Injecting Data into Simulation: Can Agent-Based Modelling Learn from Microsimulation?

Samer Hassan[1,2], Juan Pavon[1], and Nigel Gilbert[2]

[1] GRASIA: Grupo de Agentes Software, Ingeniería y Aplicaciones, Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, Madrid, 28040, Spain {samer, jpavon}@fdi.ucm.es

[2] CRESS: Centre for Research in Social Simulation, Department of Sociology, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom
n.gilbert@surrey.ac.uk

Abstract. Most agent-based models use uniform random distributions to configure the values of initial conditions in simulations. Moreover, random values are often used to distribute objects spatially, to determine unmeasured exogenous factors, and sometimes to determine aspects of the agents' behaviour. An alternative approach to the design and initialisation of an agent-based simulation is to adapt the principles of microsimulation using external sources of contextual data. In this approach, quantitative data are used in several ways: sample surveys for the initial conditions, calculated regression equations for the evolution of variables, and empirically based distributions for the calculation of new values. In this paper, we consider some of the advantages and difficulties of this alternative approach.

Key words: agent-based modelling, microsimulation, quantitative data, random initialisation, social simulation

1 Introduction

Many Agent-Based Models (ABM) aim to simulate some real-world phenomenon and their validation is usually driven by empirical data. However, the initial conditions usually do not attempt to reproduce the real world. Most often, the simulation begins with values taken from a uniform random distribution. But there are many cases where the choice of initial conditions can affect the output of the model and where a uniform random distribution is a poor choice.

Several recent initiatives have considered the introduction of empirical data into ABMs [1]. A well-known example is the model of the extinction of the Anasazi civilisation, in which empirical data are used for improving the fit between the simulation and the observed history. In this example, the exogenous factors (environmental variables) are not randomized, although the initial conditions are [2]. Another example is the water demand models of [3, 4], in which data

about household location and composition, consumption habits and water management policies are used to steer the model, with good effect when the model is validated against actual water usage patterns. A third case is Hedström's model of youth unemployment [5] in which data from surveys are imported and regression equations are used to calculate transition probabilities. From a broader point of view, there are examples such as the pedestrian flow modelling using spatial data [6] and simulations of markets such as that of the electricity market [7].

These examples show how the issue of random initialisation can be addressed successfully by gathering data and feeding the model with it. There are some similarities with a technique called microsimulation (also known as microanalytic simulation) [8]. Microsimulation focuses on the simulation of the behaviour of individuals over time. The individuals are initialised with empirical data (usually derived from a sample survey). The simulation consists of repeatedly changing the simulated individuals according to a set of transition probabilities and transition rules (ideally, both extracted from empirical data). However, microsimulation does not model interactions between individuals, each of whom is considered in isolation.

This paper encourages ABM designers to continue the data-driven trend, by merging some concepts taken from microsimulation into ABM. This approach can contribute to obtaining simulation results that are closer to observations of the corresponding target, as will be shown in a case study described in section 4. The next section presents in more detail some of the problems that result from random initial configurations. In section 3, the alternative approach is outlined, while in section 4 it is applied to a case study. The final section concludes with a few tentative guidelines.

2 Random Initialisation

As has been mentioned, the uniform random distribution is commonly applied to generate a model's initial conditions. The typical procedure is to run a series of simulations (each with a different starting random seed value) and aggregate their outputs into a mean. This is an appropriate method to check the relationships among a set of parameters in a model. However, it does not ensure that the output cannot be improved with other initial conditions, especially when there is a need to compare with real systems and precise data.

To understand why, let us assume that at least some observable elements of the real world are stochastic. Then the one instance of the real world that actually exists can be thought of as a random selection from a population of possible worlds. That means that, while the most probable case is that the real world has the same attribute values as the means of the values in all possible worlds, it is also quite likely that the real world value is not close to the mean and certainly possible that it is an outlier, far from the mean. Now suppose that due to some happy chance, we have a model that accurately represents the real social processes. We initialise the model with random conditions, run it many

times and calculate the mean behaviour. We then compare this mean behaviour with the behaviour observed in the real world. There is a chance that the two will not match. If the real world happens to be an outlier, the discrepancy could be very large. On the other hand, if we start with initial conditions that are taken from data, even if the real world is an outlier, the data will to some degree move the model in the direction of the real world, and we are much more likely to find a match between the model and the observed data.

One specific case where uniform random initial conditions are well known to be inadequate is the distribution of links between agents to form a network. Real social networks invariably display a higher degree of clustering than random graphs [9] and a number of algorithms are available to generate networks that more closely match empirical degree distributions [10].

An example of how an ABM can be improved by introducing data, in contrast with the random approach to initialisation, is a study of the Eurovision song contest [11]. This considers voting in a popular music contest in Europe, and begins with the hypothesis, "over a sufficiently long period of time the results of the Eurovision contest would approximate to random". If the hypothesis were true, a simulation with random initial conditions and random voting schema should approach the real situation. But actually it does not. It is shown that introducing empirical data, such as the distance between countries (if a country is closer, people are more likely to vote for it) or a measure of the similarity of their cultures, improves the results of the simulations.

3 A Method for Data-driven ABM

Our aim here is to propose an alternative to the classical stochastic design normally adopted in the design of ABM. The design is an idealisation of what will normally be a less clear cut process. The method could be specially useful in contexts where there are quantitative or qualitative empirical data from existing sources, or at least the possibility of collecting samples of such data.

Microsimulation [8] has traditionally been used in areas where it is easy to obtain quantitative data, in the form of surveys and censuses (for the initialisation of individual units) and equations or rules (for defining agent behaviour). Although microsimulation has been successful in some problem domains such as traffic modelling and econometrics, it has been difficult to apply in social domains that are not so well structured or where there are important dependencies between agents. Microsimulation is unable to model interactions between agents, an area where agent-based modelling is pre-eminent. Nevertheless, some aspects of microsimulation, such as basing the simulation on representative survey samples and using probability transition matrices to determine changes in the values of agent parameters, can usefully be applied to the design of ABM. Agent-based models usually follow an event-based rules approach rather than using transition probabilities. However, the limitations of modelling or the lack of sufficient data frequently make it difficult to implement explicit rules and therefore they have to turn to other solutions, one of which is to use transition probabilities, which

represent implicit rules. Qualitative information, although rarely used in ABM, can also be introduced [12].

Adopting this approach, it is possible to reformulate the classical stages of the logic of simulation, which uses empirical data just for validation [13], to include data-guided design and initialisation:

1. *Collection of data* from the social world.
2. *Design of the model*, which should be guided by some of the empirical data (e.g. equations, generalisations and 'stylised facts', qualitative information provided by experts) and by the theory and hypotheses of the model.
3. *Initialisation* of the model with static data (from surveys and the census).
4. *Simulation* and output of results.
5. *Validation*, comparing the output with the collected data. The data used in validation should not be the same as that used in earlier steps, to ensure the independence of the validation tests from the model design.

Although these stages are presented in a linear way, the design and development process is usually carried out in an iterative manner. For example, the results of the Validation stage may force changes to the design of the ABM.

The application of this procedure can present some difficulties. For some models, especially those at a high level of abstraction, appropriate data may be impossible to obtain. Another problem, common also to microsimulation, is the requirement for large volumes of detailed data about individuals. Sometimes, the lack of data stems merely from the absence of suitable surveys and other data sources. Sometimes, the problem is more fundamental. For example, agent characteristics such as their emotional states are unobservable. In some models, the agents' current state depends on their previous circumstances (this is the case, for example, in models which incorporate path dependencies, or where agents have memory). However, it is rare for such histories to be recorded systematically in representative surveys. Panel studies can come to the rescue, but these are not common. A related issue is the need for dynamic data that measures changes over time in addition to the more usual 'snapshot' data sets typically available from surveys. It is also often hard to obtain information regarding networks and micro-interaction processes, unless one is dealing with very particular domains such as virtual communities where data are recorded as a side effect of electronic interactions [14].

Some of these problems can be overcome or worked around. For example, if we want to simulate a married couple, we can find a wife in a survey based on a random sample of individuals, but we also need an agent to represent her husband. Since the data are taken from a random sample, it is unlikely that the husband will also be in the survey. Strategies for dealing with this include creating an artificial 'husband', not based on anyone in the sample; or 'marrying' the woman to a different, married man in the sample.

However, there are some unavoidable costs associated with the introduction of empirical data. In some cases complicating the model with empirical data does not bring benefits. In those cases and others, a KISS ("Keep It Simple,

Stupid") model may be better. Therefore, the decision whether to use empirical data should be made on a case-by-case basis.

The ease of understanding and communication associated with very simple models can be lost when empirical data is introduced. Another potential cost in introducing data into a model is a loss of generality, because the data is specific to one site and time, However, the alternative, to use a standard distribution, may yield a model that is a representation of no site or time whatsoever.

4 A Case Study: the Mentat Model

4.1 Context of the Model

The aim of the Mentat model [15] is to understand the evolution of multiple factors in Spain from 1980 to 2000, focusing on social and moral values. This period is interesting because of the substantial shift in moral values corresponding to the transition from a dictatorship to a consolidated democracy. The almost 40 years of dictatorship finished on 1975, when the country was far from Europe on all indicators of progress, including the predominant moral values and modernisation level. However, the observed evolution of moral values since then are analogous to those found in its EU partners. Furthermore, the changes in Spain have developed with a special speed and intensity during the period studied. The main factor proposed to explain the observed changes is demographic: the change in the age structure of the population and the influence of a younger generation. The Mentat model aims to simulate the effect of cross-generational changes, focussing on these "vertical" rather than on "horizontal" influences.

The Mentat model hypothesises that values are influenced by a range of factors, including demography, economy, political ideology, religiosity, family and friend relationships, reproduction patterns, and stage in the life course. We shall use the model to examine the effect of initialising it with empirical data, as compared with a version initialised using a random distribution. The behavioural rules at the individual level are the same in both versions. To simplify the comparison still further, we reduce the number of objective variables to the one most critical: age. Its distribution will determine the demography of the system: agents die when they are old, they search for a partner in youth, they have more or less chance to have a child depending on age, etc. Both versions of the ABM will then be validated against additional empirical data (not previously used in model initialisation).

The simulation has been configured with a population of 3000 agents and simulated for a period of 20 years (from 1980 to 2000). The agents are able to communicate, establish friendship and couple relationships, and reproduce. They form a network where the nodes are the individuals and the links can be of type 'friend' or 'family' (couple, parents, children). The more friendships exist, the more couples and families will be formed (as the partner is chosen from friends). The model includes age-related probabilities of having children (for example, a woman in her forties will have less chance than a 23 years old);

regression equations to determine whether an agent searches for a partner or not; and time-varying transition matrices for life expectancy and the fertility rate (the birth rate in Spain fell from 2.2 in 1980 to 1.19 in 2000).

4.2 The Randomly Initialised Version: Mentat-RND

The version of the ABM with random initial conditions has been named Mentat-RND, while the one with empirically based initialisation is Mentat-DAT. Both have exactly the same structure except for the source of the ages of the initial agent population. In Mentat-RND this attribute has been assigned using an uniform random distribution in the range [0, 75].

The output of the system consists of several statistics directly affected by the demographic model and the population pyramid (age distribution). We monitor the percentage of old people, the ratio of single to married agents, and the overall population growth (determined by the number of couples and their age).

The system's output is unstable, with noticeable changes between executions, so an aggregation measure is needed. The model was executed 15 times and each statistic averaged. The results are compared with empirical data and with the results of Mentat-DAT.

4.3 The Version Initialised with Data: Mentat-DAT

The agents in Mantat-DAT are initialised using data from the Spanish census, research studies and sample surveys [16]. The basic input is from the Spanish sample of the 1980 European Values Survey (EVS). The data provide a range of variables, including demographics, attitudes and financial information, for a representative sample of 2303 individuals surveyed in 1980. The data are used to generate a simulated population with the same statistical distributions of the main parameters as the whole Spanish population. Consequently, the population pyramid in the model is similar to the real one in Spain in the 1980s.

While Mentat-DAT is initialised using data from the 1980 EVS, the outputs from it (and Mentat-RND) after 10 and 20 simulated years are compared with data drawn from the 1990 and 1999/2000 European Values Surveys. The three sweeps of the EVS thus provide independent data sets for initialisation and for validation.

4.4 Comparison of Outputs

In this section we compare the results from Mentat-RND (random initialisation) and Mentat-DAT (data initialised version), contrasting them with data from the Spanish Population Census and the 1990 and 1999/2000 EVS. To put the differences of the two models in the simplest way, we have chosen three indicators for the analysis, as can be seen in Table 1: the proportion aged 65 and over, the proportion that are unmarried, and the population growth rate. The values of these parameters for the two versions of the model are shown for each point

in time for which we have empirical data. A deeper analysis of the average evolution of the main parameters can be found in [17]. The values for Mentat-RND are averaged over 15 executions to allow for stochastic variations in its output. Mentat-DAT is almost stable between executions because of its fixed initialisation and so the means shown are based on only 5 runs.

Consider first the proportion of older people. The Census shows that this has been growing, starting at 18 per cent in 1980 and reaching 21 per cent by 1999. Mentat-RND begins with almost the correct figure in the (simulated) year 1980, but the rate of growth is much faster than it should be. On the other hand, Mentat-DAT shows a closer fit to the empirical data.

The observed proportion of single people is steady over time. The number of couples in the ABM is directly proportional to the number of friendship links, so the ratio of single to married agents is a good measure of the cohesion of the network. In Mentat-DAT, the attributes of the individual agents are initialised from the 1980 EVS data, but not the couples, as there is no information about links between members of the sample in the EVS. The simulation must therefore start by creating such links to build the network structure. Only after some execution steps does the proportion of couples converge to a steady state. We can see that Mentat-DAT is again closer to the survey data than the random initialised version. Continuing both simulations beyond 20 years allows us to observe a convergence to a proportion of single people around [28,30], but this is reached more slowly by Mentat-RND.

For the case of population growth, the randomized version generates a rate of 10.1 per cent, higher than the Census (8 per cent), while the data-driven version has a growth rate slightly lower (7.2 per cent) than the Census. Overall, the data-driven Mentat-DAT provides a closer fit to the empirical data than the randomly initialised Mentat-RND for all three of these parameters.

Table 1. Validation: comparison between EVS, the random initialised version and the data-driven version

	EVS/Census*			Mentat-RND			Mentat-DAT		
	1980	1990	1999	1980	1990	1999	1980	1990	1999
% 65+ years	16*	18*	21*	19	24	29	15	19	24
% Single	28	29	29	-	45	37	-	42	35
% Population Growth	-	-	+8%*	-	-	+10.1%	-	-	+7.2%

* Source: Spanish Population Census for the years 1981, 1991 and 2001

5 Concluding Remarks

The motivation for this paper was a concern about the use of random initialisation in ABM, and the possibility of basing models more closely on empirical data. An alternative approach merges some aspects of Microsimulation with ABM and

can be useful in certain cases, such as the one presented here. This case shows that feeding a model with empirical data can improve the fit between the model and the observed social world: for example, its internal dynamics, its macro-level behaviour, and the structure of the networks linking agents.

We have suggested that exposing a model to data does not have to be left to the final, validation step, but has value at the very beginning of the modelling process. A closer look on how to deal with this can be found in [18].As a result of our experience with the Mentat model, we suggest that:

- It is valuable to explore the problem background, focusing not only on the theoretical literature, but also on the availability of data.
- It is worthwhile to compare different collections of data and conclusions from diverse sources to give a stronger foundation to the model.
- The most valuable data are those that provide repeated measurements, preferably taken from the same respondents (as in a panel survey).
- The ABM should be designed so that it generates output that can be compared directly with empirical data.
- If the data are available, it is recommended to simulate the past and validate with the present, as was done in the case study.

The effect of applying these suggestions would be to connect the majority of agent-based models more closely to the social world that they intend to simulate, at the cost of the extra effort and complication involved in injecting data into the simulation.

Acknowledgments. We acknowledge support from the projects: *NEMO: Network Models, Governance and R&D collaboration networks* funded by the European Commission Sixth Framework Programme - Information Society and Technologies - Citizens and Governance in the Knowledge Based Society, and *Methods and tools for modelling multi-agent systems*, supported by Spanish Council for Science and Technology, with grant TIN2005-08501-C03-01.

References

1. Boero, R., Squazzoni, F.: Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science. Journal of Artificial Societies and Social Simulation **11** (2005) 6 `http://jasss.soc.surrey.ac.uk/8/4/6.html`.
2. Dean, J.S., Gumerman, G.J., Epstein, J.M., Axtell, R.L., Swedlund, A.C., Parker, M.T., McCarroll, S. In: Understanding Anasazi culture change through agent-based modeling. Oxford University Press (2000) 179–205
3. Edmonds, B., Moss, S.: From KISS to KIDS – an 'anti-simplistic' modeling approach. `http://hdl.handle.net/2173/13039` (2005)
4. Galan, J.M., Lopez-Paredes, A., del Olmo, O.: Effect of technological diffusion of water conservation measures in an ABM–GIS integrated model. In: VI International Workshop on Practical Applications of Agents and Multiagent Systems, Salamanca: Universidad de Salamanca (2007) 169–180

5. Hedström, P.: Dissecting the Social: On the Principles of Analytical Sociology. Cambridge University Press (December 2005)
6. Batty, M.: Agent-based pedestrian modeling. Environment and Planning B: Planning and Design **28** (2001) 321–326
7. Nicolaisen, J., Petrov, V., Tesfatsion, L.: Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. Evolutionary Computation, IEEE Transactions on **5** (2001) 504–523
8. Gupta, A., Kapur, V.: Microsimulation in Government Policy and Forecasting. Elsevier Science Publ. Co. (2000)
9. Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. APS Meeting Abstracts (March 2002) 27012
10. Newman, M.E.J., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. Proceedings of the National Academy of Sciences of the United States of America **99** (February 2002)
11. Gatherer, D.: Comparison of eurovision song contest simulation with actual results reveals shifting patterns of collusive voting alliances. Journal of Artificial Societies and Social Simulation **11** (2006) 1 `http://jasss.soc.surrey.ac.uk/9/2/1.html`.
12. Yang, L., Gilbert, N.: Getting away from numbers: Using qualitative observation for agent-based modelling. In Amblard, F., ed.: ESSA'07: Fourth Conference of the European Social Simulation Association, Toulouse, France (2007) 205–214
13. Gilbert, N.: Agent-based models. Quantitative Applications in the Social Sciences. Sage Publications Inc. (2007)
14. Taraborelli, D., Roth, C., Gilbert, N.: Measuring wiki viability (ii). a framework to identify factors behind sustainable wiki-based communities. (Submitted) (2008)
15. Hassan, S., Pavón, J., Arroyo, M., Leon, C.: Agent based simulation framework for quantitative and qualitative social research: Statistics and natural language generation. In Amblard, F., ed.: ESSA'07: Fourth Conference of the European Social Simulation Association, Toulouse, France (2007) 697–707
16. Pavón, J., Arroyo, M., Hassan, S., Sansores, C.: Agent-based modelling and simulation for the analysis of social patterns. Pattern Recogn. Lett. **29** (2008) 1039–1048
17. Hassan, S., Antunes, L., Arroyo, M.: Deepening the demographic mechanisms in a data-driven social simulation of moral values evolution. In: MABS 2008: Multi-Agent-Based Simulation. LNAI: Lecture Notes in Artificial Intelligence, Lisbon, Springer (2008)
18. Hassan, S., Antunes, L., Pavon, J., Gilbert, N.: Stepping on earth: A roadmap for data-driven agent-based modelling. In: ESSA'08: Fifth Conference of the European Social Simulation Association, Brescia (Italy) (2008) (Submitted).