Assisted Model Building in the Social Sciences using Data Driven Simulation

Peter Lee[1], Ed Ferrari[1], Catriona Kennedy[2], Georgios Theodoropoulos[2], Chris Skelcher[3]

[1]Centre for Urban and Regional Studies, School of Public Policy, University of Birmingham, UK

[2]School of Computer Science, University of Birmingham, UK

[3]Institute for Local Government Studies, School of Public Policy, University of Birmingham, UK

Email address of corresponding author: cmk@cs.bham.ac.uk

Abstract. The complexity of predicting policy outcomes in social sciences demands sophisticated tools to assist in the decision making process. This paper proposes a framework, which integrates agent-based simulations with data-mining to exploit the advantages of both and overcome their respective limitations. The novelty of the framework is the utilisation of "data driven application simulation" (DDDAS) to enhance the reliability of the agent-based model. A prototype architecture is presented along with its initial application to a housing policy case study.

Introduction

In recent years there has been an increasing expectation that public policymaking is firmly rooted in evidence, both diagnostic (what, and increasingly, where are the problems?) and evaluative (what works?). Social scientists have of course always been concerned with questions of these sorts and have employed a variety of research tools in pursuit of their answers. The public policy impetus now demands even greater sophistication and transferability in the methods used.

It is broadly accepted that mixed method approaches to complex and interdependent policy issues offer the best way forward. In this way, researchers have strived to inform the development of quantitative techniques and modelling using better information on behaviour, perceptions and discourse, for example. Nevertheless, the processes by which qualitative insights can inform the development of quantitative models - and vice versa - are less than well developed.

Information on behaviour may be used to specify the broad parameters of models but is less frequently used to model discreet processes. Given that many observable social outcomes involve large numbers of micro-level processes, it could be seen a desirable to be able to model the impact of behaviour at a micro level. The analysis of such processes therefore inevitably involves a large data set of observations and information on a wide set of possible behavioural responses. Whilst data mining techniques can help to observe patterns in outcomes, they are less successful in guiding the researcher to which questions are appropriate ones to ask or how best to improve the model.

Alternatively, agent-based simulations may be used to model social systems and to assist policy decision-makers. “Agents” can represent individuals, groups, organisations, policy-makers etc. Existing work includes geographical decision support systems (e.g. Birkin, 2005), urban planning (e.g. Devisch, 2005) and fire evacuation (Chaturvedi, 2005). A simulation can predict the effects of candidate policies or proposed interventions (or simply the effect of doing nothing). However, such models are based on simplified assumptions and therefore their reliability for the prediction of complex policy outcomes is limited.

This paper proposes a framework, which integrates agent-based simulations with data-mining to gradually adapt the simulation to the patterns discovered in the data as well as ensuring a more focused data mining process. The rest of this paper is organised as follows. The next section describes the policy context in which our methods are applied. The subsequent section introduces “dynamic, data driven application simulation” (DDDAS), which constitutes the basis of our methodology. The fourth section describes the architecture of the proposed system and briefly discusses a case study. Finally, using information drawn from the case study and an initial trial of the model, some concluding comments and implications for future research are presented.

Policy Context

The UK housing market has undergone significant changes in its operation over the past decade. A sustained period of house price growth has led to a concern with growing affordability problems resulting in a review and overhaul of relevant policy in the UK. The government’s review (the Barker Review) of housing and planning policy was in response to the problems of supply and low rates of house building which have contributed to rising prices. The government’s proposals will lead to greater market involvement in planning with competition for residential development from developers and the sale of land being key triggers affecting housing policy. The underlying objectives of the Barker review were to reduce the impact of house price inflation on the economy and the impact this has on the UK’s ability to meet the economic tests relating to adoption of the Euro. Policy outcomes (urban sprawl and development on the Greenbelt through competition for land release) in pursuit of an economic agenda may therefore be in conflict with Regional Spatial Strategies (RSS), which emphasise urban renaissance and the delivery of sustainable communities (ODPM, 2004).

Given this policy background, there will be increasing need for urban policy makers at a variety of spatial scales to monitor the outcomes and impacts of government housing and planning policy as these changes to the planning system take effect over the medium to long term.

Characteristics of the housing market

The housing market is a unique market; one that is characterised by an unusual set of constraints and internal (micro) structures. Because conventional macroeconomic approaches to market modelling have often failed to account for these peculiarities, important structural outcomes of the housing market are often missed (Maclennan and Tu, 1996). In this way, and because the market is subject to complex multi-dimensional behaviour on the part of different agents in the market, Nordvik (2004) argues that it is preferable – indeed perhaps essential –

to give market models a microeconomic foundation. Crudely speaking, the aggregate outcomes of micro-level moves within the market, for example within chains of opportunity (Emmi and Magnusson, 1995), are different from the aggregate outcomes predicted by snapshots of macro-conditions in the market.

Two related problems

In the above policy context, we identify two problems to be addressed by the new tools proposed here.

1. Current models and macroeconomic approaches fail to sufficiently account for the unique, microscopic and behavioural aspects of the housing market, as stated above. In other words, the models are too simplistic and make assumptions that may not be true in every scenario.

2. Second, the multidimensionality of the problem together with incomplete data and significant data acquisition costs introduce insurmountable complexities for the policy analyst looking to answer questions using market modelling techniques. Hence there is a need for assistance in determining what kinds of micro-level data may be especially significant or relevant for policy goals.

Clearly, there are significant problems to the development of policy support tools using conventional methodologies and statistical models. Data needs to be related both to individuals or households and across space. The underlying drivers affecting residents housing and neighbourhood consumption patterns would seem to require an infinite array of variables to account for variations in outcomes by household type and cross-reference this to environmental factors in space.

Data driven simulation: Basic architecture

In the physical sciences, “dynamic data-driven application simulation” (DDDAS) is a method where data from a physical system is absorbed into a simulation of the system (Darema, 2005). There is a feedback system with the following components:

C1.The states predicted by the simulation can play a role in selecting the data to be absorbed. In some cases the physical system itself may be modified as a result of simulation predictions. This is also known as symbiotic simulation (Low, 2005).

C2.The predictions of the simulation are continually adjusted by absorption of new data. The underlying model on which the simulation is based may be revised as a result of data assimilation. This is the “data-driven” component.

AIMSS: A Social Science Assistant

For the social sciences, we are developing a modified DDDAS architecture called AIMSS (Adaptive Intelligent Model-building for the Social Sciences) The basic architecture of the system is illustrated in Figure 1. U is the user-defined model underlying the simulation, D is the meta-data description of what is in the simulation and is used by the manager agent (labelled “Discovery Assistant” or DA) to select relevant data and to suggest possible model revisions. Analogous to the pure DDDAS architecture, there is a symbiotic feedback system with the following components:

S1: Predicted states of the social simulation are used to form database queries and to interface with data analysis and mining tools to inquire whether there is evidence for the predicted state (Langley, 2000). The “predictions” are states that would be expected to exist now if the model's assumptions are true. If insufficient data is available, the DA agent can suggest new kinds of data that are required in future surveys.

S2: Results of database queries may be fed back into the simulation. Persistent discrepancies between the simulation predictions and the results of the data analysis is a trigger for the DA agent to suggest model revisions.

Component S1 addresses the above-mentioned problem of data selection (problem 2 in the Policy Context section). Component S2 addresses the problem of over-simplistic models of micro-level behaviour (problem 1). For the social sciences, component S2 is expected to be an interactive process with human intervention.

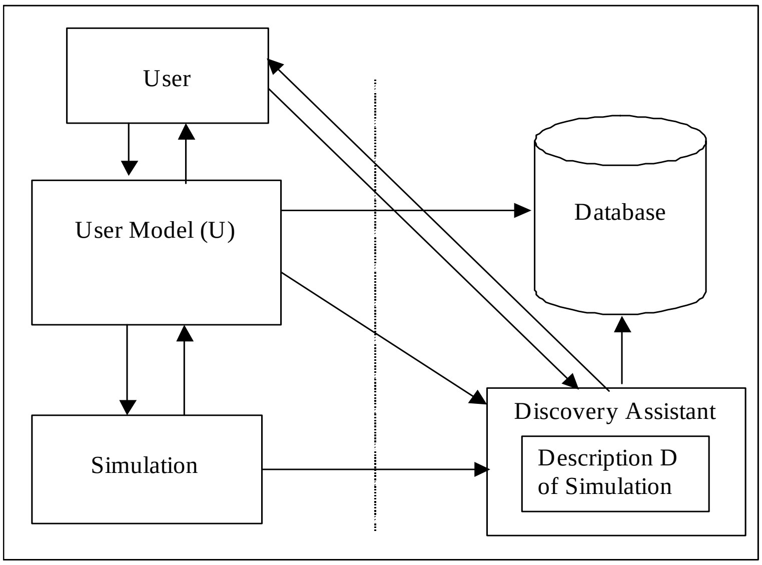

Figure 1: AIMSS Basic Architecture

Case study and current implementation

For our current feasibility study, we are using a database of tenancies in the social rented sector. This data source provides “real time” evidence of lettings made by social housing landlords. The data include stated reasons for the households’ decisions to seek housing, characteristics of the household and of the property. The Housing Corporation’s “CORE” (Continuous Recording) dataset in effect provides a continuous ledger of social housing transactions in certain areas. It will be used as a source of empirical observations of the outcomes of processes in a part of the housing market. This is a source of data against which the states predicted by the simulation (S1) can be tested.

An intentionally incomplete agent model

We have implemented an agent-based simulation in an extremely simplified and abstract environment using RePast (Collier, 2003). The aim is to represent important features of a typical city in a very generalised way. The agent “space” is divided into 4 quadrants: Q1: expensive city centre apartments, quite small, densely populated; Q2: inner city towerblocks, inexpensive, cramped, densely populated; Q3: modest suburb (medium sized with gardens), moderately populated; Q4: wealthy suburb (large expensive houses with large gardens), sparsely populated.

At initialisation, “properties” are allocated randomly to these quadrants with largest number in inner city and city centre. Precise densities can be specified as parameters. A household is treated as a single agent although it may have more than one member. Households are allocated randomly to quadrants initially with varying densities. (e.g. Q1: all properties occupied; Q2-Q4: 1 in 5 vacant).

For this exploratory prototype, we deliberately limited the rules so that we know that they are incomplete. The aim is to discover new rules that can be added to the model thus making it more accurate. The rules are as follows: an agent wants to move only for the following reasons: (1) the current property is not affordable (rent or mortgage too high) or (2) the current property is overcrowded. The first rule overrides the second. The agent will move to an overcrowded place if it is the only one that is affordable. However, it will still not be “happy”, i.e. it's critical requirements are not met. Currently these critical requirements are only affordability and space. As long as an agent is not happy it will keep on attempting to move.

Component S1: Predictions. Since the simulation is abstract, its predictions will be in the form of very general statements that should be true of typical cities. For example: “there are some middle income households in the inner city”, or “there are many vacant properties in the city centre”, where “some” is “at least one” and “many” means significantly more than in other quadrants. For example, our simulation “predicted” that moves in the inner city quadrant are very frequent and households are mostly unhappy, while city centre luxury apartments tend to have the most vacancies. Such predictions can be verified or refuted by analysing the data, assuming that the data is “typical”.

If we run the simulation manually with no DA agent, such general predictions can only be recognised by humans interacting with the simulation and recognising patterns in the graphics animation (or by looking at charts produced by the simulation). In this case, it would be reasonable to have fully interactive control of the data mining interface and the DDDAS architecture would be manual. However, for more complex and realistic simulations (for example in which interaction with neighbourhoods play a key role), human pattern recognition may miss some important emergent properties of the model. Its visual results are constrained by the attributes that are selected at design time (e.g. what attributes should be displayed and what should be included in statistical analysis?).

Therefore the automation of component S1 of the DDDAS architecture should as far as possible complement human visual pattern recognition by looking for relations between attributes that may not have been included in the graphics at design time. Of-course it may also be possible to generate new kinds of visualisation of the unexpected patterns. For this purpose we are currently investigating the application of data mining algorithms to the states of the simulation in order to find interesting predictions that are emergent properties of the model.

Component S2: Model revision. The DA should use the results from component S1 to select the relevant data and use the data analysis tools to check if the simulation predictions can be verified. So for example, if it is really true that small low-income households (1-2 persons) in the inner city areas are happy and tend not to move unless the household gets larger, the DA can query the household sizes in moves within and to/from the inner city. If for example the opposite appears to be true from the data (small households move more frequently), there is a need for model revision. To revise the model, it is necessary to find new rules that can better predict the observed data using predictive data mining tools. This will normally require integration and analysis of multiple associated datasets, but initially we are investigating the limits of using a single database. The DA can suggest that the new rules be added to the behaviour of agents or as appropriate the dynamic changes in the environment (such as for example, ageing or demolition of properties). Automation of this process may be possible, in which case the DA would directly add or replace rules in the model or modify other parameters in the simulation (such as density of properties in an area).

Conclusions

The DDDAS architecture proposed in this paper provides more than just a data mining or a simulation capability on its own. The simulation predictions can themselves be unanticipated and can lead to new ways of connecting data together and thus to novel uses of the data mining algorithms. This can in turn lead to new ways of summarising the data and improving the accuracy of the initial model.

This architecture assumes that the system can interpret the simulation and that it can use this same description to identify relevant data sources, which may involve multiple datasets being integrated together. This allows the system to associate attributes and values in the data to attributes and values in the simulation states. To do this, an ontology is required for the description of social science datasets.

References

Emmi, P. C. and Magnusson, L. (1995) “Opportunity and mobility in urban housing markets.” *Progress in planning.* 43: 1–88.

Maclennan, D. and Tu, Y. (1996) “Economic perspectives on the structure of local housing systems.” *Housing studies.* 11: 387–406.

Nordvik, V. (2004) “Vacancy chain models: do they fit into the economist’s toolbox?” *Housing, theory and society.* 21: 155–162.

Pat Langley (2000) The Computational Support of Scientific Discovery International Journal of Human-Computer Studies Volume 53, Issue 3, September 2000, Pages 393-410.
Malcolm Yoke Hean Low, Kong Wei Lye, Peter Lendermann, Stephen John Turner, Reman Tat Wee Chim and

Surya Hadisaputra Leo (2005) An Agent-based Approach for Managing Symbiotic Simulation of Semiconductor Assembly and Test Operation. Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2005), Utrecht, The Netherlands, July 2005.

Frederica Darema (2005) Grid Computing and Beyond: The Context of Dynamic Data Driven Applications Systems. Proceedings of the IEEE: Special Issue on Grid Computing, volume 93, number 3, March 2005, pages 692—697.

Mark Birkin, Haibo Chen, Martin Clarke, Justin Keen, Phil Rees and Jie Xu (2005), MOSES: Modelling and Simulation for e-Social Science. First International Conference on e-Social Science, Manchester, UK, June 2005.

Oswald Devisch, Theo Arentze, Aloys Borgers and Harry Timmermans: An Agent-Based Model of Residential Choice Dynamics in Imperfect, Non-Stationary Housing Markets. Computers in Urban Planning and Urban Management (CUPUM'05) London, UK, July 2005.

R. Chaturvedi and S.A. Filatyev and J.P. Gore and A. A. Mellema (2005) Integrating Fire, Structure and Agent Models, Workshop on Dynamic Data Driven Application Systems at the International Conference on Computational Science (ICCS 2005), Atlanta, USA, May 2005.

Nick Collier (2003) RePast: An Extensible Framework for Agent Simulation. http://repast.sourceforge.net/