of tweaking microparameters to increase the fit to macrodata. Such tweaking may compromise the *interpretation* of a putatively bottom-up simulation. We now turn to this issue.

3 Microfoundations

As our preposterous herring-simulation shows, blindly tweaking parameters can do violence not only to the predictive accuracy of a simulation, but also to whether it is "bottom-up." In the second iteration, we tuned the psychologies of the individual herring to match the macrodata, so that half the agents were artificially set to be lefties and half to be righties. This was already a bad move, just from the perspective of maximizing predictive accuracy. If, for instance, the lefties and righties change their relative positions, the macrobehavior collapses. But it also has a different problem. Tuning the model in this way smuggles the macroproperties of the school into the microproperties of the individual herring. The modeled "herring" are little more than encoded fragments of the school as a whole, rather than representations of individual actors.

Some theorists may not consider this "smuggling" to be a defect. The instrumentalist, for instance, purports to be unconcerned about anything but predictive accuracy. Inasmuch as predictive accuracy is the goal, even a strictly macroscopic model may do a better job than any extant microfoundational model at fitting the data. A simple Taylor Rule, for instance, may do a better job modeling the relation between inflation and interest rates than a model of individual agents. Likewise, the van der Waals equation or some other macroscopic state equation may do a better job modeling the relation between temperature and pressure in a gas than a model in statistical mechanics.

Predictive accuracy, however, is but one factor in model development and selection. There are a number of reasons one might want a model or simulation to be genuinely microfoundational, even at the expense of fit. Among the reasons are:

Extrapolation: Despite the better fit of a macromodel to the data, there may be reason to be more confident in a microfounded model than a macromodel, when applied to new circumstances. In economics, for instance, Robert Lucas's (1976) critique of structural macroeconomic models has led many people to favor microfoundational models. Lucas argues that macroeconomic models tend to fail when public policy changes, since the expectations and hence the choices of individual actors change in response to changes in policy. In place of structural models, he argued that macroeconomic models should be built on "deep parameters," such as the tastes and technology of individual actors, which he claims are policy invariant.

Descendent simulations: Simulations are not one-shot affairs. A microfoundational model may be preferred not because we expect that early models will be predictively superior to macromodels, but because we have reason to suspect that they lay better foundations for the long-term development of models.

Modularity: Microfoundational simulations may also be more conducive to modular construction. There are many reasons modularity may be desirable. Models do not stand on their own, but rather parts of models get incorporated into other models. There are advantages to conforming to a common ontology, reusing model-parts that others have built, and repurposing parts of a simulation in other simulations. There are also practical considerations, including scalability, design and engineering considerations, and so on. These often take strong precedence over the optimization of a particular model.

Explanatory value: Microfoundational models are widely regarded as explanatorily superior to macromodels. In general, this is not justified, as critiques of mid-century views on explanation have made clear. However, more considered perspectives still leave room for the explanatory privilege of microfoundational models, at least for certain purposes. Kincaid (1986), for instance, gives a measured defense of "individualistic" explanations in the social sciences. Similar motivations seem to drive recent arguments for mechanism-based explanations, as in Machamer et al. (2000).

Insights into microentities: It is often a central goal of models to gain insights into the properties of the microentities generating a macrophenomenon, apart from modeling the macrophenomenon. Equally, in simulations we often seek to gain understanding of patterns of aggregation. It is not only the macroresults that are of interest, nor the microentities that interact with one another, but how the macroresults are produced from the interactions.

3.1 How can tweaking take a toll on microfoundationality?

It is a puzzle how tweaking the parameters of a simulation could possibly have an effect on whether the simulation is microfoundational. The following seems like a plausible picture of simulations. A modeler chooses a set of entities to represent with elements or components of a simulation. Each of those entities is structurally described in the components of the simulation, assigning them attributes, behaviors, and so on. Initial conditions are set, and then the simulation is set off and running. For the simulation to be microfoundational, then, seems to be a matter of modeling a macrophenomenon by choosing to represent entities from some base on which the macrophenomenon supervenes (where the base is understood to exclude macroentities).

In this picture, what we might call the "modeled ontology" of a simulation is fixed by the modeler independently of *how* the microentities are represented. Changes to the representation, and in particular, tweaks to attributes or parameters, have no effect on which entities are modeled. Tweaking parameters would seem unable to make any difference at all in which entities the components of a simulation represent. And hence would seem unable to make a difference as to whether the simulation is microfoundational.

Our tweaked herring simulation, however, was supposed to show at least in principle that tweaking parameters may indeed have an effect on whether a simulation is microfoundational. The suggestion was that in the course of tweaking the parameters of the model, the macroentities were "smuggled into" the simulation.

The idea is that whether or not the modeler intends it, macroentities may nonetheless sneak into what the simulation represents. In the herring case, the modeler intends to represent herring psychology, dividing the dispositions of the herring into "lefties" and "righties." But despite what the modeler wants, there may be factors outside those intentions that determine what the components of the simulation represent. The fact that the herring-components were tuned, the way they were, to match the macrodata, makes those components represent something other than the intended interpretation of the modeler.

The herring case is a caricatured one. But there are real-world cases in which this is a real issue, of importance to modeling methodology.

Consider, for instance, "representative agent" models in economics. These models seek to balance analytic tractability with a desire to model macroeconomic phenomena in terms of their microfoundations. Because it is generally impossible to give closed-form solutions to models that include heterogeneous sets of individuals, these models treat the choices of a diverse set of agents as the choices of one "representative" individual whose choices are the same as the aggregate choices of the population as a whole.

Despite how widely these models are employed in contemporary macroeconomics, a number of people have criticized their basic assumptions. Kirman (1992, p. 118), for instance, points out that there is no reason to expect an aggregate of individual agents, even if those agents are maximizers and even if they have identical preferences, to act itself as a collective maximizer. This threatens to gut the ability of a representative agent model to guide policy.

Hoover (2006) has gone on to deny that these models should even be considered properly microfoundational at all. The representative agent, he argues, is similar to Quetelet's "average man," with 2.3 children and living in one property that is partly wholly rented, partly wholly owned, and partly on the streets. Neither the "average man" nor the representative agent is really an agent at all. Rather, "the representative agent is nothing else but an aggregate in microeconomic drag" (Hoover 2006, p. 146).

If Hoover's diagnosis is correct, a modeler may fully intend a model to be microfoundational and yet, depending on the basis for constructing the model, it may in fact be a macromodel "in drag." The same can be extended to simulations: despite the modeler's intention to refer to an element of the supervenience base, facts about how the simulation is constructed may trump those intentions.

Once we notice this problem, it becomes clear that it is potentially rampant in simulations. When dealing with putatively microfoundational simulations, it is easy to smuggle in macroproperties. Unfortunately, even some of the heuristics we might use to reduce the risk of overfitting end up increasing the threat that microfoundationality will be compromised.

Consider, for instance, TRANSIMS, a large-scale system for simulating the movement of people around a city, on foot and in cars and in public transportation (Barrett et al. 2000; Cetin et al. 2002; Eubank et al. 2004). Among the many parts of the model is a "traffic microsimulator," which simulates the flow of cars on the network of streets in a city, as they follow routes toward destinations, change lanes, and enter and exit parking spaces. The microsimulator is implemented as a cellular automaton, with the streets treated as long thin grids of car-sized cells, and cars jumping from cell to cell as they move through the streets.

In the simulation, the actual dynamics of movement from cell to cell are neglected. But suppose we wanted to add some of this texture to the TRANSIMS model. One option is to add a large number of parameters, representing many characteristics of the road. This, however, can seriously increase the risk of overfitting. A different response is to introduce a single parameter, representing the key factor or factors affecting cell-to-cell dynamics. The problem is that as we tweak such a factor, we risk finding ourselves in the shoes of the herring-modeler or of the representative-agent theorist, encoding the macrodata in "microfoundational drag." Having chosen a single parameter to modify, in order to mitigate the risk of overfitting, we risk losing control over what that parameter represents.

3.2 Guarding against smuggling

Treatments of the "tuning" of simulations, we argued above, tend to draw too simple a line between acceptable and unacceptable procedures for tweaking. Among the errors is that too much is often made of the distinction between accommodation and prediction. Here there is a different moral.

If we are to guard against overfitting, all the data is just data. But if we are to guard against the risk of compromising microfoundations, we need to treat microdata and macrodata somewhat differently. A different set of tests is called for, beyond those that evaluate predictive accuracy. Independent of the question of comparing simulations from the perspective of predictive accuracy, we should test for the effect of tweaking on what we might call the "representational integrity" of the simulation.

A good deal of testing for representational integrity is already performed, implicitly if not explicitly, by modelers in actual situations. One feature that can be straightforward to test is a simulation's modularity. This may be performed either by testing subparts of a simulation on their own, or else by swapping parts of a simulation into a different simulation of a macrophenomenon.

It is misleading to collapse all the microdata and macrodata together, without distinguishing the different testing goals they are to serve. It is easy to overlook the fact that we may have two different kinds of reasons for wanting a simulation to be microfoundational. One is to enhance the simulation's predictive success about macrophenomena.[7] But as we mentioned, there are many other reasons we might want a simulation to be microfoundational, apart from improving a simulation's predictive success about macrobehavior. These may be minor considerations or substantial ones, and we suggest it is better to consider them separately rather than simply incorporating micro-constraints into one data set.

Modularity is just one among a variety of features by which the "representational integrity" of a simulation can be tested. The notion of microfoundationality is a complicated one, and it is not clear that even if a simulation fails to be thoroughly modular, it invariably fails to be microfoundational. For instance, one might defend a representative agent simulation that fails to be modular as still being microfoundational, on the basis that it is a good idealization of agents relative to a particular macroeonomic problem (cf. Woodford 2006). Similarly, Satz and Ferejohn (1994) defend rational choice theory as an individualistic methodology in social theory despite the failure of rationality assumptions when applied on an individual basis.

Furthermore, there are tests for the "representational integrity" of a simulation that do not require modularity. For instance, in the sexual behavior paper, inferences are drawn about the relationship patterns among adolescents in light of the macrostructure of the graph of their interactions. The macrostructure of the graph provides good evidence that a behavioral constraint holds at the individual level. Such evidence may preempt the need for testing at the individual level.

In short, with the iterative tweaking of simulations, there is a real risk that simulation parameters become artifacts of the macrosystem, rather than genuinely

[7] Even this important role for microfoundationality sometimes goes unrecognized. For instance, the view of Friedman (1953) that the only goal of a science is its predictive success and that the "realism" of the assumptions involved in generating those predictions is unimportant, remains influential. Those who dispute this claim often follow the lines of Hausman (1992), which appeal to the contribution of "realistic assumptions" to a model's predictive success about macrobehavior. Interestingly, even what Hausman calls "wide predictive success" appears to be focused on the macro-predictions, rather than both the macro- and micro-predictions.

microfoundational. While it is difficult to characterize precisely when such smuggling of macrodata occurs, we may nonetheless test for the features normally exhibited by microfoundational models, and develop strategies for improving microfoundationality as an independent goal.

4 Conclusion

We have restricted our discussion to simulation and to tweaking, though to some extent the results can be extended to broader classes of models and to other processes for model-improvement.

Tweaking parameters is both common and useful. It is unnecessary to adhere to the strict code that many worried methodologists put forward, i.e., that macrodata must be isolated in one way or another from the iterated improvement of a simulation. That code is probably impossible to follow anyway, inasmuch as the macrodata affects model iterations, even at the level of the selection of basic structures to be included in a simulation. The iterative process of model development can stretch over the course of a researcher's career, and indeed passes on from researcher to researcher in the modeling community. It is a good thing that the strict codes are misguided, because to follow them would hamstring the practice of modeling. We come to praise tweaking, not to bury it.

That does not mean, however, that tweaking is risk free. One of the devilish problems of tweaking is that addressing the risks respectively involves conflicting recommendations. The simplifications that mitigate the risk of overfitting, such as treating a heterogeneous population of agents as one single representative agent, may be the very thing that compromises the microfoundationality of a simulation.

Acknowledgments We would like to thank Chris Barrett, Lisa Lloyd, Paul Humphreys, two anonymous reviewers, and audiences at the Models and Simulations 4 conference at the University of Toronto, and the Epistemology of Modeling and Simulation conference sponsored by MIDAS Center for Excellence and the Center for Philosophy of Science at the University of Pittsburgh for valuable feedback.

References

Axelsen, B. E., Anker-Nilssen, T., Fossum, P., Kvamme, C., Nøttestad, L. (2001). Pretty patterns but a simple strategy: Predator–prey interactions between juvenile herring and Atlantic puffins observed with multibeam sonar. *Canadian Journal of Zoology, 79*, 1586–1596.

Barrett, C. L., Beckman, R. J., Berkbigler, K. P., Eubank, S. G., Henson, K. M., Kubicek, D. A., et. al. (2000). TRANSIMS: Transportation analysis simulation system. Los Alamos Unlimited Release (LAUR) 00-1725.

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *The American Journal of Sociology, 110*(1), 44–91.

Bishop, C. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bondi, A. (1964). Van der Waals volumes and radii. *Journal of Physical Chemistry, 68*(3), 441–451.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodal inference: A practical information-theoretic approach*. New York: Springer.

Cetin, N., Nagel, K., Raney, B., & Voellmy, A. (2002). Large-scale multi-agent transportation simulations. *Computer Physics Communications, 147*, 559–564.

Dawkins, C., Srinivasan, T. N., & Whalley, J. (2001). Calibration. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3653–3701). Amsterdam: North-Holland.

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., & Toroczkai, Z., et al. (2004). Modeling disease outbreaks in realistic urban social networks. *Nature, 429*, 180–184.

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*, 1–35.

Friedman, M. (1953). The methodology of positive economics. In M. Friedman (Ed.), *Essays in positive economics* (pp. 3–43). Chicago: University of Chicago Press.

Glymour, C. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Hausman, D. (1992). Why look under the hood?. In D. Hausman (Ed.), *Essays on philosophy and economic methodology* (pp. 70–73). Cambridge: Cambridge University Press.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science, 55*, 1–34.

Hoover, K. (2006). A NeoWicksellian in a new classical world: The methodology of Michael Woodford's interest and prices. *Journal of the History of Economic Thought, 28*(2), 143–149.

Kincaid, H. (1986). Reduction, explanation, and individualism. *Philosophy of Science, 53*(4), 492–513.

Kirman, A. (1992). Whom or what does the representative individual represent? *Journal of Economic Perspectives, 6*(2), 117–136.

Kleindorfer, G., O'Neill, L., & Ganeshan, R. (1998). Validation in simulation: Various positions in the philosophy of science. *Management Science, 44*(8), 1087–1099.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In K. Brunner & A. H. Meltzer (Eds.), *The Phillips curve and labor markets* (pp. 19–45). Amsterdam: North-Holland.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

Mayo, D. (2008). How to discount double-counting when it counts: Some clarifications. *British Journal for the Philosophy of Science, 59*, 857–879.

Müller, P., & von Storch, H. (2004). *Computer modelling in atmospheric and oceanic sciences: Building knowledge*. New York: Springer.

Nottestad, L., & Axelsen, B. E. (1999). Herring schooling manoeuvers in response to killer whale attack. *Canadian Journal of Zoology, 77*, 1540–1546.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth sciences. *Science, 263*, 641–646.

Pierce, S., van Gieson, E. J., & Skalak, T. (2004). Multicellular simulation predicts microvascular patterning. *The FASEB Journal*, express article 10.1096/fj.03-0933fje. Retrieved December 13, 2010, from http://www.fasebj.org/content/early/2004/03/31/fj.03-0933fje.full.pdf.

Randall, D. A., & Wielicki, B. A. (1997). Measurements, models, and hypotheses in the atmospheric sciences. *Bulletin of the American Meteorological Society, 78*, 399–406.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. New York: Chapman and Hall/CRC.

Rykiel, E. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling, 90*, 229–244.

Satz, D., & Ferejohn, J. (1994). Rational choice and social theory. *Journal of Philosophy, 91*(2), 71–87.

Sokal, R. R., & Rohlf, F. J. (1994). *Biometry* (3rd ed.). New York: W.H. Freeman.

van der Waals, J. (1910). The equation of state for gases and liquids (1910 Nobel Prize lecture).

Woodford, M. (2006). Comments on the symposium on interest and prices. *Journal of the History of Economic Thought, 28*(2), 187–198.

Worrall, J. (2002). New evidence for old. In P. Gardenfors (Ed.), *In the scope of logic, methodology and philosophy of science* (pp. 191–209). Dordrecht: Kluwer.