ESRC Microsimulation Seminar Series

Seminar 4: 'Bridging the Gaps; Setting the Agenda'
University of Sussex, 11th September 2009

Microsimulation in the UK: the current state of play

Paul Williamson

[EDITED TRANSCRIPT]

Introduction

[1] In previous presentations we have heard about microsimulation in Australia and Canada. In this final presentation I want to try and give an overview of microsimulation in the UK. It's necessarily going to be a bit partial but I'll try and be as comprehensive as I can.

[2] This table lists the main currently active models that I'm aware of that exist in the UK, or that are somewhere in the being prepared/drafting stage. As you can see there are a lot of models by lots of different groups of people, quite a few static models and dynamic models, some models which are either entirely agent-based, or that integrate agent-based modelling into some kind of microsimulation framework. You can also see the wide range of applications that these models have.

If you don't appear in this list and you think you should, or if some information about one of your models is wrong, please let me know. But that's my attempt at a quick snapshot to try and underline the fact that there is a great diversity and indeed expanse of microsimulation in the UK. If I'd been giving this talk as recently as 10 years ago the list would have been much shorter, with a much narrower range of applications.

To review each of these models in detail would take far too long, so instead I'm going to try and pick out some of what I think are the UK modelling contributions which could be described as world leading - in other words, not just 'run of the mill' models (with apologies to some of us, including myself), but models where probably we're doing things that the rest of the world is sitting up and watching and thinking yes we'd like to do that as well.

As it happens most of these models, all of them in fact, were presented earlier in the series of seminars of which this is the final one. Therefore , for this talk I've just lifted a few slides from their various presentations and I'm going to attempt to summarise what I think are the key points of what people said. Therefore, although my slides have a University of Liverpool, geography logo on the side, please note that's just my template; in the main I'm talking about and presenting other people's work.

EUROMOD

[3] The first world leading model I think that I want to talk about is EUROMOD. EUROMOD is a static tax-benefit model, a number of which we heard about those this morning, in a UK government context. EUROMOD differs in that it is based in an academic centre, but what's world leading about it? Well, it covers currently 15 European Union member countries and 4 non member countries, giving a total of 19 different countries, all of which are integrated into the model. And it's currently under development to span 27 European Union countries. EUROMOD also provides the platform for LATINMOD – a static tax-benefit model which is going to cover 5 Latin American countries. EUROMOD has also been used to develop a static tax-benefit model for South Africa. So, in comparison to any other static tax benefit model in the world, EUROMOD is by far the most widely implemented, has the largest user base, and covers the most countries. So, in terms of breadth of use, EUROMOD is clearly in a league of its own.

[4] How did it get to be so widely used? Well, one of its key selling points is that is has been designed precisely to be cross national. Everything about the model has been designed to be generic, so no information about a country's tax system is hard-coded. As a result there is no hacking into SAS code or whatever to try and re-write tax policy rules. Instead the model is interface via a series of Excel spreadsheets in which you can create rules, create variables, link rules and variables together, order the way in which rules are implemented and so on. Having said that, it should be acknowledged that an Excel interface isn't easy to learn in the sense of being able to sit down and learn it in an afternoon, but it's easy to learn in the sense that within a week you can be up and running and developing your own tax benefit model for a new country.

The second selling point for EUROMOD is the ease with which others can gain access to it. For example, EUROMOD is the only tax benefit model in the UK you can get your hands if you're not the producer, although you would of course have to obtain licensed access to the underlying data – easy for academics via the UK Data Archive; potentially more challenging for others.

Spatial microsimulation

The second area that the UK might possibly think of itself as well leading in is in terms of spatial microsimulation. Justine McNamara talked about how NATSEM is catching up with the UK in this respect and, in many senses, actually now outdistancing the UK in terms of actual application on the ground in a policy environment. But in terms of the origin, development and validation of spatial microsimulation applications the UK is arguably still at the forefront of international endeavours. I'll endeavour to support this case with a few illustrations.

[5] *Williamson*

This slide summarises various aspects of my work in this area. The first question to be answered is 'what is spatial microsimulation?'. The bedrock of spatial microsimulation is the creation of some synthetic, spatially detailed micro data, where spatially detailed in this case typically means sub-regional or even sub-city. These data need to be synthetically generated because geographic coding is normally the first thing stripped off survey data to protect respondent confidentiality. In the UK we have experimented with various approaches to creating this synthetic, spatially detailed microdata, and in essence they nearly all equate to re-weighting survey data, one way or another, to fit known local area constraints (typically derived from the Census). Once created, estimates can be derived from synthetically generated spatial microdata of distributions of interest, such as the count of socio economic group by household composition and so on. In terms of validation of these estimates, the bottom left-hand table reports the % of times, out of 100 stochastically derived estimates, that the estimated distributions are statistically significantly different from the observed distribution. As the table shows, despite the fact that these distributions weren't used as constraints in the re-weighting process, they've been estimated accurately for small areas, for different types of areas ranging from 'middling England', easy to estimate because it's very similar to the national (and hence survey) average, through to the highly atypical 'deprived urban area'. In fact this histogram plots each of these areas types on distance from the national average. As you can see the deprived urban area (B4) is highly atypical, so it should be hard to estimate. Even so the synthetically generated microdata clearly provide reasonable estimate of the unknown interactions. This top right-hand graph shows another comparison of a score calculated from synthetically estimated microdata to the equivalent score calculated direct from available Census data. As you can see, the fit is very good. In short, we have managed to demonstrate that there is a technique for re-weighting survey data down to local area constraints which works well enough to act as a basis for spatially detailed microsimulation.

[6] *Anderson*

Ben Anderson, based at Essex, has done some work, taking exactly the same kind of approach – in other words, taking different sorts of surveys and re-weighting them to fit local area constraints, typically derived from the census. This map is of predicted expenditure on telephony, derived via re-weighting expenditure survey data on telephony to census area constraints, yielding an estimate of the amount of money households spending on their phone bills. The plot on the right-hand side shows how much British Telecom thinks households are spending on telephones in comparison to the estimates and, as you can see, the fit is reasonably good fit given the spatial precision of the estimates (UK wards).

[7] A second example of Ben's work is the re-weighting of Time-Use survey data, yielding an estimate of the amount of time people spend at work. The estimate is again mapped, and in this case plotted against census returns in which people say how many hours a week they work. This validation exercise again provides some confidence the process of synethetically estimating spatially-detailed microdata. Indeed, as Ben has shown, so long as you use the right re-weighting constraints, constraints that actually are predictive of the thing you're trying to estimate, the technique works well. If you re-weight using variables with little or no predictive power – such as eye colour – on the other hand, the technique will clearly not work well. In other words, you have to use the right constraints in the survey reweighting process.

Ben is currently working on taking successive censuses, extracting counts for spatial units that are stable over time, and then projecting them forwards. For example, you could take the age structure and then project that forwards over time, to give an estimated age structure for an area in 2020 given how it's changed over the last 30 years. The plan is then to take these 2020 constraints use them as survey reweighting constraints to produce synthetically estimated microdata for 2020.   This is, of course, very much like the classic static microsimulation model approach to updating or uprating, but stretched out over a longer time-period and applied to smaller spatial units.

Separately from this, Tony Lawson, under Ben's superivions, is doing currently creating a regional dynamic microsimulation model which will be run forward over time, but then aligned using these rolled forward census small area constraints to produce small area estimates.

[8] *Ballas*

A third example of spatial microsimulation is Dimitris Ballas's work on happiness. Recently some economists have decided that income is no longer a good measure welfare, and have argued that we should be thinking about some broader, more subjective term, such as happiness. The argument is that,you know people are happy if they say they are, and it is this measure, rather than some cruder measure, such as income, that should be used to help inform policy formulation. So Dimitris has taken the British Household Panel Study, which includes questions on mental wellbeing, such as "how happy are you" and, using the same approach as Ben and I, re-weighted BHPS to a set of census-based local area constraints - in this case at district level. Then, in the same way as Ben, Dimitris has projected forwards from 2001 to 2011 the district-level determinants of happiness, such as age, and come up with a projected estimate of the future % of persons who will self-report as being more happy than usual. And as you can see from his results, it would appear that Scotland's about to get a lot more cheerful (LAUGHING) …

[9] Dimitris has also analysed the BHPS to find out how key demographic events are associated with happiness. In this slide I've just picked out from his results the top 5 positive and negative associations. A clear potential next step is to integrate this information into a classic dynamic microsimulation, in which births, deaths, marriages and other demographic life events are modelled, and use this to investigate how people's happiness changes over their life course.

[10] *Bell*

A fourth example of spatial modelling, a lot of which is going on in the UK at the moment is OPERA, which stands for the Older People's Resource Allocation Model, created by David Bell from the University of Stirling. In fact, having listened to his presentation I think this is a bit of a misnomer because in reality David follows myself, Dimitris and Ben in re-weighted a baseline static microsimulation model to whatever constraints you're interested in for the problem in hand. David has used this approach in a number of applications funded by the Scottish Govt, Welsh Assembly and so on. One example, presented here, is an investigation, for the Scottish Government, of what the distributional impacts would be of purusing their favoured policy of changing from a property tax (council tax) to a local income tax. As presented here, the analysis looks no different from any other classic static tax benefit model application. In fact, David did use spatial reweighting to pull that out to what the impact would be across districts in Scotland; for example, to investigate whether the Highlands and Islands lose out compared to cities under this proposal.

[11] A second model that David has developed focuses on the changing care costs likely to be associated with the treatment of dementia over time. To do this he took a static model, imputed dementia rates given personal characteristics, then re-weighted it forwards to future age, sex and other totals. He then mapped on to that the estimated projected cost of care for dementia-sufferers in various circumstances, disaggregating those costs by different elements of expenditure.

The key thing to highlight about OPERA is that its applications have been directly influential on policy formation, funded as they have been by the policy makers concerned.

[12] *MOSES*

My final example of spatial microsimulation is MOSES. MOSES is a dynamic spatial microsimulation model. It starts off, as with the previous examples, by generating an initial, synthetic, spatial-detailed population micro database. However, it then ages these people forwards not by reweighting, but by the application of annual probabilities of mortality, fertility, nuptuality etc. A key point to note is that MOSES, unlike other dynamic microsimulation models, uses a 100% population representation so that sufficiently robust and distributional estimates can be derived for each small area.

As well as being spatial and dynamic, MOSES is based on a computing-grid architecture, which involves the pulling together of data sets stored on one computer and analytical software based on other computer and mapping tools based on a third computer, and delivering them in an apparently seamless web-based front-end to the user's desktop. Hence this diagram shows the model architecture that allows you to use different algorithms for re-weighting the survey data to get you to your base data etc. In other words, the interface is designed to enable you to undertake model development in a 'plug-and-play' environment. To my knowledge this is clearly world leading, as all of the microsimulation models I'm aware of internationally have a fixed, one-system, computer architecture.

[13] *NCeSS*

Tangentially linked to microsimulation I should note in passing that Mark, because of his interest in developing computer infrastructure for social simulation, has just landed a grant, £2 million, 18 person years, to set up an e-infrastructure for social simulation, and he's currently casting around for ways of spending that money and those person years. One of the things he wants to do is embed MOSES into web-based portal and then make it open to a wider user base so that people remotely access and use the model. But this infrastructure is also meant to provide a portal for people to exchange information, ideas and so on. In our discussions this afternoon, therefore, we might want to think about some proposals we would like to make to Mark about how he might spend some of his resources on facilitating improved communication amongst the UK microsimulation community.

Integration of ABMs

Student migration in MOSES

[14] OK, so the UK arguably leads the work in cross-national and spatial microsimulation. A third area where the UK perhaps leads the world is in starting to try to integrate microsimulation and agent based modelling. For those of you who don't know, agent based modelling is modelling where the individuals and the model follow a set of rules like if it's hot I'll go outside and if it's cold I'll stay inside, that kind of thing. Note that, in its purest form, there will be no empirical basis to these rules; rather they will be theoretically derived. The agents within the model then interact with each other, each perhaps using a different sets of rules; possibly they might learn from one another, copying the rules used by more successful agents. And then you see what emerges from the model in terms of behaviour. The classic application of an Agent Based Model is a stock-market trader model, where the traders (agents) have rules, such as "I'm going to sell if the market's gone up 3 days in a row" or "I'll sell if the market has gone up 2 days in a row". The agents then interact, buying and selling shares, learning from each other. And the result is the emergence from the model of the classic stock-market booms and busts even though the model involves not a shred of empirical data.

In contrast, as we know, microsimulation is all about data, surveys, empiricism. Is there, therefore, any scope for these two approaches to meet in the middle? Well in MOSES, the dynamic spatial microsimulation model, attempts were first made to model student migration using the conventional type of empirically-based approach, which is summarised in this slide. Transition probabilities for migration were derived from available data, given age etc. Individuals selected stochastically as experiencing a migration event where then allocated to a new zone, taking account of housing vacancies, using a spatial attraction model, empirically fitted such that shorter moves are more likely than longer moves, and so on. Remember that MOSES represents 100% of persons and housing stock). Now, this modelling approach was highly sophisticated, highly complex, data heavy and – guess what - it didn't work; at least for migration flows involving incoming and outgoing students. In fact, it led to exactly the same problems that any small area population projection model, micro or macro, has in realistically representing student migration behaviour. This is mainly to do with the poverty of data on student migration in the UK - it's simply not very well captured in official statistics. As a result the models ended up leaving students in situ after graduation, leading to a large, ageing, bulge of people in a place where they should, in fact, have moved out of. Conversely, for student incomers, the model over-disperses new arrivals – perhaps in part because of the failure of in situ students to move out upon graduation, leaving no housing vacancies for incoming students to occupy.

[15] What Mark Birkin and Belinda Wu have done is to tackle this problem using an agent based model. For most ages the conventional, empirically-based approach to modelling migration was retained, but for students they applied instead a rule-based (ABM) approach. As a result, in the model, a set of rule is applied which treats differently the likely behaviour of undergraduates and postgraduates according to their year of study (first, second or third). Underlying all of these rule sets is a basic principle which says you're going to look for some housing, the housing is going to be near the university, it's going to contain other students, and then when your course has finished you're going to leave.   This type of behaviour sounds intuitively reasonable, but note that is based not upon empirical data, but upon a theoretical understanding of stuent behaviour. The result of applying this ABM approach to student migration is illustrated on this slide. The observed distribution of students is on the left, the disbribution projected by the empirically-based dynamic microsimulation in the centre. This shows that the empirically-based model , under-concentrated students through failing to move out enough people. In contrast, the outcomes from the agent based model on the right, perhaps over-concentrates students, but overall proves to be far more successful in capturing the real life behaviour than the conventional empirically-based approach.

Modelling individual consumer behaviour

[16] A second example of integrating an ABM into microsimulation is that of Kirk Harland and Alison Heppenstall, both at Leeds. This is work that's just started on trying to model consumer behaviour, again starting off with the creation of a synthetic spatially detailed micro data set. The model then goes on to represent people as customers looking for somewhere to spend money. It will therefore sets up retail outlets, and then either empirically or possibly through rule based mechanisms, model their shopping behaviour to generate flows, starting off, if I remember rightly, with petrol stations, and later adding in a road network to help influence choice of retail outlet. The reason they're doing this is that conventional approaches to this kind of retail demand modelling use aggregate models, and they reckon they can do better by breaking things down a bit into individuals and their individual characteristics, modelling different and heterogeneous behaviours for different sorts of people. So for example the elderly might have totally different retail patterns to the young.

ABM v Microsimulation

[17] In this seminar series we have previously listened to a talk by Edmund Chattoe-Brown from the University of Leicester who is an agent based modeller. I asked him to come along and talk about the possible links between agent based modelling and microsimulation modelling, because I've always argued that it's a continuum from a 'pure' microsimulation model, totally fitted to empirical data, on the one hand, to a 'pure' agent based model, entirely based on theory driven rule sets on the other, with somewhere in the middle, models based on more or less empirically informed rules when either the available empirical data or theoretical understanding is clearly less than perfect. Anyway, that's my theory. Edmund came along and argued that agent based modelling is totally different to microsimulation and that 'never the twain shall meet'. I think I'm going to continue to disagree and argue that the MOSES handling of student migration is a good example of why I disagree. But he did make some useful points which are worth rehearsing quickly now.

Edmund's major challenge to microsimulation from an ABM point of view was that prediction doesn't equal explanation. Just because a model predicts well in the sense of managing to reproduce current survey distributions this doesn't mean that the model is actually correct. Dangers include over-parameterising to the particular survey data you've got; aligning to projected future totals in attempt to get around the fact that the underlying dynamic processes aren't working correctly; ignoring non-linear, second round and exogenous effects – all practices which it must be confessed routinely occur in microsimulation models. And yet we still claim that the model results are potentially useful in some way. Edmund concluded by giving us quite a challenge to think more carefully about devising methods of validation which try and address some of these problems a bit more head on.

Global models

So that's cross national and spatial microsimulation and the integration of agent based models into microsimulation models.

[18] The fourth area where the UK leads the word in microsimulation is in terms of scale. I think we can safely claim to be the home of the world's largest microsimulation model, which is the global epidemic simulator created by Neil Ferguson, Wes Hinsley and others at the Imperial College. This microsimulation models the entire 6.5 billion people of the world. Its aim is to model the spread of infectious diseases through direct contact. This slide shows some screen shots from their model as currently developed. How is the model implemented? Well the use LANDSAT population data, in which the global population is estimated by age and sex for each 20 km grid square on the earth's surface. That's the yellow squares at the top of the top image on this slide. To model disease spread, individuals are randomly sampled from each 20km grid square and randomly brought into contact with a person somewhere else in the world. The probability of contact is driven by a distance decay function so you're very likely to come into contact with somebody else in your own 20 km grid square, moderately likely to bump into somebody in the next grid square along and decreasingly so going out from there. And then when you get a bit further away, for computational reasons, the rest of the world is treated not as 20km grid squares, but as 320 km grid squares (the blue squares in the top image). So far the main development work for this model has concentrated on devising a software architecture capable of modelling the global population whilst still delivering reasonable run times. The solution adopted involves the splitting of the computational task across multiple computer processors, with each major world region modelled on a separate CPU.

What they're working on at the moment is improving the empirical realism of the modelled flows between countries, local areas and so on, using data on plane routes, border crossings and so on. Another planned development, already implemented in previous smaller scale modelling work (small as in 'only the entire population of the United States) is to build in schools, workplaces, places where people meet and thus have points of contact. Note that both this earlier work, and the subsequent global model, work in time-steps involving fractions of a day, rather than the annual time-stpe used by most conventional discrete-time dynamic microsimulation models.

MAP2030: Care for the Elderly

[19] One final area where the UK might be leading the world is a collection of projects collectively known as MAP2030, a series of five related projects trying to model the needs and resources of older people. This project comprises a series of work package, illustrated here, and most of which involve microsimulation somewhere. I want to pick out 2 of them, mainly because those are the 2 that came and talked to us, so I actually know something about them. But I also happen to think that, in themselves, both of these contributions are world leading in different ways.

[20] *Modelling extended kinship patterns*

The first of these contributions is work being undertaken by Mike Murphy at the LSE using SOCSIM. SOCSIM is a microsimulation package originally written in Berkeley, but which I know Mike has tweaked fairly substantially since then. SOCSIM is a closed, discrete-time, dynamic microsimulation model that has been designed in particular to keep track of kinship relations. So, once a person leaves a household the model keeps track of them, which enables you to do things like start counting surviving stepchildren and the like.

The major problem faced in modelling extended kinship patters is that there are no data on kinship networks; no government survey which asks 'how many grandchildren have you got?', 'how many stepchildren have you got who don't live in your house?' etc., as they are all household-based surveys. So the first thing that needs to be done is to synthetically estimate the current kinship patterns. To do this, Mike starts the model in 1751 with 40,000 people. He then applies historically plausible fertility, mortality and nuptuality rates, and rolls that forward through time. Of course the kinship patterns captured in the initial years of the model run are unusable as the model is just getting warmed up. But by the time it hits the 20th century the patterns of kinship relations, the demographic patterns and so on, contained in the model actually look like those observable in the real world. In these illustrative results, the dark line plots the model estimates. Reading across the bottom of the graph is a person's age. On the vertical axis is how many surviving grandparents, grandchildren, siblings, children and so on they they have. The green dots are the survey estimates from the one, small-scale, survey that has been conducted on kinship patterns. As you can see, the fit of the model-based to survey-based estimates is remarkably good. So starting back in 1751, and rolling the model forward to the present, generates a set of plausible kinship networks.

[21] Having run the model to the present you might as well leave it running into the future, again using plausible demographic rates. The result is a set of projected kinship patters for 2050. This graph, for example, shows how, over time, the number of surviving grandchildren and grandparents you are likely to have a given age will change – a reflection of the complex interplay of changing fertility and mortality rates into the future. Why would you care future kinship patterns? Well, in a policy context the government might think that grandparents can be expected to step into the child-care breach caused by 'back to work' campaigns for lone parents; or that surviving children can be called upon as the primary source of care for their ageing parents.

Dynamic cohort modelling

[22] The second MAP 2030 project I want to talk about briefly is CARESIM. This is a model run by Ruth Hancock at the University of East Anglia. As currently constituted, CARESIM is a dynamic cohort microsimulation model of the cost of long term care for the elderly. In that sense it's not world leading, and I don't think Ruth would claim that – although it is clearly UK leading. But what Ruth is doing at the moment is, I think, potentially world leading. What she and her team are doing is turning their attention from the current 65+ cohort to the following, 45-64 cohort **[PPT]**, retro-fitting pension and retirements data not captured in the survey via imputation **[PPT]**, and then rolling the cohort forward. As a result they can now analyse either the whole population 65+ as far forward in time as 2027 (now that they've estimated the missing retirement and pensions history of current 45-64 year olds); or they can focus their analysis on the cohort aged 85+ as far forwards as 2047 **[PPT]**. And this kind of serial cohort extension is, I think, relatively novel.

Conclusion

[23] That concludes my quick romp through some of the key highlights of microsimulation in the UK, but as I showed you at the beginning there are lots of other things going on. For example, I have deliberately omitted reference to the government models not because they're not necessarily world leading (in the case of the GENESIS code generator for PENSIM2, they clearly are), but because I knew we'd be hearing about them this morning so that's saved me time.

In conclusion, microsimulation in the UK is burgeoning; it's world leading in a number of fields which I've enumerated; there are some challenges which I think we face, some of them obvious ones, faced by all models, such as maintenance, updating, upgrading, validation – exactly the kinds of things we've been hearing about pretty much all morning.

There are, however, some other challenges I would like to highlight:

Increased collaboration – there are currently 6 dynamic microsimulation models in the UK that exist or are under development and, a bit like we've just been hearing about for Stats Canada, a lot of the work is the same work. The application's different, some of the algorithms will be different but a lot of the basic stuff is the same. So maybe there's a bit more scope for collaboration.

Increased user base – Excluding the commercial traffic microsimulation model S-Paramics (which is arguably an agent based models anyway), and EUROMOD, in the UK the user base for every single microsimulation model, academic, commercial and government in the UK I would say is very narrow compared to, say NATSEM and Stats Canada. So I think we need to think a bit about that perhaps as a community.

Academic models – As academics I would say that our policy impact collectively has been pretty low. This isn't to say that our modelling efforts aren't policy relevant – they clearly are. But other than in a very few select cases (TAXBEN; OPERA; CARESIM), we have had nowhere near the level of policy influence enjoyed by, for example, NATSEM. So perhaps we need to think more about how we engage with the policy making community.

[END OF RECORDING]