Social Simulation: investigating population-scale phenomena from the bottom up

Alex Voss[1], Andy Turner[2], Jing-Ya You[3], Eric Yen[3],
Hsin-Yen Chen[3], Simon Lin[3], Ji-Ping Lin[4]

[1] School of Computer Science, University of St Andrews, United Kingdom
[2] School of Geography, University of Leeds, United Kingdom
[3] Academia Sinica Grid Computing, Academia Sinica, Taiwan
[4] Academia Sinica Center for Survey Research, Academia Sinica, Taiwan

Abstract: Agent-based social simulation methods build on and extend existing social science research methods aimed at understanding and forecasting social phenomena in real-world contemporary societies. This paper provides an overview of the process of developing such models and argues that recent developments in information technologies have made it possible to model social phenomena at population-scale. A feasibility study for a model of migration in Taiwan is presented.

Introduction

Agent-based modelling techniques are widely used in a number of disciplines to study the behaviour of individual social agents and the phenomena that such individual behaviour gives rise to at the level of groups of agents or even whole populations. For example, agent-based models are widely used in ecology to study the interactions between animals (of different species) and the environment they live in (Grimm and Railsback 2005, Topping *et al.* 2005). In the social sciences, researchers have started using agent-based modelling to conduct studies that are based on theories about individual-level behaviour and that aim to investigate how such behaviour can give rise to social phenomena of interest. Applications exist in areas as diverse as public health (Epstein 2009, Stroud *et al.* 2007), the study of criminal behaviour (Malleson, Heppenstall and See 2010), crowd behaviour (Richmond and Romano 2008) or the study of financial markets (Lux and Marchesi 2002).

Social simulation (Epstein 2007, Gilbert and Troitzsch 2005) builds on and extends other research methods and builds on existing theories about human behaviour and social phenomena. Some conceptual investigations may be done without reference to data about real-world societies but increasingly researchers are interested in developing simulations that provide insights into the workings of contemporary societies as well as forecasts about social phenomena of importance to policy making. Such models need to be based on data about the populations of interest such as census data or data from surveys. The obvious starting point for the construction of the conceptual model is to use existing theories about societies to develop a model of how society members interact with each other as well as with their environments. The conceptual model is then translated into a computational model that can be executed on a computer system. Simulation runs produce output data that can be analysed to produce knowledge about social phenomena or to forecast social trends.

While simulation has been introduced as a research method in the social sciences as early as the early 1970s (Schelling 1971), the production of complex models involving a multitude of variables, social processes and real-world population sizes is only now becoming a reality as the availability of technical resources increases. Furthermore, software engineering methods have matured to a level where it is possible to formulate simulation models at sufficiently high levels of abstraction to facilitate effective collaboration between social scientists responsible for the conceptual model and technologists responsible for its implementation in a computer system.

Agent-based Modelling

Agent-based models are based on representations of individual members of the population being studied and specifications of their individual behaviour. The representations of population members are called *agents*. It is important to note that an equally important part of agent-based modelling is the representation of other entities that exist in the *context* these agents exist in. These may be features of the natural environment or social constructs such as organisations or families. Where these entities exhibit behaviour not just in response to interactions with population members, they may themselves be modelled as agents.

Agent-based simulations are usually executed through a progression through discrete time steps. That is, a central controlling model component progresses the model time and then triggers any behaviour associated with agents that needs to be executed at this time step. The prescriptions of the behaviour are often not deterministic but controlled by probability distributions and the precise ordering of events is often also randomised to avoid bias being introduced by the fact that individual processing units in a computer process instructions one by one. Agent-based models are therefore related to *Monte-Carlo* methods, which are used in many disciplines in the natural sciences as well as engineering (Liu 2004). Their scientific value is based on the law of large numbers, which ensures that for a large enough population, the randomness produced at the individual level does not matter as the aggregate behaviour will be close to an 'expected behaviour' and any measurements of aggregate variables will be close to their expected values.

Data Use and Data Protection

The empirical data underpinning a simulation model can be taken from a number of different sources such as census programmes or large-scale surveys but also from other sources such as transactional data collected by governments or industry. In addition, researchers may collect data specifically to inform the production of simulation models. Another input to the modelling and simulation that is often not explicitly acknowledged is existing theories of social phenomena that form the basis for decisions made about the selection of data and the various aspects of developing the simulation model.

Raw data from census programmes and surveys by definition is personal data and therefore in most countries falls under data protection legislation that regulates its collection and usage. Census data in particular is considered to be highly confidential and therefore digital census outputs are often subject to disclosure controls that often limit the use that social scientists can make of them. Clearly, a balance must be struck between the interests of the individuals to have their data protected and the interests of the public in general to support social scientific research and its applications.

Often, the full census data at individual level is not made available to researchers for the reasons mentioned above but instead is either made available only in 'safe settings' (Rahman, Jirotka and Dutton 2008) or in a restricted format as census outputs. Safe settings enforce physical access control and usually require a vetting and training process. As the data cannot be exported from these settings, all research has to be done on-site, making it difficult or impossible to conduct analyses that require significant compute resources, other resources or that need to be done collaboratively by different researchers.

Digital census outputs generally made available for research have been subjected to disclosure control measures that result in the removal and aggregation of variables. Two common types of digital census data that are released for general research are Census Area Statistics (CAS) and Samples of Anonymised Records (SAR). Figure 1 illustrates the relationship between raw census data depicted as a table of records at the centre and the CAS depicted to the left and SAR outputs to the right. In the following, we describe the construction of census outputs in the UK, which has a mature census dissemination programme[1] where census outputs are made available to researchers as downloads that require registration and the acceptance of the terms of a usage license.

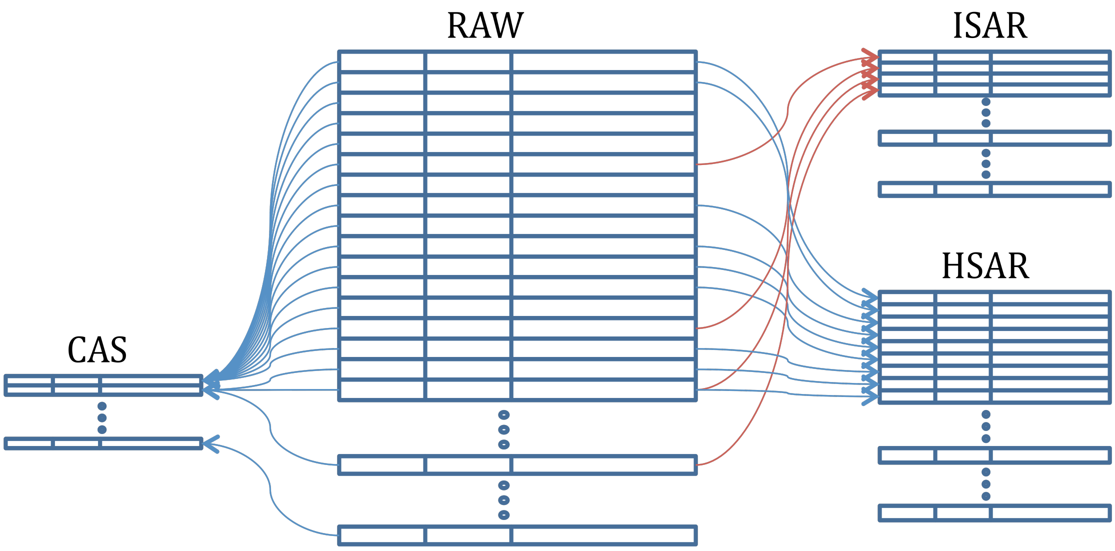

Figure 1: Relationship between raw census data and different forms of census outputs.

[1] http://www.census.ac.uk/

For the census area statistics, a set of geographical areas are defined and data are produced on this basis to describe geographical variations. Aggregate statistics for groups of records are specified (counts and averages) and spatial aggregations of records are produced as CAS. Removal and further spatial aggregation of output by area code may result in CAS at a range of spatial scales. Census area statistics for the UK are made available directly by the Office of National Statistics[2] and are available through a web-based system from the Census Dissemination Unit[3] at the University of Manchester. In Taiwan, data from the 2000 Population and Housing Census is available at the level of aggregation of prefectures[4].

While the CAS data contain statistics on households in specific geographic regions, the samples of anonymised records contain information at the level of the individual person (ISAR) or household (HSAR). As these outputs contain information at the level of individuals or households rather than aggregate statistics, measures are taken to protect individuals from identification through disclosure control. A first step in census data disclosure control involves pseudo-anonymisation whereby each record is given a unique identifying code by some form of secret randomized procedure and the key identifying variable(s) such as personal identification number and name and perhaps some details of address are removed or aggregated. For example, the most detailed address location identifiers for the record may be reduced to partial postcodes or mapped to an output area code describing a geographical area with a population sufficiently large to ensure that the individual person cannot be identified on the basis of knowing their location and the other attributes contained in the record. After pseudo-anonymisation, further disclosure control may be deemed appropriate. Further processing can involve aggregation of variables (generalization), aggregation (grouping) of records, sampling and even the removal and perturbation of variables.

In the UK, the individual-level SAR[5] (I-SAR) contains a 3% sample of individuals for all countries in the UK and is available through the end-user license. A household-level SAR[6] (H-SAR) that allows linkage between individuals in the same family and the same household contains only a 1% sample of households containing less that 12 individuals and is available only under the more restricted accreditation scheme. The degree to which data is selected and aggregated in these datasets differs because the household linkage significantly increases the chances of identification of individuals through data about people related to them.

In summary, the application of disclosure control measures are an attempt to reduce the likelihood that information about an individual can be identified from the data made available without significantly reducing the usefulness of the data for researcher. It is perhaps worth noting that anonymised census data records are usually still more detailed than many surveys and where they are only samples, the sampling is done carefully. A final point is that census programmes for the nations that have them are mandated to collect data for all people and

[2] http://www.statistics.gov.uk/census2001/censusareastats.asp
[3] http://cdu.mimas.ac.uk
[4] http://eng.stat.gov.tw/ct.asp?xItem=8465&ctNode=1629
[5] http://www.ccsr.ac.uk/sars/2001/indiv/index.html
[6] http://www.ccsr.ac.uk/sars/2001/hhold/

tend to be far more complete than other social surveys. Bias abounds in any form of survey, but a census is probably the best source of data that is representative of the whole population. However, the statistical disclosure control mechanisms applied may well mean that the census data that is available to researchers fall short of this ideal, a fact that researchers need to carefully consider. In the end, it will depend on the specific research purpose whether the data available can be considered to be of adequate quality.

In addition to the census data, researchers will normally need additional data covering topics of interest not covered in the census. For example, in order to model the demographic development of the population, a model of reproductive behaviour will be needed that allows the model to accurately represent fertility. For a model studying migration, researchers will need to include a model of income levels in different regions as migration is influenced by economic considerations. Such additional data is often available through the same sources as the census data but is likely to be based on surveys rather than a full census approach, so issues about potential bias in the data become relevant.

Population Reconstruction

In most cases, it will not be possible to source a single dataset that provides an individual-level description of a population so that it can be used directly to define the population of agents in the computational model. Some data may be available only in the form of aggregate statistics while other data may exist as a sample of individuals. Depending on the use of the data in the model, it will be necessary to use the data to create the individual attributes of the agents in the model – to reconstruct a population (Birkin, Turner and Wu 2006). Similarly, attributes of the environment may need to be assigned to entities in the environment. Other data will be used directly in aggregate form. For example, fertility may be determined by an overall probability distribution that defines the probability that a female person becomes pregnant, dependent on individual characteristics such as age and health status. That is, while fertility is not treated as an individual characteristic, it is dependent on individual attributes.

The basic process of population reconstruction is described by Birkin, Turner and Wu (2006). At its core is a genetic algorithm, which evolves an initial candidate population, then evaluates the result according to a 'fitness function' that tests whether the new population has a stronger correspondence to the given data about the real-world population. The process continues until pre-defined criteria are met that indicate that a solution has been found that is 'maximally consistent' (*ibid.*) with the empirical data. It is important to note that in the general case, there is no guarantee that a single such population exists or that it will be found by the algorithm. Nevertheless, statistical tests can be applied to ensure that the solution found is adequate for the given purpose. Depending on the specific research questions, different parameters may be assigned different weights to ensure that a high degree of consistency is ensured with regard to them while others conditions may be relaxed to ensure that a solution will be found and found within a reasonable timeframe.

Verification and Validation

In the construction of a simulation model, it is important to ensure that the implementation of the computer program accurately follows the specifications laid down in the conceptual model. This process of *verification* corresponds to the general computer science problem of reasoning about the correctness of an implementation. In contrast, the *validation* of a simulation model involves ensuring that the model outputs can be taken to represent an adequate forecast of the behaviour of the social system given the input conditions.

The validation of agent-based social simulation models can carried out using a number of different methods. Louie and Carley (2008) distinguish between:

- validation of the conceptual model, asking whether the underlying theories and assumptions are adequate given the purpose of the model;
- validation of the data, asking whether the data sourced as input to the model is of an adequate type and quality;
- operational validation, asking whether the model outputs correspond to phenomena observed in the real-world population that is modelled.

The issues above are not unique to agent-based social simulation but are concerns in many quantitative social scientific investigations. Similarly, the problem of verifying the code is also not unique although perhaps the complexity of the codes involved compared to other social scientific codes makes it a more pressing concern. One question is important in relation to the stochastic nature of the model – do repeated model runs indeed produce the expected behaviour, following the law of large numbers? It is important, therefore, to conduct a *stability analysis* to test that the model code does not introduce outlier behaviour that might throw doubt on the validity of the model.

In general, the methods used to validate an agent-based social simulation model are methods that are drawn from the set of methods familiar to any social scientist working quantitatively and are therefore firmly grounded in disciplinary traditions.

Computational Tractability

Simulation models of real-world populations involve the representation of millions of agents, each with a set of attributes and relationships with other agents and other entities in their environment. The storage requirements of such representations grow with increases in the number of agents but also with increases in the number of attributes of agents and with the number of relationships. In addition, there is often an overhead involved in managing the data and making it efficiently retrievable. The processing power required to manipulate this data and execute the behaviour associated with each agent is equally significant. This means that simulation runs can require large-scale computational resources beyond what is available on a researcher's desktop.

In the past, this has meant that scaling agent-based models to the size of whole populations of individual countries, of regions and, ultimately, the whole world, has been impossible. Thanks to recent developments in the computer industry, this is now changing as the processing speeds and storage capacities of commodity hardware continue to increase. Of the universe of conceivable social scientific investigations, a growing proportion is not becoming *computationally tractable*, that is, the capacity of the infrastructure is large enough to effectively execute simulations and their running times are short enough to make their practical use feasible. In addition, we can see that in the short to medium term, the compute capacity required for a significant number of investigations will be available as a commodity in the form of a single server or powerful workstation. This has the advantage that the development of agent-based social simulation models becomes easier as there is no longer a need to use high-performance compute infrastructures that require specialist skills to operate and program. Such skills are scarce in many areas of science and, given that social scientists do not tend to be heavy users of advanced IT infrastructures, this skill shortage is particularly pronounced in the social sciences. The change in the skill sets required to develop social simulation models is therefore arguably most significant. In the next section, we discuss a study we have undertaken to investigate the feasibility of studying migration in Taiwan using the agent-based approach.

The SimTaiwan Feasibility Study

Migration has been an important factor in the social and economic development of Taiwan since the 1600s. Based on previous work by Lin (2006), we are developing a simulation model of migration in Taiwan drawing on the decennial census and other social scientific datasets. The model consists of individual agents representing the population described by the 2000 Population and Housing Census.

We have developed a simplified model that draws on freely available data from the 2000 Population and Housing Census Statistical Table and other datasets available from the Directorate General of Budget, Accounting and Statistics[7]. The use of these datasets makes it possible for us to run the model in different environments in Taiwan as well as the UK without violating data protection legislation. The code developed therefore includes a population reconstruction element as described above to generate an individual-level population as input to the model.

Once the simulation model has been fully developed and tested to the extent possible using these datasets, we will then run the model with individual-level data available at the Academia Sinica Center for Survey Research. One aspect we are interested in is whether a model developed on the basis of aggregate statistics or a sample of anonymised records can be validated against the full census data. The availability of the full data at Academia Sinica will allow us to do this.

The SimTaiwan model consists complex agents, each with an individual set of attributes describing their social and economic status. As current theories of migration focus on economic variables as possible explanations for migration decisisions, our model also reflects this emphasis. However, we are aware that life-style choices and biographical circumstances are also important factors in individual decision making, so we hope that we will be able to reflect these in future versions of the model. For example, we know from previous analyses of

[7] http://eng.stat.gov.tw/lp.asp?ctNode=1629&CtUnit=779&BaseDSD=7

census data that there is a recent trend for elderly people of mainland Chinese origin to migrate back to their places of birth due to an increased openness shown by the Chinese government to visitors from Taiwan (Lin 2006).

These variables influence peoples' decisions to consider a migration move. We follow the traditional approach to split the migration model into a 'departure' part and the 'destination choice' (Lin, under review). Consequently, the pair part to the individual-level representation of society members is the representation of Taiwan's prefectures and major cities. Data from the 2001 Industry, Commerce and Service Census[8] data and the Labour Force Survey[9] provide the basis for modelling the destination choice part of the model.

Stability Analysis

As agent-based models are stochastic in nature, researchers need to investigate the degree of variation of the behaviour they produce. As mentioned before, the law of large numbers should guarantee that the individual random events within the model produce the expected behaviour when aggregated over millions of agents and large numbers of decision points as the model runs through simulated time. However, because it is impossible to guarantee perfect random behaviour within a computer system, there is always a danger that computer implementations introduce systematic errors.

Consequently, an important aspect of model development is an analysis of the model stability, i.e., an assessment whether repeated executions produce similar behaviour and of how large the variation is. We have run this analysis for our model using resources available within the EUAsia VO[10] of the world-wide EGEE grid. As the code is under development, our aim was to not simply produce a single investigation but to set up an environment in which it is easy to produce these kinds of analyses repeatedly. The availability hundreds of CPU cores within

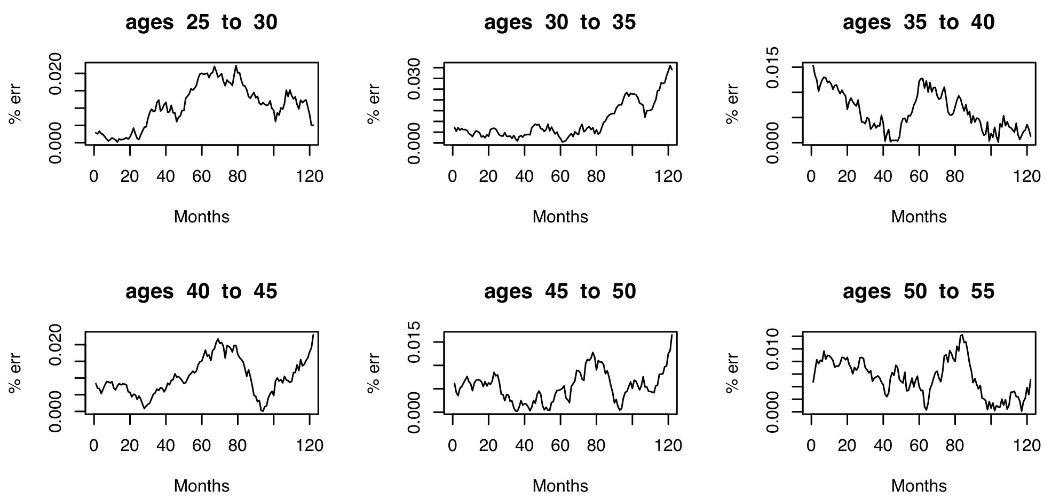

Figure 2: Stability analysis using count of female agents over time for a number of age groups. Based on 80 runs of the model with 1 million initial agents and running over 10 year of simulated time

[8] http://eng.stat.gov.tw/ct.asp?xItem=8260&ctNode=1624

[9] http://eng.stat.gov.tw/lp.asp?CtNode=1607&CtUnit=757&BaseDSD=7&mp=5

[10] http://www.euasiagrid.eu/

the grid makes it easy to produce repeated runs that are executed in parallel, so in the best case, the turn-around time is the time it takes to execute the longest of the runs. Unfortunately, the compute resources on the grid do not meet the memory requirements of full runs, so we had to limit them to 1 million agents in the initial population. This, however, is not a serious issue as the law of large numbers suggests that if model runs are consistent with each other for a smaller population then they will be for the full population.

Figure 2 shows the results of the analysis run with 1 million agents over a simulated time period of 10 years (in one-day steps). The percentage errors for the measure used of less than 0.1% indicate that the model runs are in agreement with each other and not just for the final result but throughout the simulated period of time.

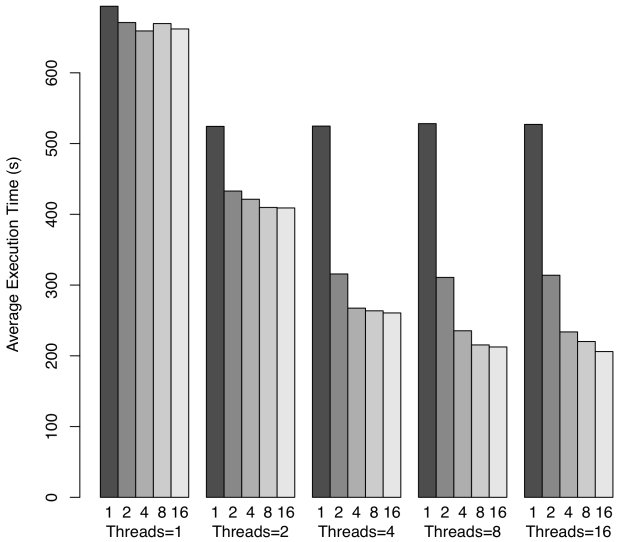

Figure 3: Average execution times dependent on number of worker threads and partitions of the data structures.

Model Performance

In order to assess the performance of the model and its resource requirements, we also conducted a series of runs with varying numbers of agents in the initial population. The initial implementation of the model was written so it would use only a single processor core and running times for even a relatively small population were in the order of many hours. As modern servers have multiple processor cores (up to 48 in large commodity servers, more in specialised machines), we modified the code to utilise them through partitioning the agents into groups and submitting these groups to a work queue that is serviced by the different processor cores of the machine (Voss *et al.*, under review).

Figure 3 shows the average execution times measured on a server with 8 processor cores. The number of worker threads in combination with the number of partitions of the data determine the maximum degree of parallel execution in the code. Threads get mapped to individual cores by the operating system during runtime[11]. The results show that the modified model can make effective use of multiple cores although there are limits to the speed-up achievable. Overall the performance compared to the initial model implementation shows an improvement of a factor of five, making it possible to run much larger simulations on the same machine.

An important question, then, is what the limit is for the size of the population under study. This will be determined both by the time a model run takes to

[11] Note that the last measure was taken with 16 threads. It is sometimes useful to define more threads than there are processor cores because operations such as output to file can mean a thread has to wait for slow disc operations to finish, so the operating system can schedule another thread on the same core.

execute and by the memory requirements. Figure 4 shows the running time for different initial population sizes both in terms of the overall running time (the 'wallclock time') of the model and the amount of CPU resources used (CPU time). For a model run with 20 million agents, the running time is 1460 seconds or 24.5 minutes for a single year of simulated time, so even on the relatively modest 8-core machine we used, simulations at population scale and over dozens of years will be feasible.

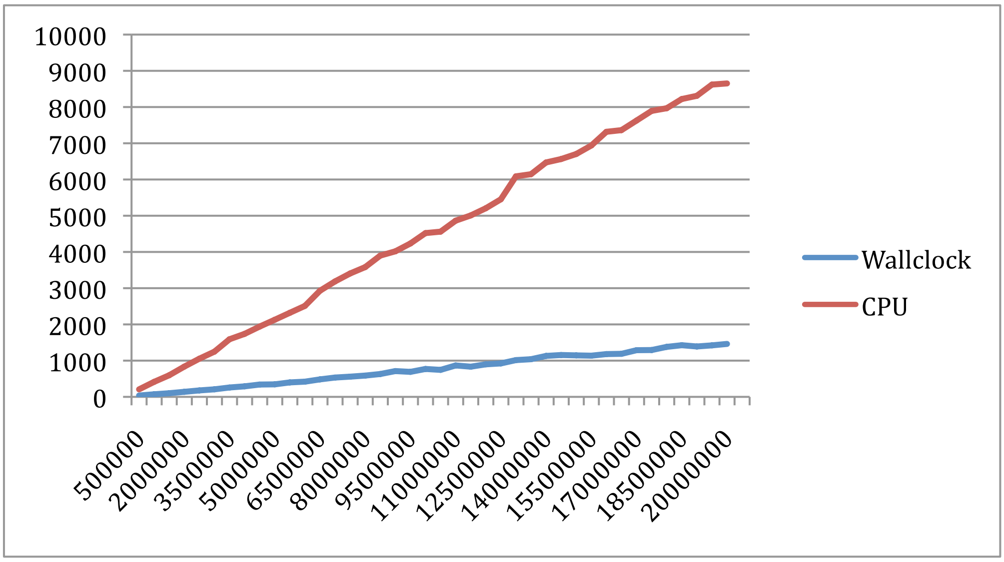

Figure 4: Running times for different initial population sizes (simulated time is 1 year; using 8 worker threads)

Similarly, the growth of the memory usage of the model indicates that it will be possible to run population-size models on the compute resources available. We tested the effect of introducing additional variables into the definition of our agents and the added memory usage is in line with expectations. For a run with 20 million individuals, less than 5GB of heap space[12] are used, again suggesting to us that we will be able to run our SimTaiwan model on the given hardware. While, in principle, it should be possible to run the model on a high-capacity workstation, a server shared by multiple researchers provides better value-for-money as its average utilisation will be higher. While the model used in the experiments was a relatively simple one, we do not expect the memory or CPU requirements to grow beyond the capacity of the server available to us (8 cores, 48GB or memory), giving us the confidence that we will be able to model the full population of 23 million people in Taiwan.

[12] The overall memory usage is determined by the heap space plus space for the program stack, the code and other data structures. For our purposes, the heap space is the dominating factor.

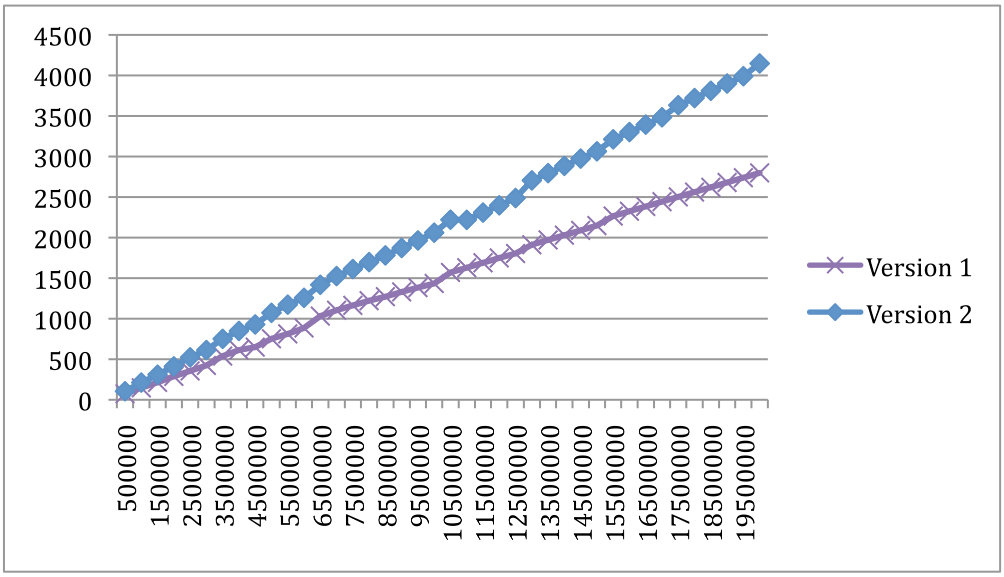

Figure 5: Heap size in megabytes after population initialisation for different initial population sizes

Conclusions

The paper has outlined the principles of agent-based social simulation and argued that an increasing number of social science investigations can now be carried out using this method. Social simulation plays an important role in the UK's e-Social Science programme (Halfpenny and Procter 2010, Birkin *et al.* 2010) and it seems reasonable to expect that there will be a level of interest amongst social scientists worldwide. An important factor in making agent-based social simulation available as a research method or as an instrument for policy makers will be the provision of virtual research environments in which researchers can collaborate to develop and use models. It is important to note that we see the development of social simulation models and a necessarily inter-disciplinary endeavour involving social scientists (possibly from different disciplines) working in collaboration with computer scientists and statisticians. Only rarely will it be the case that a single person combines all the skills necessary to effectively develop the conceptual model, manage the data inputs, develop the model implementation and analyse the results of simulation runs.

We hope this paper has contributed to the development of interest in agent-based modelling amongst social scientists and will help to establish fruitful collaborations with computer scientists to explore the possibilities opened up by the increasing technical capability of commodity hardware. It is important that in this relatively new interdisciplinary field, a corpus of knowledge is generated that will eventually bring agent-based social simulation recognition as one of the modes of investigations available to social scientists and will lead to increased uptake while ensuring appropriate usage.

References

Birkin, M., Turner, A. and Wu. B (2006). A Synthetic Demographic Model of the UK Population: Methods, Progress and Problems. *Proceedings of the 2nd International Conference on e-Social Science*, Manchester, UK.

Birkin, M., Procter, R., Allan, R., Bechhofer, S., Buchan, I., Goble, C., Hudson-Smith, A., Lambert, P., De Roure, D. and Sinnott, R. (2010). Elements of a computational infrastructure for social simulation. *Philosophical Transactions of the Royal Society A*, 368. pp. 3797-3812.

Epstein, J.M. (2007). *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.

Epstein, J. (2009). Modelling to contain pandemics. *Nature*, Vol 460, p. 687.

Gilbert, N. and Troitzsch, K.G. (2005). *Simulation for the Social Scientist*. 2nd edition. Open University Press.

Grimm, V. and Railsback, S.F. (2005). *Individual-based modelling and ecology*. Princeton University Press

Halfpenny, P. and Procter, R. (2010). The e-Social Science research agenda. *Philosophical Transactions of the Royal Society A*, 368. pp. 3761-3778.

Lin, J.P. (2006). The Dynamics of Labor Migration in Taiwan: Evidence from the 1990 and 2000 Taiwan Population Censuses. *Geography Research Forum* 26. pp. 61-92.

Lin, J.P. (under review). Immigration 'Push' and Native-born Labour 'Flight'?: Evidence from Taiwan.

Liu, J.S. (2004). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics, Springer.

Louie, M.A. and Carley, K.M. (2008). Balancing the criticisms: Validating multi-agent models of social systems. *Simulation Modelling Practice and Theory*, 16. pp. 242-256.

Lux, T. and Marchesi, M. (2002). special issue on heterogeneous interacting agents in financial markets. *Journal of Economic Behaviour & Organization*, Vol. 49, no. 2.

Malleson, N., Heppenstall, A. and See, L. (2010). Crime reduction through simulation: An agent-based model of burglary. *Computers, Environment and Urban Systems*, Vol. 34, no. 3. pp. 236-250.

Rahman, M., Jirotka, M. and Dutton, W. (2008). Lost in Reality: The Case for Virtual Safe Settings. *Proceedings of the 4th International Conference on e-Social Science*, Manchester, UK.

Richmond, P. and Romano, D.M. (2008). A high performance framework for agent based predestrian dynamics on GPU hardware. *Proceedings of EUROSIS ESM*, University du Havre, Le Havre, France.

Schelling, T.C. (1971). Dynamic Models of Segregation. *Journal of Mathematical Sociology*, Vol. 1, no. 2. pp. 143-186.

Stroud, P., Del Valle, S., Sydoriak, S., Riese, J. and Mniszewski, S. (2007). Spatial Dynamics of Pandemic Influenza in a Massive Artificial Society. *Journal of Artificial Societies and Social Simulation*, Vol. 10, no. 4. Available at: http://jasss.soc.surrey.ac.uk/10/4/9.html

Topping, C.J., Sibly, R.M., Akçakaya, H.R., Smith, G.C. and Crocker, D.R. (2005). Comparison of a life-history model and an individual-based landscape model of skylark populations affected by a pesticide". *Ecotoxicology*, Vol. 14. pp. 925-936.

Townend, P., Xu, J., Birkin, M., Turner, A. and Wu, B. (2009). MoSeS: Modelling and Simulation for e-Social Science. *Philosophical Transactions of the Royal Society A*, 367. pp. 2781-2792.

Voss, A., Turner, A., You, J.-Y., Yen, E., Chen, H.-Y., Lin, S. and Lin, J.P. (2010). Scalable Social Simulation: Investigating population-scale phenomena using commodity computing. Under review for the 6th IEEE e-Science Conference, Brisbane, Australia.