ESRC Microsimulation Seminar Series

Seminar 2: 'Adding behaviour'
London, 6th May 2009

MODELLING INDIVIDUAL CONSUMER BEHAVIOUR

Kirk Harland

[EDITED TRANSCRIPT]

Thank you, as you have probably gathered, I'm Kirk, not Alison, Alison's at the back, she has made the journey with me today. We both work on a project called Modelling Individual Consumer Behaviour which deals with some of the getting the hands dirty issues that we've just been hearing from Edmund.

Unfortunately, the project is just starting off so we can't deal with some of the major issues that we've just been discussing, but the discussion we've just been having is quite a nice setting for this presentation, so maybe we can take some input on the ways that we should go in the coming months.

Today we're going to talk about the project as an overview, where we're at, the motivation for actually undertaking this project, and I'm going to give you a rundown on some of the results that we've just been getting from the first stages, how the model's set up and the framework it's going to operate within, and then the future research, the bits that are coming which are really the nitty gritty bits.

So the motivation really is looking at retail models, they are very mature, it's a mature environment, but they tend to work on aggregate populations, which don't give any, or pay very little attention to individual decisions. And the whole environment, social environment has proved much more complex in recent years, internet shopping, multi purpose journeys, the pace of life has just speeded up and the pace of decision making has speeded up. There's increased data and increased computational power which allows us to maybe increase the complexity of our models to take account of some of these changes in the social world.

So recent work into agent based modelling, which actually came out of Alison's PhD thesis, took a retail model and disaggregated it into three layers, at the bottom you had an agent based, if you like, supply side to the model which each petrol station, this was a petrol station or fuel supply context, so each of the agents in the supply side was an actual petrol station which employed different rules to try and maximise its profits if you like, its pricing strategies. The demand side was still represented by the spatial interaction model and the distance travelled within that spatial interaction model was represented by a very simplistic travel network which used Euclidean distances between population weighted centroids of zones. So the main problem with this model is that the consumers are still aggregated. It works quite well, it works very well producing very interesting results but this area now is what this project's trying to address, trying to look at introducing individual consumers.

So this is a very basic project plan looking at what we regard as the milestones for the project as it progresses, and to give you an idea of where we're at and how early on we are, we're about there, we're sort of just entering this red one and we're just writing up some results from this first stage, so we've still got a long way to go.

Looking at how we envision the model, well how it will all hang together hopefully at the end, looking to create a population data set, an individual population data set, looking at just pure attributes, no behaviour in there, trying to synthesise an individual population from aggregate data.

Then we go to look to other surveys, consumer behaviour data sets, and derived rule sets that will hopefully incorporate some form of behaviour within a behavioural framework within the agents. Now the question mark is this bit here which is going to be one of the tricky bits, joining them two together, that's one of the nitty gritty bits that we've got to come yet and the question that, if anybody has any input or advice on how we can, or anything we should look at in order to accomplish that, it would be much appreciated.

These two data sets when they join together they form our individual consumer agents which will replace the spatial interaction modelling layer and then hopefully if we have enough time, we'd like to introduce a realistic road network, because in a retail model transportation and travelling is quite an important factor. So we'd like to introduce a little bit more realism than that Euclidean distance between the population weighted centroids if we can, if we have time.

The original petrol stations which actually can be reconfigured to represent schools, retail outlets or any form of supply side location, will remain the same.

So phase 1 requires us to generate the population and we need to be able to generate robust populations but that are really tailored slightly to what we're looking at, the sector that we're looking at. It's no use producing a population for a health study and then applying it to retail modelling because the attributes that we're interested in are probably completely different.

So in this frame we're looking at or examining three different ways of synthesising the population. These are deterministic reweighting, the conditional probabilities model and combinatorial optimisation, namely simulated annealing.

A bit of a description of what each of these are. Deterministic reweighting operates in two stages really, the first stage weights the likelihood that a person from a sample population is likely to appear in a zone. The second stage re-weights that derived weighting to represent the actual populations against the populations of that zone. It's very quick to execute which is a good thing. You can tailor the model using the order of constraints to incorporate geographical inconsistencies. The example I'm using here is an economically active person in the city centre, maybe less likely to own a car than in the suburbs and therefore less likely to make a multipurpose trip. This model allows us to get some of that tailoring into the population which again is an added bonus. But that can also be detrimental because determining the order of the constraints to bring that tailoring in requires more pre-run model set up and lots more analysis, it takes a lot more time. If you need to produce quite a few populations, and again there's limited resources on this project, this can be quite a problem.

The second method that we've been looking at is conditional probabilities model. This, basically, is the synthesis model introduced in the late 1980s by Mark Birkin and Martin Clarke, and to be quite honest with you Mark's talked me through this quite a few times, I'm not quite 100% certain that I understand what it does, so I'm not going to go too deeply into it. As I understand it the probability distributions are derived from the preceding constraint. So each probability distribution is conditional to the preceding constraint probability distribution and they need some fancy re-weighting and pops out the population at the end. The algorithm is slower than deterministic re-weighting but it's not hugely problematic, it's not that slow to cause major problems. It's not dependent on having a sample population which is a real bonus. If you don't have the sample population to synthesise from you can actually generate your own from the aggregate date without sampling. That though, can also cause a problem because when you do have a sample population you want to constraint to, you don't want to go outside of those bounds, the algorithm can actually produce people that don't appear in that sample population which is then a problem.

Simulated annealing, basically executes in two big loops. The population, shall we say, is generated at random from the sample data, the sample survey. Changes are then made to this generated population at random and changes are either accepted or rejected to whether they make the population a better fit to the, to what we're trying to get to at an aggregate level. So taking that individual population, aggregating it back up, do we fit what we observe or is it getting worse? Are we getting better or worse? Now it also employs the Metropolis algorithm which will accept backwards steps if you like, a worse change. This is dependent on how far down the annealing algorithm you are, the closer you get to the end, the less likely you are to take a backwards step, so you keep taking forwards steps the closer you get to the end. This avoids some of the issues of hill climbing, getting stuck in sub optimal solutions, but it does have the penalty of having a huge number of calculations required. For example, creating a population for Leeds required just short of a quarter of a million iterations of this outer loop without these iterations here, or the calculations of the fitness or anything like that. So it is computationally very intensive. By applying some indexing strategies we can speed up the execution of the model, so we're not actually cycling the data all the time, we just jump between arrays which speeds the model execution up quite a lot. And one of the big bonuses or one of the big benefits of this approach is there's very little pre model set up. You clean your data, configure the model and hit run pretty much, you avoid a lot of the set up that you need for the other two approaches.

So the experiment or the area that we've been experimenting with surprise, surprise is Leeds, and we have 715,402 individuals, this is the residential population for the 2001 census. And we're using an extract from the sample of anonymised records, the small area of microdata file for the sample population to generate our synthesised population from. We're using aggregate data from the 2001 census for the constraints, these were the experimental constraints, just six univariate attributes taken from the census tables, and then a selection of both univariate and cross tabulated attributes were taken from the census tables for evaluation purposes, to see how that synthesised population aggregates back up, how well we're representing the aggregate numbers.

Small cell adjustment method or SCAM did cause a few issues with the data because the tables didn't all add up which is a bit of an issue. So I took the actual total populations for each of the zones and re-weighted these tables so that they matched and the proportions stayed the same but the actual total numbers add up to the actual residential population in each of the small areas.

So to the initial results. As Voas and Williamson state the constraint table should be well matched by all methods, but the simulated annealing method was the only one that produced a perfect fit at the univariate level anyway, it actually didn't misclassify any of the people within the constraint tables which is quite pleasing. Conditional probabilities model was very close, there were very few [misclassification errors] and the deterministic reweighting was a little bit out with the qualification and age but not so bad with the other constraint tables. I must stress as well that the deterministic reweighting [method] didn't have any swapping nor simulated annealing on the back end which is quite common practice in [some other studies] they actually have the swapping algorithm on the back end, this one is just purely reweighting.

So evaluating the results then, taking the attributes that we're not constraining the model by and looking at some of these relationships within the synthesised data that we produced. How well are we representing the sex and marital status variables when you consider the numbers that fall into both of them categories at the same time? And we can see that we actually don't do too badly. Conditional probabilities model and the simulated annealing are quite close in all respects.

Finally this is just looking at how we do with attributes that lie outside of the constraint categories. So these are attributes that are not incorporated in the constraints, apart from the sex by hours worked, that was put in there just simply to look at how well we projected the hours worked constrained on sex alone, so just constraining on one of the variables of the cross tabulation. Here again we have varied results. The interesting one is tenure, we're looking into that in a little bit more detail at the moment but unfortunately I just didn't get a chance to finish some of the things I am doing. I've got a sneaking suspicion that's to do with geographical location within the city, so some people of a certain economic activity or social status are maybe less likely to own their flat in the city centre, are more likely to rent, whereas in the suburbs [a comparable person is] more likely to own a house than to rent a house. So that could have some bearing on that [issue].

Now one of the things that we're quite interested in, from our particular application area, is not so much projecting outside of the constraints, but just looking at how well we synthesise the population within our constraint boundary. But we actually want to have quite a few attributes in our model. So we're looking at how far we can push this [number of constraints], the constraints within the model before it starts to break down, before we start to see significant deterioration in how well we fit our constraints.

So this is some results pretty much hot off the press this morning, I actually put this graph together on the train on the way down here, I haven't had a real vast amount of time to pull this apart, but this here is the number of people misclassified in each of the constraints. So this is out of the 715,402, and this is the percent error on the side, the red line tracks the percent, the blue bars are the misclassified people for each of the constraint categories that we pushed through the model. And I think the highest ones still rest at 0.55% error which we're quite pleased with. I haven't had a chance to examine the inter relationships to see how well we're actually representing them yet which will be an interesting test, hopefully one I'll do in the next few days.

So this is the framework that we've been working on which houses this populated synthesis model, the simulated annealing approach and the deterministic reweighted approach, it also has a spatial interaction modelling element in there, and we're actually going to be developing the agent based model into this framework in the coming months. Basically it's all written in JAVA and each of these plug-in modules if you like are flexible, you could take them away, put them in, develop new ones, plug them in and it picks them all up at run time, so there's no hard coding in there. The only bit that is hard coded is this top bit which is the framework [which] deals with all the screen control and things like that, and then we have various databases that hold our data.

Screen shots, literally just to show the interfaces, so how you put together the models, it's all drag and drop, you literally drag tables in, create links, select what variable, what statistics you want to pull out, hit the run and wait for your output. And there's one thing I must say is [that] there's an incredible amount of statistics it can spit out, right down to the cell level so we're getting variance on how each category is actually performing, each constraint is performing, but I just haven't got through all the statistics yet, all the tests done, so they'll be coming in a paper that we'll be writing in the next few months, there'll be a lot more detail in there, in the approach and in the test.

So on to the nitty gritty bit, the human behaviour. We're interested in consumer behaviour but only the important aspects because really we haven't got the computational power to represent human behaviour in its entirety, nowhere near, to be quite honest with you the frameworks that we have at the moment are simplistic, they're just not capable of handing that [all human behaviour]. These are two of the practical frameworks for representing human behaviour in agents, they've been discussed quite widely so you might be quite well versed with these. Beliefs, Desires, Intentions is probably the widest known, introduces practical reasoning, decides goals and how to achieve them, so literally the agent can stay out to produce strategies on how to produce goals. Beliefs [refers to] their knowledge about the world. The desires are the goals that the agent is trying to achieve. And the intentions are really the most important, balance, so the priority. Through these, the implementation of these three attributes if you like we produce a strategy for achieving the goals. The drawback is that it is very unrealistic. It creates completely rational agents, there's no irrationality in there, they rationalise and they do that.

The PECS framework, Schmidt and Urban 2000, is a little bit more flexible. Basically it's Physical condition, Emotional states, Cognitive capabilities and Social status. This introduces a bit more of an irrational element. This has been successfully applied by one of the PhD students at Leeds [University] at the moment for a crime simulation. Nick Malleson is one of Alison's PhD students and this is, an image he's produced and we've borrowed it from him, supervisor privileges and all that. It just shows how he's implemented the different states, the different attributes to produce the actual actions and the priority of the actions that take place to represent the behaviour. It takes in the state of an agent at a given time, calculates or applies intensity functions to that state, produces an intensity which goes into an action guiding function and then produces a set of actions that will be executing or try to be executed in a certain order. This has worked very well, and there is a URL for his [Nick Malleson's] blog spot here. He's produced some very good simulations and some very realistic movement and behaviour of the burglars within a certain area of Leeds. The drawback to this model is that it's computationally very intense, he's got the agents moving around a realistic road network, interacting with buildings, interacting with each other, but it's computationally very, very, very intensive. He's got a small number of agents and it takes quite a long time to run. So one of the problems, well not problems, challenges, look on the bright side, one of the challenges we have is to upscale our model so we could have our agents moving round and interacting but at a much faster rate and at a much quicker pace.

So that's about as far as we've got. There's a list of references here if anybody wants any more information on anything I've talked about. I'll try and answer any questions, I'm not guaranteeing I'll be able to, but I will try and answer any questions that you have.

QUESTIONS

Q: A quick technical one. You fitted the various population generated algorithms on your whole dataset, you didn't do half and then check it against the other half or …?

Kirk: No, basically what we did is used the six constraints, see if I can go back, there, used the six variables there as constraints, so they go into the model and they're used to constrain the population, when you're selecting your population you fit into the numbers in these tables.

Q: Yeah I'm not clear whether you did the fitting exercise on the whole population or whether you did a hold back.

Kirk: Yeah, no we did it on the whole population. The validation if you like from the re-aggregation of that data set, the synthesised data set takes into account some variables that weren't put into the data, so looking at the relationships in between, the attributes.

Q: It's not really a question. Tenure is probably geographically as well as socially concentrated in the UK, so with regards to social status, it's not [AUDIO UNCLEAR]?

Kirk: Yeah, well one of the things that we're actually looking to do and we just haven't had a chance to do it before today because literally the paint's still drying on all these results, is to map them and look at the geographical disparity and then looking at where they fall in context, one of the reasons why we use Leeds is because we know quite a lot about the city, maybe what we'd expect to find and see if what we expect to find we actually do, it's quite interesting.

Q: A slightly odd question, but what makes the architecture of the youth behaviour and architecture? I mean BDI has kind of been out there years, again computer science literature, PECS is just another, I mean why that one, why not …!

Kirk: Well that's just one that's successfully used, the two of them around, but I agree, what does make them successful architecture and what makes that one more successful than this one over here …

Q: I just kind of wondered how many are out there that would be well defined architectures for your needs?

Kirk: Funnily enough we're actually talking to some, is it Helen Jackson, Educational Psychologist and she's been working with some agent based modellers at Southampton, working through some of these issues, looking at team working and we're having a meeting with her next week to actually talk through some of this in preparation for the next phase, which is looking at these architectures.

Q: Because SAW is another big one used in the literature, but I've no idea whether it would generate the kind of outputs you needed.

Kirk: But it would be nice to have the time to actually test some of these, like implement the different architecture to see what happens, what performs, what doesn't and why, you know ask some of them questions. I don't know if we will have that time because like I say we've got limited resources and a finite amount of time, but it's certainly something I'd be really interested in doing.

Q: Just to pick up from Edmund how you're going to go about, when you've got your agents running round, how are you going to validate …?

Kirk: Good question! One of the things that we are actually planning to do is, I did my PhD with, in the education sector and the education sector has some incredible data which is really under used, vastly under exploited. And one of the things I'd like to do is take that data and it has attribute information, apply some behavioural architecture as we were talking about on the top, and then test what we get out from the model in behavioural characteristics, to what they observe in the preference data they collect. That gives an insight into the actual behaviour that you will observe. So that gives a method of validating if you like the actual behaviour, how the behaviour's playing out, but for actually looking at where people go to shop and things like that it's very difficult because although we've got access to a commercial data set, recently agreed access to a commercial data set to look at some of the behaviour for shopping, we've talked to ASDA, we've talked to various supermarket retailers, very protective of their data! So where do we get the data to actually validate these? It's a very good question in the retail market.

Q: Subsidiary to my earlier question, I guess in a way what you want to do, given what your limited resources is to cover the field of architectures as effectively as possible, BDI's got big problems in microsimulation as economic models, PECS I don't know very much about. Do you not want some sort of learning model in there, some sort of neural net? Because that's a whole other family of agent architectures would be a reactive learning agent without any kind of behaviour, it just adapts.

Kirk: That's one of the things we've talked about at great length and yes, the answer is yes we'd love to, but again there's time restrictions. One step at a time, we're going to try to take a nibble and try and get some behaviour there and get some realistic movement, and then if we can get more funding, more resources on the project, try and get some learning model as well. The gaming industry, you've only got to look at the gaming industry, miles ahead on that, role playing games, they have learning algorithms in that they react to players doing on this side of the console, tailor how the game reacts. So actually making that, or incorporating some algorithms to actually learn and develop the behaviour would be fantastic.

Q: But I guess you wouldn't want to spend too much time on PECS and then discover that it wasn't actually that different, because there's this sort of issue about your metric and how broad a class of architectures and things really are.

Kirk: No, yeah, I see what you're saying, you don't want to concentrate your resources in one place and then find that …

Q: If you're only going to look at two architectures that actually they're both broadly rather similar.

Kirk: Yeah, well that's one of the reasons why we could do with looking at a broader range of architectures, see if any do stand out, if they don't stand out you know ….

END OF RECORDING