ESRC Microsimulation Seminar Series

Seminar 3: 'Moving beyond tax-benefit and demographic modelling'
Leeds, 2nd July 2009

Retail , expenditure and time-use

Ben Anderson

[EDITED TRANSCRIPT]

In the spirit of giving an overview of the work that we've been doing, I've really got essentially five things that I want to talk about. The first is essentially what's the problem space that we're operating in and it will be clearer to some of you, perhaps less so to others, and I want to talk a little bit about what we've doing in terms of income modelling at small area level, expenditure modelling so consumption if you like from a sociological point of view, time use and then I want to talk a little bit about the current activities that we're doing which maybe form the basis of some future collaborations.

So, before we go into some of that so I wanted to, just for the non geographers amongst us, presumably there are some I hope, to give you an idea of what I mean by small areas, this is Oxford I'm afraid because I haven't got decent maps of Leeds or anywhere else other than Oxford at the moment. So those would be our Parliamentary constituencies, then within those and around those we have local authority, unit 3 authorities and then wards, and wards would the first area level that I might refer to as small areas and I'll show you some wart level modelling, and then the lower layer super output areas which are the ones which the census data is now out in as well, and output areas which is the smallest one you can get out of the census, so within the LSOA. So I'm operating, we're operating with data estimates, real data at these levels of geography, so the ones in italics, the wards, the LSOAs and also the OAs, because census data which is predominantly what we've been using is not available at postcode level or postcode unit level, we won't be talking about postcodes.

So why would we want to bother with it? Well there's various reasons we might want to use that sort of level of data for applied reasons if you like, whether it's public policy or strategic service delivery management or things like that, and I'll talk a little bit about some of that a bit later, and there's also research reasons for why we might want to as well. The ones that are particularly of note for this talk are deprivation indices, inequalities that we might be interested in within the small areas, and to some extent spatial small area levels of consumption patterns as well, all of which have both research and actually applied interest as well.

But of course the problem that we have is that whilst we have census data down to the OA level, the census of course collects only certain kinds of data, it doesn't collect much that is actually particularly sociologically interesting when it comes to consumption. There's lots of things of course about employment, a certain amount about qualifications, there's no income in the UK census from 2001 at least, I understand possibly there may be income in the Scottish, next Scottish census and maybe Stephen will talk a bit about that. So the data that we have at these area levels are broad and shallow and sometimes if we're interested in linking to administrative and private data, the latter can be of let's say unclear quality rather than uncertain because in many cases it's difficult to work out exactly how these sorts of data have been collected or imputed and so on.

So in order to try to deal with this what we've been trying to do as part of this wider programme of work really in the UK at least is to build synthetic censuses, which are small area estimates which include those variables that are not in the census but they're the ones that are of interest to us, they may be income for example or they may be kinds of consumption that aren't collected by the census. And if possible we'd like to validate them, so it's all very well producing estimates but if you've got no idea whether they're actually reasonably close to the truth then they're not going to be of much use to you. And as part of the validation process in taking that forward we're also interested in being able to model, I don't really like the word 'forecast' but perhaps projections or scenarios further out into the future so that we can look at potential impacts of different kinds of interventions on spatial distribution of various kinds of things, and I'll talk a bit about some of that later one.

So how do we do it? Well in essence we're using spatial microsimulation process which by now is probably quite well known in people coming to these seminars, and our approach is to take survey data cases, so let's say for example it's the family resources survey, we take all of the households in that case from that and we put them into a zone which could be a ward or an LSOA or even an OA if you like, we take the households from the

particular region that they came from in the survey, so we don't put Londoners into York, we put people from York and Humberside into York, we could do it nationally but we've done it this way. We do the same thing for each of our zones. Now of course if you simply do that you just get replicates of the FRS across the country which is pointless, so what we do is we take the census data as constraints to re-weight these cases, and there's all sorts of different ways of doing this, for example we use iterative proportional fitting which is deterministic, you always get the same answer which is helpful when explaining why things have gone wrong, there are others such as combinatorial optimisation which Paul has been working with. And the whole idea is that you make these households, you re-weight these households so they fit what we should expect from the census. So if our original sample here had an average of say two cars per household, we want to re-weight it so that it comes out with whatever the census said, should be in that place. So you end up therefore with a whole load of re-weighted households for each of the zones you're interested in, obviously the smaller the zone the more weighted cases, data you've got and everything starts to get very busy.

So the question or one of the questions becomes how do you select which constraints? Do we want to use the age of the household response person, do we want to use the number of cars, the number of people, what have you, because of course we could use all of them and the more we use the longer it takes and some of them may be very inefficient. So what we tend to do is try to work out which ones are going to have the strongest effect, in other words which are the best for re-weighting for our purposes? OK? So if I give you an example of this, we use regression to select the constraints from the survey data, so just thinking about income and this is, that's the definition of household income deprivation that DWP use, so it's, well equivalised household income being below 60% of the UK medium. And what we do is we put in all the potential constraints from the survey data and we run a very simple regression model analysis and it tells us a number of things. First of all it tells us in the blue bars what the effect of the particular constraint is on its own, so that's a bivariate regression. So you can see in this case the number of earners in the household has the strongest effect, followed by employment status and competition, so that's blue. But then if we run that as a set of nested regression then it enables us to see what the effect is when they're all combined into one model, so gradually over time you can see that we're not actually getting much additional benefit from the variables down the bottom here. OK. So each one adds a little bit more to the explanatory power of the model, that's the yellow, and the sum of it is here. So after about let's say there, things aren't changing that much, so in terms of the constraints we might want to use for this particular variable, it's probably going to be those. And the thing about this is if we change the variable of interest, we might well have to change the constraints in order to make it work as effectively and this is actually a bit of a pain because it means you can't run one model for lots of variables, you actually have to think about testing this for each variable that you're interested in so that you can then say to people for example we don't think we can generate a good model of that because we're not getting a very high explanatory model in the survey itself. OK. So if an r^2 comes back of about let's say 2% I'd probably say well you're going to get a random result basically, it's probably not worth doing, which gives us a method of filtering which is very helpful as well.

So having established what our constraints are, we then run back through this process and the outcome is something that looks a bit like this. So for each of our zones those are LSOAs in England, that's one LSOA in England which has a region which is 3 which is probably the North West, and you can see that we've put the households from that region into it, and these are the weights, so the weights are fractional, OK, so in each zone we've got all of these households, they've got fractional weights and we're not going to chop those weights off we're going to leave them as fractions, so we keep all households with all that variation and we keep the weighted, keep the weights as they are and it's brought in with it, so that's the weekly income of each household identified by the household ID here from the Family Resources Survey. So we've got the weighted income which is quite easy, you multiply that by the weight, and then we've got the likelihood of them being, or the fact that they are poor in the sense of that income is lower than 60% of the median, and then you can do any kind of weighted statistic you want on that data set. We've got two over there, one of them is the weighted mean income and the other is the proportion of households whose income is below 60% of the median. So you can imagine for every zone, in whatever number of regions you want to do, you can generate these sorts of statistics for each zone. There are other things you can do and I'll come back to one of those in a moment.

So following on from that I just want to give a little bit of a discussion of the kind of income modelling we've been doing really for two reasons. The first which was funded by the DCLG was the potential or exploring the potential to replace the income, the main score in the current indices of deprivation which is based on benefit counts, and what they want to do is to try to line it up with the DWP's poverty indicators which are these there so we were trying to explore whether or how well we could estimate those. And we were also asked to look at deprivation patterns for a number of devolved administrations and the Commission for Rural Communities as well, so we're using some of the things I've just been talking about in terms of the data. And I'll just give you a

flavour of a couple of the results, this is for the East of England, eastern region, I'm afraid there's very little in the way of physical geography on there, it's just outline, and these are regional results summarised from those sorts of zone levels. And on the charts here we've got three things. We've got the blue bar with the error bars is what you get from the FRS itself at the regional level, so that's usual survey data analysis. And then the pink is what you get if you model everything after you've selected the constraints at the English level, so we run that regression model for England, we use those constraints and away we go. And then the last one, the yellow one is what you get if you select those constraints region by region on the basis that spatially those constraints may vary in terms of which ones are the best predictors and therefore the best to use as re-weighting constraints. We can only do it at the region level using the FRS because we don't know geography beyond that within the FRS which is unfortunate but true. And in general our results, well really the spatial simulation results follow the FRS and the regional approach follows it closest which is a satisfying result but again it means more complication because you have to keep re-running it for every region. And another example is the result for pensioners, we've got the difference between before and after housing costs. But we've got another bar in here which is this one and that one is what the DWP publish as their results, and I don't know what they do, exactly what they do, but I cannot get any of our results to match them exactly, it may be the way they re-weight some of the data at the top end of the FRS distribution, it's difficult to tell from the report exactly what the mismatches might be, but for all of the models that we've tried to run we consistently under estimate what the DWP publish, but we pretty much match what we get from the vanilla FRS that you can get from the data archives.

In terms of validation, because of the interest in trying to use this as part of the IMD we compared it with the income domain score of the IMD and by and large you get a reasonable correlation, it varies in some places, for example using the English model or even the regional model in the South West before housing costs it doesn't actually match that well with some places, and some of those places turn out to be areas which have got people in them who are not picked up by benefit counts such as household response persons who are students.

So that's a sort of work in progress and I don't quite know yet what they want to do with it and whether they are going to implement it or not. I have a feeling that the Welsh and the Northern Ireland stats offices are probably going to do this, the CLG I think is still making its mind up.

Remember I said that we can do other things with that weighted data, well bear in mind that we've got here a weighted income distribution which is the kind of thing you get in any income survey, and from that we can calculate the gini coefficient which is of course a very well known way of representing inequality across income distribution. And we've been doing this within small areas and as far as know nobody's ever tried it but it take an awful long time because you're using these weighted data sets. So if we calculate the gini coefficient for each zone across the eastern region of England, that's what it looks like, that's the income distribution itself and I just want you to keep your eye on that little part of East Anglia which effectively is rural High Suffolk and North Essex which is predominantly farming land use, because it appears to have quite high income inequality. If we look at the next one we can actually see that. So overall the UK gini coefficient is about 0.36, the mean in Eastern England is 0.35 but we can find areas, for example central Cambridge which is about 0.40 and areas in South East Ipswich which are much much lower. And when you then drill into the data you can tell stories about it that start to make sense.

So Cambridge is the blue one, so here's a very high gini coefficient but what it happens to have in that zone is a lot of people in the highest economic work employment category and a number in the lowest which will include students, and then we have a slightly lower gini coefficient, you can see they're all pushed down towards one end. So you can see why in the Cambridge case you would expect to see this inequality, this income inequality, but in the Ipswich case you wouldn't because they're all, everybody's pushed down to one end, it's more homogeneous if you like.

So what we've been using this to do is characterise income and equality and looking at it across different zones particularly, we started doing this in Northern Ireland but obviously the Commission for Rural Communities were very interested in this as well, in the sense that for rural areas you get a much denser distribution but also in some cases a much higher income inequality which is in some cases what you would expect.

So I want to switch now, to thinking about expenditure and again one of the reasons we've been doing this is to look at inequality, so re-using the gini coefficient but instead of looking at income, looking instead at expenditure, so the idea of expenditure equality. And also an awful lot of the work that we've been doing has been looking at ways in which infrastructure providers can essentially steal a share of your wallet, that's they way they rather crudely put it, they want some of the income that you're, some of the revenue you're spending on other things

they want to come to them and we can look at ways of doing that through demand system models. But that's in grey because unfortunately I can't really show you much of that at the moment because it's part of this work, hopefully we'll be able to talk about that in the next year or so.

But I just want to start off by looking at some of the consumption and inequality work we've been doing. So for example we've been looking at water consumption through the expenditure and food survey because that has, records in it the amount of money you spend on water and the mode of paying as well. So we looked at the distribution of expenditure on water across East Anglia and we've also been doing it using this definition of water poverty which is sort of similar to the definition of fuel poverty as well in order to try and look at whether there's any socioeconomic distributions that we might need to be worried about. And one of the things we're doing with this with the Department of Biology is to try to match it up with known localised water resources as well to try to model demand and supply of water in the locality because the East of England uses a lot of water but it's got virtually none of its own, other than through bore holes which are rapidly becoming difficult to extract from.

Going back to more technology oriented, one of the reasons that I wanted to get involved with telephony is of course that we're able to do things like validation through data sets which can be provided to us by large infrastructure providers like BT. So we estimated at the zonal level, in fact we did it at ward level for this, this is LSOAs, the distribution of expenditure on telephony and then tried to validate that against BT's own internal data, and it turned out to work remarkably well as you can see from the correlation coefficients there at the ward level. There were some wards where BT didn't have any customers but we predicted income, so if you like those are areas where, it doesn't cover Hull where of course there'd be no BT customers at all, but it enables you to start analysing what you might see as latent demand, so it's places where an infrastructure company hasn't got customers, hasn't got revenue but it could have, and I'll come back to a little bit about that.

Now one of the, it would be nice to say that we had some sort of role in this latest nonsense about putting 10p on a telephone line to pay for Broadband, but one of the things that we did do, and this is a fictitious version of it, was to look at what you might call a return on investment model. So we were trying to say well let's imagine that if we want to wire up all households in the country with some sort of level of Broadband, let's say 2 mega bits, it's going to cost you £500, let's assume that, that's about £500 per household, let's assume that that's going to be clawed back through an increase in internet subscription revenue which is not across the board but it goes up by a certain percentage for each household, OK, so this is not an average, it's a microsimulation model because each unit has a different amount that they spend, if you increase it by 20% we're going to assume 50% is used to offset that investment and we're also going to assume that once one area has paid back you're not going to transfer it to another, well that's obviously complete rubbish but let's just assume that that's how it works for the moment. This is an indication of how long it would take to pay back. So for example North Norfolk is highly unlikely to get wired up under this kind of model because it's not going to pay back in any kind of time frame than an infrastructure provider can plan for, eight years for example is too long in most time frames. And there are other areas as well which aren't necessarily rural but they are perhaps deprived areas as well. So that's an example of the sorts of strategic modelling that we've been doing for BT in that case.

I want to switch now to time use which is, it's all using similar sorts of methods and techniques but the, the sort of sociological interest is varying. One of the first things that we did was to look at work time, again because it's relatively easy to validate it against the census because in 2001 the census asked about work time and we were able to simulate it, estimate it using the time-use survey 2001 and the census, so this is contemporaneous, almost contemporaneous data which is very helpful, you're not looking at two different time sets, and you can look at the distribution of work time which basically follows the usual ring effect out of London as you'd expect, and it also matches quite well onto the income distribution that we saw before and the validation worked reasonably well. So again this gives us some confidence that where we're able to select constraints which are reasonably good predictors, we think we're probably going to produce something which is a sensible result, OK, so it doesn't always work, it's not going to always work, but it will work in some cases which we are able to algorithmically select. The flip side of that is many ways is media usage which is dominated by television, TV watching, and it turns out that if you look into the actual data itself, these are predominantly areas dominated by retirement populations here on the Essex coast, here on the North Suffolk coast and we've got some North Norfolk coast as well. Now it turns out that if you're interested in delivering let's say digital services to these people it's going to cost you the most to get it to there even though those are the people with the most time to use it, not necessarily the most money to spend on it, but certainly the most time to use it.

And now just to think about scenarios, and this is a very simple instantaneous traditional static microsimulation model. So we've taken demand systems which are well known in economics and I mentioned them briefly with

the expenditure work where you've got an inter relation between, in this case it's time use but it's usually done as expenditures, which have a knock on effect upon one another, you can't assume they're independent. So if we change the time being spent on one thing it's going to have a knock on effect somewhere else, on that matrix of time-use that we have for a given day. And you need to estimate those using a demand system model which are fairly well defined now in economics, and having built the models we're then going to make some sort of change within it, re-estimate it to see what's sort of bubbled and rippled through this demand system and then run through the spatial microsimulation to see what the spatial effects might be. And again just thinking about television, when you do a microsimulation just within the survey you don't see much of a shift at all within the distribution, there's a very very minor shift. Sorry I probably should have said the model was in 2001 switching everybody to have internet access, well in 2001 very few people actually had it, so you're not going to see much of an effect, if we did this now you might see a much more substantial effect for the way in which television is delivered, possibly not if you add a television as an experience that's actually being watched. But then when you look at the spatial effects of it, you can see that there's actually quite a shift in some of the distributions, so these are the scenarios for the different urban types and you can see how for example they've all shifted rightwards quite substantially, perhaps more so than you might expect from the basic microsimulation itself. So it implies that when you play this out through the spatial distributions you may actually pick up some more interesting effects than you would have done if you'd simply looked at the whole population which is there.

There's a lot of words on there. Basically what it's saying for a start is that estimating demand systems takes an awful long time, we've been having, we've been running some of our models on grid clusters and they take up to a week to estimate, sometimes they never converge at all which is not very useful when you've got a commercial client saying I want the answers last week please.

Also of course instant change is unlikely, we need some sort of pattern of growth, so not everybody's going to get switched on to something overnight, this is not a tax change you know where you can do these sorts of modellings, sort of modelling, and as I said almost certainly we are looking at historical effects, if we were to do this again now we might find some quite substantially different results.

And the final thing I want to just give a brief overview of are two things, and actually I'm going to talk about them in reverse order, one is projecting the censuses and the other is Tony Lawson's work, PhD student, on agent based microsimulation. I mentioned right at the start one of the things we're interested in is pushing our, as with MOSES and SAGE and a lot of the other pension projection work, is pushing out some of these scenarios into the future and to do that we need fixed geography over time. So we've spent, wasted an awful lot of time trying to link up censuses over time in sensible ways to be able to project them forwards, and what we're working on at the moment is to take the enumeration district zones for those years and turning them into, or aggregating them into LSOAs for 2001 and then assuming they're going to stay constant over time, they probably won't but we're going to assume that they will, in order to give us a constant geography. So we're working at that level. We've been doing it, trying out a number of ways of doing it and some of that work is just sort of seeing the light of day now.

But then perhaps more interestingly and more importantly, we're also looking at ways to, not to dynamically model the whole population perhaps as MOSES was trying to do, but to take our sample data sets, whether it's FRS or the Expenditure Food Survey or whatever, taking that as a sample and rolling that forwards in a dynamic microsimulation model which is what Tony's work's been doing. And then what we do is for every year that they match up, we drop out the two data sets and we combine them using the spatial microsimulation approach that I mentioned before. So we're not trying to dynamically model people in space, we're doing them separately and then combining them, so that's what we're doing at the moment.

Tony's work uses Netlogo which turns out to be an extremely effective way of building, quite rapidly building an agent based model and he's used for example the BHPS transitional probabilities, he's also taken lots of results from the SAGE project and re-used those, save ourselves some time and the usual actuarial tables and prices as well so you can build expenditure modelling into it. Then he's validated it by taking the BHPS as a seed and running it through to 2005 to see how they've compared. So the dark line, fixed line, solid line are the proportions of households or individuals who are married, divorced or widows and the dotted lines are his model results and for most of the demographics they match up reasonably well. So what's underneath this appears to be reproducing the patterns that we would expect, that's cohabitation for example and the different age groups as well.

You can look at some of the models, I seem to have lost one of my slides, but he's got three models which are on the web which you can take a look at and our ongoing work which I was actually talking a bit earlier about is to try to do something with expenditure modelling with work location and fuel consumption in mind. So the idea is if you can shift work to people rather than people to work, so it would be the whole home work, tele work ideal, how might that change the distribution of people, work, travel, fuel consumption in the context of, and peak oil and high oil prices. And you can find nearly all of this stuff there at the moment, we have moved now into sociology so that will change but that will still stay there, and our new stuff will be on a new website. OK

QUESTIONS

Male question 1 – I've got a question, you know what it is! The income distributions that you talked about…I have an idea that when you produce synthetic income estimates they're never going to be as varied as they are in reality, so even when if you, you know when you're looking at correlations you may have a perfect correlations but the variation in the synthetic estimates isn't as great. I mean I just wondered if it was possible in your comparison between the synthetic estimates and the IMD to comment on that or to test that, I'm not quite sure what's in the IMD income indices.

Ben Anderson – It's a score, a single score for each zone.

Male question 1 – So presumably you could reproduce that score synthetically and look at whether the variation was great?

Ben Anderson – Well yes in a sense if you were to do that you'd be synthetically reproducing the benefit count score which you could do because the benefit, whether or not those benefits are being taken is in the FRS for example, so you could do it from the FRS to see if they match up. But I think what you were asking was, I mean that perhaps would be, another way to do it would be to compare the variation in each zone that you're producing synthetically with the variation from either a ground check data set like, well the income question in the census but if it wasn't banded, but also whether you could extract some data from say Experian which is real data within a zone. Or the other way you could do it is the FRS is clustered and I think the DWP would tell you not where they come from but which ones come from which cluster within a post code unit, so you could compare the variation within the sample at the small area level with an estimated variance.

Male question 2 – Right, I've actually got some quite good ground check data for research use so it would be quite interesting to talk to you about whether we might run a test on that. I mean just a very quick comment if I may. I mean one was when you were looking at the urban town village thing, I mean I just wondered if you'd considered looking at geo demographic classification of that like OAC for example, it might be a bit more, would be another interesting way of actually portraying those variations because that seems like a, you know you're effectively looking at sort of geo demographics in a fairly crude way, so that was one comment. The second was that when you were talking about the 71, 81, 91 censuses I had a feeling that Danny Dorling had done quite a lot of work on that two or three years ago …

Ben Anderson – They updated it to 91.

Male question 2 – I think he took it to 91 so they were …

Male question 3 – Didn't he update it, they were linking their tables but not projecting it I think, so it was matching the geographies between the existing data but not …

Male question 2 – Yeah I thought that's what you were talking about at the top part of the …

Ben Anderson – No, effectively what we've taken is the LSOAs from 2001 as our fixed units and then used aerial interpolation or whatever to turn the data from here into those boundaries, I don't think Danny ever did that for 2001. The Geo Convert Service offers some parts of this but not all of it, so we've had to implement some of it.

Male question 2 – And then the third thing was when you were talking about the forecasting and so one thing it does, I mean it would be relatively easy to take you know say MOSES and produce area level projections for these areas and then reproduce your kind of synthetic thing from the survey record or what have you, so that would be one way of validating we might potentially talk about.

Male question 3 – Thanks for that, it was a very interesting overview, quite a lot of work you've done there. I was very interested in the agent based modelling at the end and the, I was wondering do you model, what kind of interactions do you model between the agents? Do they interact and how does it work?

Ben Anderson – At the moment no! The interactions that there are is that it's in the sense that it's a closed sample, so cohabitation forms from within the sample I think, but that works on a probability basis, there's no interactions at the moment between the agents as such that the behaviour of one can affect the behaviour of another.

Male question 4 – (inaudible)

Ben Anderson – Yes, but for example your spending choices don't affect that of your partner, and if you've got a new partner then it doesn't, theirs aren't affected by your spending choices except to the extent that the way Tony's implemented the expenditure modelling is that when for example you cohabit with a new partner you will inherit the spending characteristics of a couple from last year with your characteristics. So your spending pattern would change through that re-allocation of expenditure profile. But that is not explicitly interaction between the agencies itself.

Male question 3 – So when you dynamically project them you don't have a modelling process of choosing a partner for someone who is single to get married someone from inside the model - you don't do this sort of thing?

Ben Anderson – How does the matching work?

Male question 4 – You match them to someone of appropriate age, educational status and so on…

Ben Anderson – Yes; there's a lot of stuff about marriage markets that could be implemented!

Female question 1 – Can I just have a quick question on this? So what do your agents do in the system and in which way the two approaches are combined?

Ben Anderson – I'll let Tony answer the first one! The second one if I could just, the way we combine them is that the agent based simulation produces a new synthetic survey every year in effect because it creates a new sample, so we then take that out, well we haven't done it yet but what we're planning to do is to take that out as a survey data set and push it into the spatial microsimulation method. OK. What do the agents do?

Tony Lawson – The agents… (inaudible) …so a household at a certain time and to a certain... (inaudible)

Female question 1 – Yeah but in which way you call them agents, agent based modelling rather than microsimulation? What makes it the distinction?

Tony Lawson – I don't think there's very much distinction really. We call them agents because we used… (inaudible)

[END OF RECORDING]