A Bayesian Approach to the Validation of Agent-Based Models

Kevin B. Korb, Nicholas Geard and Alan Dorin

Abstract The rapid expansion of agent-based simulation modeling has left the theory of model validation behind its practice. Much of the literature emphasizes the use of empirical data for both calibrating and validating agent-based models. But a great deal of the practical effort in developing models goes into making sense of expert opinions about a modeling domain. Here we present a unifying view which incorporates both expert opinion and data in validating models, drawing upon Bayesian philosophy of science. We illustrate this in reference to a demographic model.

1 Introduction

Agent-based models (ABMs) are computer simulations of numerous, heterogeneous "agents". The models' microbehavior is determined by explicitly programmed rules, while their macrobehavior is not, instead emerging from the collective behavior of the population of agents, usually in very complex ways. This kind of simulation has grown from early efforts in ecology and artificial life into one of the most widely applied computer methods across the sciences today. Nevertheless, skepticism about the interpretation and epistemological standing of these models remains widespread and will do so until at least the fundamentals of ABM validation are agreed upon. Here we present and defend a Bayesian approach to ABM validation.

As many have remarked, the theory of how to validate ABMs is vastly underdeveloped compared to its practice (e.g., Klein and Herskovitz, 2005, Kleindorfer

Kevin B. Korb
Monash University, e-mail: kbkorb@gmail.com

Nicholas Geard
University of Melbourne, e-mail: nicholas.geard@unimelb.edu.au

Alan Dorin
Monash University, e-mail: alan.dorin@monash.edu

et al., 1998). This is unsurprising given how rapidly ABMs have grown from a niche computer application in the 1980s to a leading research technology for ecology (Grimm and Railsback, 2005), economics (Tesfatsion and Judd, 2006), epidemiology (Auchincloss and Diez Roux, 2008) and dozens of other sciences (see http://jasss.soc.surrey.ac.uk/JASSS.html for a full range of examples).

In this paper we take some of the principles of Bayesian theories of scientific method and develop them into an account of validation practice for simulation science. The Bayesian approach to philosophy of science explicitly recognizes the distinction between the current understanding of the behavior of a system (prior belief) and the data (likelihood), which provides us with a framework for integrating both qualitative and quantitative approaches to validation. In essence, our prior belief about the model is updated in the light of experimental data gathered from our simulations.

Bayesian inference tends to accord well with an Ockham-like favoritism for simplicity (e.g., Wallace, 2005). By contrast, both the systems under study and the ABMs themselves tend to be complex, nonlinear and high dimensional. This complexity raises some special epistemological questions about Ockham's Razor, which we address in §3.

After developing a Bayesian approach to validation in the abstract, we illustrate it in reference to the demographic submodel of an epidemiological ABM.

2 Bayesian Philosophy of Science

Klein and Herskovitz (2005) have presented a case that Karl Popper's falsificationism (Popper, 1959) be made the basis for the epistemology of simulation. Popper's account of methodology has many virtues, which have made his name prominent throughout the sciences and perhaps even seem synonymous with philosophy of science. For example, Popper's emphasis on "severely testing" theories — pitting them experimentally against an alternative, such that one or the other must become falsified — is very agreeable to the empirical spirit. Likewise his emphasis on the fallibility of scientific method agrees with both the history of science and traditions in scientific education. Regardless, there are many difficulties standing in the way of a Popperian theory of method. Kuhn (1962), Lakatos (1970) and Feyerabend (1975) all demonstrated with numerous historical examples how in a great many cases unexpected results were rationally held to be *anomalous*, rather than *falsifying*, demanding, not rejection of the theory under test, but instead the discovery and elaboration of new auxiliary hypotheses which could explain the discrepancies between theory and observed reality.

In view of the importance they place on accounting for the accumulation of scientific knowledge, more troubling for Klein and Herskovitz (2005) will be the fact that Popper never gave any reasonable account of the *growth* of knowledge. His reliance strictly upon falsification left any account of support, confirmation or growth of things *known* at best open. To be sure, Popper talked much of "corroboration", even developing a measure of degrees of corroboration. That is, theories that have survived more severe tests are meant to have higher degrees of corroboration than theories that have survived less severe tests. This was supposed to fill the vacuum left by an epistemology exclusively reliant upon refutations, but a vacuum filled with a fictional aether is still just a vacuum. Popper insisted not just that all such corroborated theories were lacking any empirical support, but also that they were, in point of fact, *false*. On Popper's repeated account, all synthetic universal hypotheses are false (e.g., Popper, 1959, Appendix *vii) and simply waiting for their refutations to be found![1]

2.1 Bayesian Confirmation Theory

Bayesian philosophy of method has grown from the ashes of Popperianism. Bayesianism has been propelled by numerous factors: in artificial intelligence and statistics by the development of new methods for exact inference (in Bayesian networks; e.g., Pearl, 1988, Korb and Nicholson, 2010) and approximate inference (in MCMC simulation; e.g., Friedman and Koller, 2003); in cognitive neuroscience (Glimcher, 2004); and generally across many sciences through the explosive growth of the accessible computational power needed for these kinds of analyses.

In philosophy a driving force for Bayesianism has been a string of successful Bayesian re-analyses of Popperian insights into method, combined with an approach that supplies what Popper could not: a theory of theory confirmation.

All of this originates in Bayes' theorem (Bayes, 1763), which simply describes the posterior probability of a hypothesis (conclusion) and in terms of its prior probability and likelihood. In particular,

$$P(h|e) = \frac{P(h) \times P(e|h)}{P(e)} \tag{1}$$

This is an analytic theorem. Bayesian confirmation theory goes well beyond it: it asserts that the proper way to assess confirmation is to adopt the probabilities conditional upon the available evidence — as supplied by Bayes' theorem — as our new posterior probabilities. This move to a posterior distribution is called *Bayesian conditionalization.*

Given this view, the simplest way of understanding the concept of the confirmation or support offered by some evidence is as the difference between the prior and posterior probabilities of a hypothesis; that is, e supports h just in case $S(h|e) = P(h|e) - P(h) > 0$ (cf. Howson and Urbach, 1993, p. 117). A second measure of support, the ratio of likelihoods e given h over e given not-h, is equally defensible (Good, 1983):

[1] This was Popper's extreme skeptical "solution" to Hume's problem of induction: stop inducing! And never mind that his statement itself is a synthetic universal. Popper was no more bound by the petty hobgoblin Consistency than any inductivist!

$$\lambda(e|h) = \frac{P(e|h)}{P(e|\neg h)}.$$

It is a simple theorem that the likelihood ratio is greater than one if and only if $S(h|e)$ is greater than zero. $\lambda(e|h)$ (or, simply, λ) can be understood as a degree of support most directly by observing its role in the odds-likelihood version of Bayes' theorem:

$$O(h|e) = \lambda O(h).$$

This asserts that the conditional odds on h given e should equal the prior odds adjusted by the likelihood ratio. Since odds and probabilities are interconvertible ($O(h) = P(h)/P(\neg h)$), support defined in terms of changes in normative odds measures changes in normative probabilities just as well as $S(h|e)$. λ has a significant advantage over $S(h|e)$ however: it is easier to calculate. Since hypotheses often describe how a system functions given initial conditions, finding the probability of the evidence assuming h is often straightforward. What a likelihood ratio reports is the normative *impact* of the evidence on the posterior probability, rather than the posterior probability itself (which would require also the *prior* probability of h). However, confirmation theory is concerned with accounting just for rational *changes* of belief, and so λ turns out to be the best tool for understanding confirmation, as we show now with two examples.

(1) Likelihood ratios make clear why Karl Popper's (1959) insistence that scientific hypotheses be subjected to severe tests makes sense. Intuitively, a severe test is one in which the hypothesis, if false, is unlikely to pass; that is, whereas the hypothesis predicts some outcome e, its competitors do not. Since the hypothesis predicts e, $P(e|h)$ must be high; since its competitors do not, $P(e|\neg h)$ must be low. Together these imply that the likelihood ratio is very high. So, a severe test will be highly confirmatory if passed and highly disconfirmatory otherwise — providing the most efficient approach to testing a hypothesis, as Popper pointed out.

(2) Another example is the preference which experimental scientists exhibit *ceteris paribus*, when confronted by two possible tests of a theory, for that test which is most different from one previously passed. For example, Eddington had two alternatives to testing Einstein's general theory of relativity (GTR) in 1919: either repeating Einstein's analysis of the precession of Mercury's perihelion or checking the predictions which GTR made of a "bending" of starlight by the mass of the sun, observable during a total eclipse. Despite the fact that astronomical observations of the motion of Mercury are cheaper and simpler, Eddington famously chose to observe the starlight during the eclipse over the Atlantic. Intuitively, we can say that this was because a new result, as opposed to a repeated experiment, offers a more severe test of the theory. For formal Bayesian analyses of this case, see Franklin (1986) and Korb (2004).

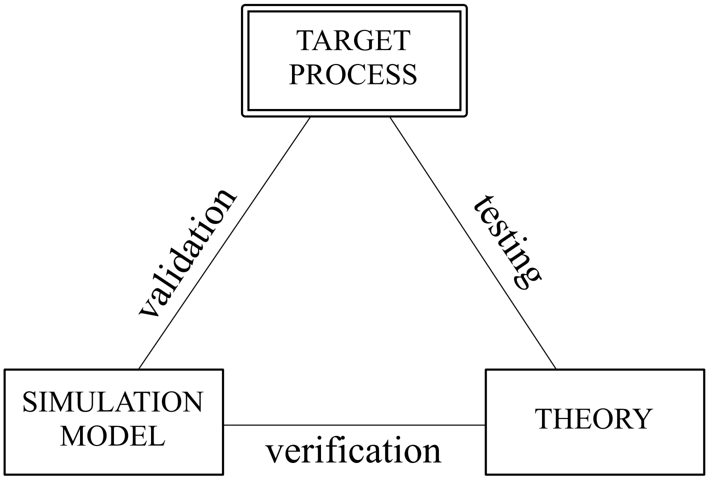
More comprehensive accounts of Bayesian method can be found in Howson and Urbach (1993) and Korb (1992). For our purposes here, it suffices to point out that λ provides a tool for understanding the direction and degree of confirmation or disconfirmation, allowing guidance for validation techniques even when a full probabilistic account is unavailable. We now proceed to a qualitative account of Bayesian ABM validation.

2.2 Bayesian Validation

The goal of empirical validation of computer simulation — its central epistemological question — is to determine whether the simulation is telling us the truth about some target process in the world — whether the theory which it instantiates is true or false.

Some researchers take an unnecessarily narrow view of this process. For example, Windrum et al. (2007) suggest that in order for a validation process to be *empirical* it must directly involve data, which may become an excuse to downplay expert opinions and intertheoretic relations between the theory behind the simulation and related science. But, while empirical knowledge ultimately rests upon sensory experience, it does not have to *directly* rest upon it. We see empirical validation as encompassing both statistical tests using data and expert opinion, which itself (hopefully) derives largely from experience. Bayes' Theorem, in fact, provides a natural form in which to combine these: expert opinions are readily interpreted as providing prior probabilities of a model being correct, while statistics can be used to measure the fit of the model to the data — i.e., its likelihood. This division does not perfectly divide model validation activities, but it does work roughly and, more importantly, serves to reinforce the importance of *combining* expert- with data-oriented validation methods.

ABM simulation is widely understood to involve a tripartite relation:



The central epistemological question can be answered once we know the status of any two of these relations (Mascaro et al., 2010, Chap 3).

Verification.

As the goal of validation is to determine the representational accuracy of a model, the process logically begins with the construction of the model. Assuming that a simulation strictly reflects some underlying scientific theory, then whatever probability that theory has must be shared with its simulation. So, activities which contribute to the confirmation of theories, together with activities which verify that a simulation is true to some theory, also are properly considered an aspect of validation and are most naturally accommodated as contributing to an assessment of the model's prior probability.

Calibration.

In a similar vein, calibrating a model contributes to its probability of being true. This may be an uncommon observation, which is a natural consequence of the calibration of *some* models being trivial. For example, calibrating a binomial model to fit the observed tosses of a coin is trivial, and it also doesn't obviously contribute to the binomial model being true, since *whatever* the bias of the coin, the binomial model could have been appropriately calibrated. Falsificationism suggests that a model which can accommodate *any* data cannot even be tested, let alone be regarded as a true (or good) scientific theory. However, this suggestion is misleading. For one thing, if there were ever any models which could *not* have been calibrated to fit the data, then successful calibration rules them out. Whatever probability *those* models started with must be redistributed, raising the probability of a successfully calibrated model being true. Also, it is worth keeping in mind that models may be either parameterized (fitted), partially parameterized or unparameterized. What can be calibrated to fit any frequency of heads is the unfitted binomial model, and it cannot be disproved (or proved) by any frequency of heads. A fitted, or partially fitted (e.g., with an interval specified for its parameter), binomial model, however, will be more or less probable given some frequency data, and so confirmed or disconfirmed by those data.

We may distinguish between calibration and testing, but that is not to say calibration has nothing to do with the probability of truth. It is, as with verification, properly accommodated in a prior probability.

Emergence.

The emergent (bottom-up) character of ABMs has important epistemological consequences. In general, the most interesting behaviors an ABM might show are macrobehaviors which have not been explicitly programmed, but emerge from lower level rules which have been explicitly programmed. Normally, the higher level behavior could be realized (at least qualitatively) in multiple distinct ways at the lower level. In philosophical terms this relation is supervenience: higher-level behavior supervenes on one or more lower-level supervenience bases.[2]

As the higher level, emergent behavior is often the target of interest, the behavior we should like to predict or explain, it is also the behavior we should like to validate our model against. So, wherever possible, it makes sense to preferentially calibrate with lower level data and validate or test with higher level data.

3 Simplicity or Complexity?

The complexity of ABM models raises the question of the status of Ockham's Razor. Edmonds and Moss (2005) have argued influentially that ABM modeling, contrary to the usual methodological advice, should start out complex and devolve towards simplicity. Ockham's Razor, of course, suggests the opposite: that we should add complexity only in the face of some evidential setback — specifically, that where two theories do equally well with the data, the simpler is to be preferred (Keep It Simple, Stupid). Edmonds and Moss claim that this rule has little or no merit in ABMs and, more specifically, that simplicity confers no epistemological virtue to a model. ABMs are aimed at understanding complex phenomena, and, according to them, should aim to represent them in the "most straightforward way possible", meaning as descriptively detailed as possible (and so their "Keep It Descriptive, Stupid").

Striking just the right balance between simplicity and fit to the data — what Grimm et al. (2005) call finding the "Medawar zone" — is always going to be difficult. And it may well be that many overemphasize simplicity to their own disadvantage. But we disagree that simplicity confers no epistemological advantages.

Undoubtedly, Edmonds and Moss's starting point, the presence of so much complexity in the systems being modeled, can seduce people into over-specifying their models, but that's a danger, not an essence nor a virtue of ABMs. Methodological simplicity, on the other hand, has a number of real, if modest, virtues:

1. The KISS approach is at least a *possible* inductive strategy. Adding complexity where required by evidence is a possible path to the truth, as Reichenbach's (1949) vindication of induction argument suggests. The inverse approach in most domains, where complexity is unbounded, doesn't even begin to make sense, since there is no beginning. And choosing one model from the multitude having complexity comparable to some target system can hardly be justified at the *start* of a research program.
2. Starting out with a complex model implies having a large parameter space. This is not only operationally inconvenient, it is hugely methodologically suspect, since over-specified models fit noise and fail to generalize.

[2] This is often, and wrongly, characterized as micro-reduction. On supervenience, see McLaughlin and Bennett (2011).

3. As a simple model with features added only as needed, a KISS model is far more promising as a vehicle for the consilience of induction, i.e., we can try to adapt it to new and related domains. For example, a KISS measles model might well be usable in a pertussis problem with minimal (and motivated) changes. A KIDS measles model will always only be a measles model.

Perhaps the main virtue that has been put forward for Ockham's Razor, and the one Edmonds and Moss (2005) contest most vigorously, is that simplicity *ceteris paribus* corresponds to higher probability. While widely regarded as true, by both Bayesians and their opponents, this would be exceedingly hard to prove — or to disprove. We don't have any convenient, unbiased collection of examples for testing it. As the probability of simplicity is an exceedingly complex matter, and the advantages of simplicity above are independent, we pass over it (but see Wallace, 2005, for a Bayesian defence of simplicity).

Modeling is not much different from theorizing, as its epistemology shows. It's simply cognitively impossible to start out with a theory that is as complex as the phenomenon. One starts out with a central idea or two, which then get enhanced. Furthermore, the *goal* is to end up with a theory that is at least somewhat simpler than the phenomenon at issue, that can explain it, rather than simply reproduce it. Ockham's Razor is methodologically inevitable.

4 ABM Validation Methods

Here we present a number of recognized types of validity for ABMs, characterizing them in terms of both prior and posterior considerations. We suggest that, as in the case of Bayesian analyses of scientific methods mentioned above, Bayesian analyses of these validation methods can be made and may well improve their usage.

We don't propose that each kind of validity considered here needs to be adopted in every ABM study, however these varieties will generally be worth considering. Here we consider them in the abstract; in §5 we consider some of them relative to our own simulation.

1. **Expert opinion (prior).** This is the usual starting point for constructing and refining computational models. We suggest that this covers most kinds of validation which do not directly involve data, corresponding to what Pitchforth and Mengersen (2012) call **nomological validity:** establishing that the model fits within its wider scientific context. Some of the terminology comes from psychology, by way of Pitchforth and Mengersen (2012). That study focuses upon Bayesian network simulations, however the concerns of simulation epistemology are strikingly similar across ABMs and Bayesian networks.

    a. **Face validity:** Does the model look right to an expert? While face validity is a weak kind of test of a model, it is nevertheless central to most modeling endeavors. Models that look wrong are often abandoned without further ado, something which often causes headaches with machine learned models, since learning algorithms rarely incorporate any kind of aesthetic sense. Face validity should be examined throughout the modeling process, analogously with agile software development processes, where end users provide continuing feedback on the adequacy of software.

Aside from a holistic assessment of a model, all the other forms of validation under *expert opinion* are similarly subjective assessments. Frequent reviews from different experts provide an opportunity for those with varying assumptions about both the model and the domain to provide feedback. Such reviews are also an opportunity to negotiate validity criteria, perhaps including exemptions, when unrealistic aspects of model structure or behavior are deemed less relevant for validation.

b. **Content validity** considers whether the most important factors and relationships between variables noted in the literature are present in the model. Expert opinion will be the primary guide here, but focused reviews of the literature will also be useful.

c. **Case analysis** takes specific instances and examines how the model deals with them. This shares conceptually again with software engineering, where "use cases" are often applied to review software usability, etc. The specific instances may (should) include both normal and extreme cases. They also may be constructed from setting specific initial conditions (as in a historical case study) or from setting parameters that govern relations between individual roles within the simulation (e.g., reproductive or immigration rates) or, of course, from both.

A thorough validation might take further inspiration from software engineering and do an equivalence partitioning of initial conditions to generate a suite of cases that looks at all (or many) varieties of normal and abnormal conditions. Since the results need to be judged by an expert, the value of this depends also upon the patience of the experts available.

d. **Internal validity** examines whether variation in the model's variables is reasonable (Sargent, 2010). This could specifically consider covariation between sets of variables, to determine whether changes in some variable either cause or are codependent with changes in others, in ways which are judged sensible by experts; this is generally called **sensitivity analysis**. The inverse process of **robustness analysis** aims to identify features of the model that are resistant to varying initial conditions (Grimm and Railsback, 2005, Sec 9.7).

2. **Data (likelihood).**

a. **Predictive validity** is the primary way of validating in many discussions. If we were to take "prediction" literally, then even the use of historical data not employed in calibrating the model would be (improperly) excluded (what has been called "retrodiction").

Measuring the fit to data of a model — i.e., predictive accuracy — is again often the only way considered of assessing predictive adequacy. However, predictive accuracy has limitations; see, e.g., Korb and Nicholson (2010, Sec 7.5) for a discussion.

Regardless of the measure used, testing picks up wherever the calibration left off. Reusing data used to calibrate a model to "test" it is generally just an error, since what is then being tested is only the ability of a model (with tuned parameters) to remember what was used to train it. A possible approach to getting the most out of a finite pool of data would be to adapt cross validation methods from machine learning, e.g., using randomly selected splits of the data to repeatedly calibrate with one split and test with the other. The difficulties of calibrating ABMs may limit the utility of this approach, however.

For any measure, some account must be made of the *degree* of accuracy required of the model. It may be that the model is intended to fit data to some precisely specifiable degree of tolerance. Perhaps more common is a requirement that some qualitative aspects of the data be matched, what in economics are called **stylized facts** (Kaldor, 1961, p. 178) and Grimm et al. (2005) call **patterns**. An example from economics would be the positive dependency between support for public education and GDP per capita (e.g., Barro, 2002). This is well established for industrial societies, so a model of modern economies allowing for exceptions would be reasonable, but a model showing no such tendencies would not.

3. **Other.** Not every technique cleanly falls into data or expert opinion, but has aspects of both.

   a. **Convergent validity:** how similar are the model structure, discretization and parameterization to other models that are intended to describe a similar system? Where divergence between models in their assumptions or methods suggests a divergence in results, then we have **discriminant validity**.

      The judgment of the similarity (and relevance) of other models and their features will have to be made by experts, but may well be made in part on the basis of statistical features of data generated by those models.

   b. **Visualization; traces; animation.** Different ways of visualizing the results of simulations may support expert judgments of convergent validity, sensitivity analyses, etc.

   c. **Fruitfulness.** As with the assessment of scientific theories themselves, the fruitfulness of a simulation, its successful adoption by other researchers in application to related problems, is an indirect measure of its validity. In particular, a model which is widely and successfully (re)applied in related problem areas cannot be an entirely wrong-headed model across these domains.

5 Validating a model of household demography

We now briefly illustrate how the validation techniques discussed above might apply in a real ABM, using as a case study a model of household demographics, developed as a component of a larger epidemiological simulation. This model is relatively simple, exhibiting emergence of household and population-level dynamic patterns from individual-level demographic processes.

Households are an important focus of disease transmission with a special relevance for childhood diseases, with the probability of transmission known to be affected by family size and composition (Viboud et al., 2004). Existing models typically assume a static household distribution. However, this is inappropriate when dealing with long-term patterns of disease and immunity for endemic diseases like measles or pertussis (Glass et al., 2011), during which dramatic shifts in underlying demographic rates may occur. However, accounting for the variety of household types and the transitions between them in a mathematical model would be extremely difficult; hence our ABM.

The primary requirements of our model were that it capture the composition and dynamics of households containing children in a plausible fashion over extended periods of demographic change, and that it be amenable to calibration using data from a variety of different developed and developing countries, allowing for international comparison.

Our model represents a population of individuals, defined by their ages, sexes and the households to which they belong. At each time step, depending on their current attributes, individuals can experience one or more of the following demographic events: death, birth of a child, leaving their family home, forming or breaking a couple with another individual.

For some parameters of our model, such as mortality and fertility, age and sex specific rates were directly available (Australian Bureau of Statistics, 2010a,b). However, for other parameters, such as the probability of leaving home, and the formation and separation of couples, data were not readily available. We adopted relatively simple rules for estimating the probabilities of these events occurring, which we subsequently adjusted by calibrating simulation performance against the data that was available. For example, to determine parameters for couple formation we tested our model's output against survey data on the percentage of people at particular ages who had never been in a couple (de Vaus, 2004). This process involved adjusting the age at which an individual becomes eligible for forming a couple with another individual as well as the probability of an eligible individual forming a couple. A similar procedure was used to calibrate parameters corresponding to couple dissolution.

Having calibrated our model using statistics concerning individual-level events, our validation exercise focused primarily on population structure and household dynamics. The quantity of data against which we could validate varied according to country and year, so a broader approach than just data comparison was required. Space precludes a complete description of our validation methods and results (a paper on this is in preparation); instead we describe how each of the categories in Section 4 could be applied to our model. Note that some of the validation processes were more straightforward than others, and, in general, any one validation process may be more or less relevant depending on the particular model.

1. **Expert opinion (prior).**

    a. **Face validity.** To some extent we are all familiar with the varied dynamics of population and households and our own intuitions provided a first point of contact for face validity. The field of demography (via both expert researchers and literature) provided more specialized perspectives on what constitutes a model that looks ‘right’. One important point is that experts from different disciplines may judge the same model differently. This validation process therefore provided an opportunity to negotiate an appropriate set of criteria for further validation, as well as to identify ‘exemptions’ — aspects of model behavior that may be unrealistic, but are deemed unimportant in the context of the research question. For example, our model does not currently allow for the existence of ‘group households’ (e.g., student share households); however, as these types of household typically do not contain young children (the focus of our research question), this was considered an acceptable omission. In the context of a different research question (e.g., the epidemiology of sexually transmitted diseases), this design choice may render the model invalid.

    b. **Content validity.** As mentioned above, engagement with domain experts and literature provided the check on the completeness (or reasoned omission) of factors and relationships in our model. Particularly helpful were documents such as the Australian Institute of Families report (de Vaus, 2004), which aggregated and contextualized census and survey data on households under chapter headings that matched the types of individual life transitions we wanted to capture in our model (e.g., chapter titles include “Marriage and remarriage”, “Transition of young people to adulthood” and “Lone parent families”).

    c. **Case analysis.** During the development and verification of our model we used Australian data collected in the last decade. Despite keeping our calibration (individual level) and validation (household level) data sets separate, we were aware of the possibility that we could consciously or unconsciously be ‘designing’ our model to reproduce a very specific pattern of behavior. To guard against this, subsequent to final development, we validated model behavior against two new cases, using previously unused data sets: historical Australian data from 1921 and Zambian data from 2000. Both of these populations differed from the modern Australian population data along several dimensions. For example, the average household size in Australia in 1921 was 4.3 individuals, as compared with 2.6 in 2000. The success of our model in passing validation tests on this data, without requiring new adjustments to the underlying mechanics, strengthened our confidence in the general model.

    d. **Internal validity.** We took two approaches to assessing the internal validity of our model. First, we re-collected output data on distributions of the individual events whose probabilities we had calibrated. As calibration was performed on individual model components, comparing these output distributions against the calibration data provided a straightforward way of checking that interactions between components were not producing any unexpected side-effects in the combined model. Our second approach was to conduct a sensitivity analysis on the input parameters governing household formation and dissolution (i.e., leaving home and the formation and separation of couples). Compared with the easily available mortality and fertility rates, these parameters required more indirect estimation from available data. Therefore, assessing the sensitivity of our model output to these parameters provided an indication of how critical these values are and how successful our estimation had been.

2. **Data (likelihood).**

    a. **Predictive validity.** The general principle we adopted in separating calibration and validation data was to calibrate the probabilities of individual events (birth, death, couple formation, etc.) and validate against higher-level properties of households. Data available for validation included the distribution of household sizes, distributions of household types occupied by individuals of given ages (couple households with/without children, lone person households, etc.), and household transition matrices, mapping the proportion of individuals in a household of type X who had been in a household of type Y at some point in the past (Wilkins et al., 2011). Each of these constituted a set of data that was clearly distinct from our calibration data, against which model output could be compared in a quantitative fashion.

6 Conclusion

Our data-directed and expert validation efforts have shown that the demographic model is doing a reasonable job of recreating long-term demographic patterns in our target population (currently Australia), supporting our planned use of it as a platform for developing epidemiological simulations.

The simple Bayesian message we would like to finish with is that a validation process that concentrates on expert consensus to the exclusion of collecting statistics from data, or, equally, one which tests against data but ignores expert opinion, is incomplete. It is only by combining prior probabilities with likelihoods that we obtain a balanced picture of the empirical merits of a model.

Acknowledgements

This work was in part support by ARC grant DP110101758. We would like to thank our project partners Jodie McVernon, James McCaw, Kathryn Glass and Emma McBryde for their support. Alessio Moneta, Roman Frigg, Jay Pitchforth and referees also provided helpful comments.

References

Auchincloss, A. H. and A. V. Diez Roux (2008). A new tool for epidemiology: The usefulness of dynamic-agent models in understanding place effects on health. *American Journal of Epidemiology 168*(1), 1–8.

Australian Bureau of Statistics (2010a). *Births, Australia, 2009*, cat. no. 3301.0, viewed 18 August 2011.

Australian Bureau of Statistics (2010b). *Life Tables, Australia, 2007–2009*, cat. no. 3302.0.55.001, viewed 18 August 2011.

Barro, R. J. (2002). Education as a determinant of economic growth. In E. P. Lazear (Ed.), *Education in the Twenty-First Century*, pp. 9–24. Palo Alto, CA: The Hoover Institution.

de Vaus, D. (2004). *Diversity and change in Australian families: Statistical profiles*. Melbourne, Australia: Australian Institute of Family Studies.

Edmonds, B. and S. Moss (2005). From KISS to KIDS: an 'anti-simplistic' modelling approach. In P. Davidsson (Ed.), *Multi-Agent and Multi-Agent-Based Simulation 2004*, pp. 130–144. Springer.

Feyerabend, P. (1975). *Against Method*. Verso.

Franklin, A. (1986). *The Neglect of Experiment*. Cambridge University.

Friedman, N. and D. Koller (2003). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning 50*, 95–125.

Glass, K., J. McCaw, and J. McVernon (2011). Incorporating population dynamics into household models of infectious disease transmission. *Epidemics 3*, 152–158.

Glimcher, P. W. (2004). *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. MIT Press.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. University of Minnesota.

Grimm, V. and S. Railsback (2005). *Individual-based Modeling and Ecology*. Princeton: Princeton University Press.

Grimm, V., E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science 310*, 987–991.

Howson, C. and P. Urbach (1993). *Scientific Reasoning: The Bayesian Approach* (2nd ed.). Open Court.

Kaldor, N. (1961). *Capital Accumulation and Economic Growth*. Macmillan.

Klein, E. E. and P. J. Herskovitz (2005). Philosophical foundations of computer simulation validation. *Simulation & Gaming 36*, 303–329.

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

Korb, K. B. (1992). *A Pragmatic Bayesian Platform for Automating Scientific Induction*. Ph. D. thesis, Indiana University.

Korb, K. B. (2004). Bayesian informal logic and fallacy. *Informal Logic 24*.

Korb, K. B. and A. Nicholson (2010). *Bayesian Artificial Intelligence* (2nd ed.). Boca Raton, FL: CRC Press.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago.

Lakatos, I. (1970). *Criticism and the Growth of Knowledge*. Cambridge University.

Mascaro, S., K. B. Korb, A. E. Nicholson, and O. Woodberry (2010). *Evolving Ethics: The New Science of Good and Evil*. Imprint Academic.

McLaughlin, B. and K. Bennett (2011). Supervenience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011 ed.).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Pitchforth, J. and K. Mengersen (2012). A proposed validation framework for expert elicited Bayesian networks. *Decision Support Systems and Electronic Commerce*. Submitted.

Popper, K. (1934/1959). *The Logic of Scientific Discovery. Translation of* Logik der Forschung. New York: Basic Books.

Reichenbach, H. (1949). The the pragmatic justification of induction. In H. Feigl and W. Sellars (Eds.), *Readings in Philosophical Analysis*, pp. 305–327. New York: Appleton-Century-Crofts.

Sargent, R. G. (2010). Verification and validation of simulation models. In *Proceedings of the 2010 Winter Simulation Conference (WSC)*, pp. 166–183.

Tesfatsion, L. and K. L. Judd (Eds.) (2006). *Handbook of Computational Economics, Volume II: Agent-Based Computational Economics*. Amsterdam, The Netherlands: Elsevier.

Viboud, C., P. Y. Boëlle, S. Cauchemez, A. Lavenu, A. J. Valleron, A. Flahault, and F. Carrat (2004). Risk factors of influenza transmission in households. *British Journal of General Practice 54*, 684–689.

Wallace, C. S. (2005). *Statistical and Inductive Inference by Minimum Message Length*. Springer Verlag.

Wilkins, R., D. Warren, M. Hahn, and B. Houng (2011). *Families, Incomes and Jobs, Volume 6: A Statistical Report on Waves 1 to 8 of the Household, Income and Labour Dynamics in Australia Survey*. Melbourne, Australia: Melbourne Institute of Applied Economic and Social Research.

Windrum, P., G. Fagiolo, and A. Moneta (2007). Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation 10*.