ESRC Microsimulation Seminar Series

Seminar 1: 'Scaling up; scaling down'
London, 2nd April 2009

E-infrastructure for large scale spatial simulations

Mark Birkin

[EDITED TRANSCRIPT]

OK thanks Paul. OK something slightly different perhaps! The theme of the day I think originally was scaling up and scaling down and Paul's aware that I'm involved in a number of projects involving something called e-Infrastructure and I'll come back and say a bit about what I mean about that later on. So my theme is reasonably close to what I think was originally advertised but e-Infrastructure for large scale social simulation and I've put the name of myself and Andy Turner are both researchers at the University of Leeds on this presentation.

So what I'd like to talk about is broadly speaking four things, I'd like to give you some background to a couple of projects that I'm currently interested in and then talk a little bit about the kinds of features and capabilities which we've developed to deliver on these interests. And then hopefully if I get the timing anything like right, I'd want to spend most of, or a good proportion of the talk speaking about current development plans and priorities.

So background, well I'm talking about two projects specifically, one is something called MOSES which is well ran from 2005 to 2008 and the follow up project using a program called GENESIS, I imagine there are quite a lot of GENESIS projects going on around about the world at this time, which has been running since last October and that's actually a partnership with a group at the Centre for Applied Spatial Analysis at University College London. OK, both of these projects are funded by the ESRC, our sponsors today, the Economic and Social Research Council as part of its NCESS, National Centre for e-Social Science, so again the e-Social Science is focusing on essentially projects which use advanced computational technologies of various kinds to deal with social science issues and problems.

So this particular project, one of its aims is essentially about cities and regions, so it's microsimulation at the urban and regional scale, so we're interesting in looking at cities as they are, as they will be, as they could be.

So if I move straight on and then try to unpack that a little bit, so the specific sort of components of these projects we're trying to develop realistic representations of cities, so again at a micro level we're thinking about what kinds of individuals live in particular locations, interact in what kinds of ways within urban areas. And this project is concerned with projections and forecasting as in the previous presentation. I'd describe this as medium projections, I work in a Geography department so you know we have people in our department who work on you know ice cores from 200,000 years ago and longer ago than that, so I could hardly call these long term projections although they might be considered so in a social science context. So basically the sort of modelling we're concerned with so far is running to about the year 2031, so we're running over a sort of 25 year time period, so quite a long time for social projections.

We're interested in things like changing behaviours and activity patterns on problems like service utilisation, resource planning, scenario based forecasting, so things like particularly issues about Asian populations, of course there would be changing economic structures, questions of that kind. And then a fourth thing which I've said again already, but particularly the kinds of technology interfaces that we can put together to support these sorts of activities.

Now I'm not going to say a lot about the kind of technical aspect of the first of those three points and indeed my colleague Belinda Wu is going to talk a little bit more about the dynamic modelling aspects of this project after lunch. So in terms of background I've just got I think four slides which are giving a very general flavour for the sorts of thing that we're trying to do and I think Paul was probably going to talk a little bit about this thing as well after lunch I would guess. So for example if we're trying to recreate the population at an urban level, so here's an example where the starting point in our simulations is to try and recreate the population of a city or a region. Now I asked a question of the presentation about representation because interestingly what we're actually trying to do in our simulations is to represent a complete population, so in this example here this is just a map but what underpins this is an attempt to represent all the individuals and households in Leeds as what is sometimes referred to as a synthetic population. So we've got imaginary

representation of the entire population of Leeds, we're not trying to identify particular individuals or households but in principle there's complete representation in there.

OK and then we can do, so this particular example we can read the structured population, typically we go right back to the 2001 census because that's our most reliable data source for this kind of work and then in that reconstruction we can interrogate that synthetic data if you like to ask questions that we wouldn't have known from the original data that it comes from. So for example this one here is looking at multiple deprivation, I think that relates to people who are elderly and co-dependent, don't own their own cars, have housing problems and so on and so forth, so this is a kind of a social services kind of a slant on that.

OK so then the second element as I say is trying to look forward to this, so this is, so we have some kind of dynamic model in here, as I said this goes forward to 2031, so for example we can look at on the left hand side there what does the population structure of Leeds look like at the moment, that's the elderly in particular and where's this population going to grow, well it's going to grow everywhere but where is it going to grow particularly and what will that distribution look like as we move into the future?

The third thing we want to try and do when we've built these representations and these forecasts is then to say things about infrastructure and service provision, so for example here's something which is taking that simulation as we go forward in time to 2031 and saying OK well if we interface that, what is essentially a demographic projection with something like a transport or a traffic model, so this is taking actually a well known piece of transport modelling software, something called OMNITRANS and it's taking data from 2001 in the top left there and then assigning these, basically it's got demographics in there, it's got jobs and so forth and then saying OK so what's the utilisation of the road network in the area look like at this time, and then how does that develop in a year, up to the year 2015 until the year 2031, assuming that the traffic infrastructure stays the same. Any of you know the Leeds/West Yorkshire area, so this, well this bit going down the, is that going down the eastern side, you've got the A1 going down there, that's the fat red road, we ?? (08.04) on there and along the bottom here we've got the M621 leading to the M62 going over to Manchester there. So anybody who knows this area particularly well would be quite frightened by this illustration because you already can't move on the M621 on a typical morning and evening and pretty much going to get a lot worse as the population grows over the time period of the simulation. That's a by the way, I don't really want to talk about the substance of these simulations.

But that's the kind of thing and then we might represent that in a different kind of way, so a lot of our friends who are working with us on this particular application interested in not just congestion but impacts on air quality and things like that, so we can start to look at OK how does that, how do those patterns themselves start to impact in terms of what are we looking at here, OK well average speed, so the network's getting more congested so the traffic's moving less quickly and, well it isn't shown here but that then as I say feeds into things like air quality, pollution, issues of that kind.

OK, so those are the things that broadly speaking we're interested in on the project; if anybody's interested in the technicalities of that I can certainly provide lots more details. But I want to take a slightly higher level kind of view of what's going here for present purposes. So in terms of features I just want to note two or three features of what we're trying to do here.

So the first thing I want to point to is that we really have quite a substantial requirement here for processing and storage, I think actually the previous two speakers have referred to demanding kind of processing requirements at various stages. But because, I mean it's partly because our model is, well because it's spatial and because we're representing a lot of individuals and households in these simulations, is that they're quite demanding in terms of the processing and storage and this relates to both the reconstruction element of the simulation and also the dynamic forecasting component of it. Again I mean just to take one aspect of that, so if we said, looking at Leeds, if we take typical forecast so well 730,000 individuals in Leeds at the moment but that's going to grow over the next 25 years, let's suppose we're dealing with something like 30 individual characteristics, we might be taking single year simulations so that's say 30 time periods and we might have a number of different scenarios, we actually want to look at how the population develops under various assumptions about where people put their housing or how the transport infrastructure develops or what have you. I haven't done the maths in this particular example but I know we worked out one or two cases a little while ago and you could end up with quite a few terabytes of data to actually manage you know purely in terms of being able to you know kind of capture and archive the outcomes of those sorts of simulations, and that's just looking at Leeds, I mean we're actually running this model on a, or it's available on a national basis. So there's really, and it's partly because you know

obviously as you're aware you know simulations you're looking not just at kind of the base stage if you like that you start with but you know this idea of scenarios in particular you know is that one can start to generate all sorts of versions of that data and the whole thing can mushroom and expand really rather quickly.

OK the second feature which we're particularly interested in is that we're not, I think what I'm going to come on to is that again in contrast perhaps to both the previous two talks is we're not trying to build like a desktop version of a simulation, we're trying to build something, so this idea of the infrastructure is thinking about making simulations that can be more generally available across you know wide communities of users, so potentially almost to anybody, so one of the things that we're thinking about here is using different data sources, pulling these into our simulation but again not pulling them into databases and then feeding them in to a particular piece of code, but actually being able to access in some more general purpose kind of way different sorts of data sources, and there are quite a lot of these different things. So again the basic sort of population reconstruction is based around a sense of smaller statistics and a sample of anonymised records, so it kind of synthesises those two different sources. But in terms of the applications and in terms of the dynamic modelling we're also looking at things like the special migration statistics, various kind of ONS vital statistics, you know fertility, mortality that kind of thing, and things like BHPS, General Household Survey, Health Survey for England, so lots of kind of different survey data sources about people's activities and behaviours, you know we might be representing that in various spatial forms, so using things like mapped boundary sets and you know all these things come from different formats, from different kinds of places. So that's the second feature that we're trying to accommodate in this work.

And, sorry this is an awful slide, I got back from Athens some time on Tuesday, I can't remember when because I'm a bit scrambled, but I know I got up early and I was there for a long time and so I had to put some slides together in a hurry, so this one I didn't have time to tidy up so I apologise, but just in general terms what we, you know we want some kind of capability for interrogating our simulation and then being able to present that information in a variety of ways. So as I've said already we want to be able to produce maps of distributions, we might want to produce you know charts and tables, and again this is very similar to what you've seen in the previous illustrations, we might want to produce you know design reports for particular users who are interested in particular issues.

So what we've been interested in doing in this project is try to put together and OK so this, I mean the other characteristic I guess of this project is it involves a combination of, I'm a social scientist, a geographer, so I'm interested in the simulation aspect of this, the geographical and social aspect, but it's a partnership between social scientists and computer scientists, so the computer scientists with us are trying to work on the more general what they like to refer to as 'architectures' that actually help us to put these things together. So this is an architecture that they originally designed to support these MOSES models. So it basically comprises, I'm going to come to talk about this a little bit more, in the pink there we've got what they call a 'portlet', so what I would myself a spatial decision support system, so we've some, or you know people talked about you know kind of user interfaces and that whole kind of side of things. So we've got the bit that people actually haven't used there and that in itself is made up a series of different components which again I'll touch on briefly in a moment. And then I mentioned this issue here, so we've got, with these portlets themselves sitting on various different data sources which are shown down at the bottom here, these, so these data sources get mangled in various ways by the simulation; the blue bit is showing us how we can store that information in various kinds of location; and the yellow bit is access to high performance computational resources. So we're looking at the problem I mentioned before that we need quite substantial processing power, so we're trying to actually access that from things like grid resources at both a national and European level, so big clusters of computers basically that can make these things run in a reasonable timeframe, at least in principle.

So that's a version of an architecture, as I say we've got computer scientists on the case which was a mistake because then they go away and endlessly amuse themselves by playing with architectures and representing things in different ways and so on. But this slide's actually saying something, well it would be saying something quite interesting, I'm not sure if I'm intelligent enough to interpret it. But this is, so, a quick digression? Yeah, does anybody know anything about grid computing in the room at all, is that something that's in anybody's consciousness? No. As I say, OK, so big computing, this is the e-Science thing that's been going for about the last seven or eight years and actually the Govt and the Research Council have pumped enormous amounts of money into it, three or four years ago I saw that the figure was about £340 million I think in terms of Research Council investment which is absolutely enormous, and it's basically, it's the idea that you can set up huge parallel computers basically. And so there are a number of these big grid computing facilities that tend to get used by the physicists for processing the resource from their you know particle acceleration experiments, although quite a lot of them are used in biomathematics and that kind of thing. Sorry that's digressing you don't really

need (mumbling) … So people used, so when we started this project people were thinking about big parallel computers and how could you use those to do things like run simulations effectively and efficiently? People these days are starting to talk more about something called The Cloud, actually being able to use, to connect together individual PCs that sit on people's desks, but the other thing they're talking about a lot is this idea of services. So one of the things that our computer science friends are now looking at in terms of these simulations is how can we disaggregate the various different components of a simulation into different kinds of services? So these are individual components that you can disaggregate and then connect together and they can talk to each other in various ways. Again I think this idea of services is implicit in some of the things that we've heard about already, this one here is phrased in terms of analysis charting, mapping, those kinds of functions, but equally when you come to look at the next ?? (18.21) so actually a simulation itself, then you might be thinking about OK, well how can I disaggregate the different components of that simulation so that again they can be maintained, updated and all the rest of it? So let me just mention that feature in passing.

So what does this thing look like? Well again the idea of it is that we can take, we can develop these kinds of services and components and then actually integrate them in different kinds of ways. This is a little bit of an old version of the system, again I'm ? (18.57), computer scientists they like to deal in kind of you know high level concepts and things, I say to them you know can you produce anything that's a bit nicer for me? They tend to say that's not very interesting and I say well it might help you to sell it a bit. But anyway, so this is like a version 1 of the system, version 2 will be available soon, but it's essentially what I would call a spatial-decisions support system, so what happens is you've got two tabs across the top here which actually do various different things, so for example what we would get here is just what we call a selection portlet and what it allows you to do is to go into a particular area, pick out some areas in green that you're interested in and ignore some others that you're not. So that's just a selection. Then we can do things like look at different kinds of interrogation of what's in the data, so we're looking at here, no idea, provisional of social care versus social grades, OK, so that's just looking at, it's essentially interrogating a micro-database to say tell me what's the distribution of people with particular socio economic characteristics and particular needs for, as I say I'm not sure if they're using or providing social care, but one of those two things anyway.

Here's another example, so here we're looking at diabetes, so we've got simulation people, we've then taken some further information about the distribution, the likelihood of people having diabetes both now and in the future and we can interrogate and tabulate that kind of information.

Then we can start to think about knitting together different sorts of things and to, for example, to represent our information in different kinds of ways, so in this case I mentioned the kind of maths before, so here we've got shown in red some outputs from the simulation, it's forecast to the year 2031 of the elderly population and I've pulled in something you recognise there, the Google map just as an overlay or underlay for that data, to give it some more context in geographical terms.

Here's an example where we've done something similar with Google Earth, this is actually a transport simulation I think which is looking at changes in congestion under a particular scenario, development of a tram network I think in Leeds, again the detail's not important. And so again we can start to inter, the point here is that we're going, it's taking information from our simulation and we're then putting in other information from Google, from Google Earth and knitting those things together in terms of this application.

And then we can pull together bits from different parts of that architecture if you like, so again you know maps, charts, table sorts of thing and the idea here is that we can design particular outputs for particular users, so lets say I'm someone who has decided to open a new hospital in Leeds shall we say and I want to see the impact of that in various different kinds of ways, then that's maybe the sort of thing that I want to do.

OK so in terms of plans and prospects, well I mean I raise this just because I thought it might be interesting if you like for the point of view of this audience, it's certainly a good opportunity for me. So the third project I want to talk about actually started yesterday in theory and it's called The National Infrastructure for Spatial, actually that shouldn't be Spatial Simulation, it's the National Infrastructure for e-Social Simulation, a little bit of a slip there, so certainly it's not restricted to geographers and that kind of thing. So this is a project that's been funded GISC under something called their Information Environments Programme, as I say it started yesterday, they decided to spend a couple of million £s and so naturally they told me about a fortnight ago that they wanted to start, so as I say in theory it started yesterday. And so it's quite a big project, we've got about eighteen man years of effort to extend in this. The objective of this project to produce

production level simulation tools and services, social simulation exemplars, integration of tools and repositories, blah, blah, blah, those things, but rather than kind of try and work through all of those in detail let me just try and give a big picture view of what I think we're trying to achieve in this project.

So one of the things that the National Centre for e-Social Science has created is something that you just call the NCESS portal as it says in the, towards the top left of that screen there. OK so basically this is, a portal is a web portal, it's, you can see there's a user ID and a password there, so it has regulated access to this thing, but basically it's an attempt to create the beginnings of, in those GISC terms, an information environment or a virtual research environment for people that are interested in e-Social Science, and so social science computing if you like in crude terms.

So what can we do in NCESS, well as we go along this is what faces me if I go along, I log in and so this portal actually provides me with a series of kind of capabilities or functions on projects that I'm associated with, OK. So the CSAC here is actually the research cluster that I'm associated with at the University of Leeds, ESRC EIP is another ESRC project that I've not talked about today, I've got GENESIS which is a work site for that ESRC project and GISC A2 is actually the work site for the project that I'm just talking about here. So I've got a series of different projects there and I can do various things. So for example I can access, OK this is what I get in the ESRC, the infrastructure project, I can get general information about that project, I can go along to the information centre and pick up news, the latest activities that are going on that are relevant to this group. I can access shared resources, so for example file store which again is shared and all of these tend to be multi institutional projects, so I can access code, minutes for meetings, all kinds of things that have been deposited here by the various different project partners.

Now this kind of, you know these kinds of portals, information environment have somewhat facetiously sometimes been referred to as you know kind of Facebook for researchers and I won't bother with the survey in the room but it's something that's more relevant for my children than me, but you know it's kind of, so you know the criticism is potentially you know it's very interesting, but you know you've got chat rooms and you're setting the environments for people to talk to each other and you know is it just kind of a posh website kind of thing at one level.

I think what's interesting and what we want to try and get towards in this new project is actually say well can we take things like the MOSES portlet or a simulation portlet, this support system and actually embed them in something like one of these portals. So it actually gives people the capability to come along and to actually run simulations and take results from them and maybe take those way again. And then if we say OK if we can begin to do that, can we then take it onto the next stage where we're not just placing our own services if you like in those terms or any applications within these portals, but we might actually have a community of people who are coming in and putting their own services, components into this portal as well, and having the ability to be able to integrate these things flexibly together, which is kind of the question that Belinda was touching on, the previous speaker about you know can we start to you know exchange information between applications and so on and so forth.

OK I haven't a huge amount of time but just, so this is, no I won't dwell on that actually, that's loads of different elements and the packages within this actual application, but let me just focus again, I'm just trying to say something about the general concept. So if for example if we were trying to run you know what I've talked about as a population recreation model, so what I might be doing is taking a piece of data and then typically taking a piece of data, applying method to that data to produce some outputs which I might then take on to use in some kind of dynamic model shall we say. Then what I might start to explore is to say well actually there are a variety of different ways that I might, different procedures that I might use to actually go about this process and I would then want to evaluate them in various kinds by looking at spatial distributions, by looking at reports, looking at various kinds of statistics, telling me whether my model is particularly effective or not. And so you know so I might have a kind of a use case of this, you know this activity saying you know that allows me to put together all these various different things to do that sort of work.

So you know again, and the way that computer scientists would deal with this probably if they designed things that they called work flows which would basically say, if they saw a work flow through here, is to say take some data, run a model, run some analysis on it and evaluate the results. The other idea which is starting to emerge here which is quite nice but I don't fully understand myself and haven't really got time to go into, the idea of a research object is you know whether you could actually take that kind of a piece of analysis and then start to annotate it, and people are now talking about actually publishing, this is quite established in biomathematic publications, starting to publish the results of simulations in this kind of environment, so you could actually annotate and publish the results as a way of sharing experiments and again this is,

this idea is really taking hold of things like biomathematics and those kind of applications in kind of contrast to the standard academic approach of having to publish everything in papers that are very slow to referee and all the rest of it. And then building those individual elements up into these kind of portlets, these applications.

Let me just see what else I want to see. Oh OK, not too much now, just a couple more things.

I wanted to make the point, OK, the way GISC works is that if you want to get something funded by them you have to show that lots of people actually want to get engaged with this activity, so the partners in this project, it's being led by ourselves at Leeds but it's involved Manchester, Southampton, UCL, Glasgow, Daresbury and Sterling, so it is quite a bit multi institutional project and it has a number of people who said they'd support us both within academic communities and user communities and also the infrastructure service providers and stakeholders. So for example ? the people that provide the census data, oh we haven't, yeah ESRC, UK ? Data Support Service, all sorts of kind of data providers, map providers, those kinds of people.

I won't go over that one because I haven't got time and that one actually, let me come straight to this one.

You know the point I wanted to make is that, so the National Centre for e-Social Science has been put together these sorts of elements of a platform of a research infrastructure, you know the idea that we can get together these portals, that we can populate them in our work with ways of accessing data, with ways of accessing simulations, ways of archiving those results, but where, oh and so we have funding to actually, to implement if you like some kind of version of these technologies, to get together some sort of social simulation portal, virtual research environment, virtual organisation, virtual community, call it what you will, I think that where people have struggled so far on these projects and the NCESS has been going for three and a half, four years now, is that they've done better on technologies than they have with user engagement and community engagement. And so I'm very keen to try and get off to a good start on this particular project and you know to be able to engage with potential users of this sort of technology. So I'm just hoping we might be able to have some conversations over lunch, later on in the break out sessions or what have you, you know I'd just be very interested in people's ideas about how you might engage with an activity of this type. And you know certainly I'd be hoping over the next three years to try and build up the users of our infrastructure, of our information environment and clearly people that I'm, well I'll be looking to nail people to try and get involved I think(!) at least as you know kind of registered users, some kind of system, and you know hopefully to think about developing this in accordance with your skills and interest.

Right I'll leave it there.

QUESTIONS

Paul Williamson – Thank you Mark [CLAPPING]. Again any quick questions to Mark at this stage? Can I just ask one then? If you're looking for people to suggest models, I might have bolted in to your infrastructure or …?

Mark Birkin – Sorry say that again?

Paul – You're looking for people who have models bolting into your infrastructure or data or …?

Mark Birkin – Yeah potentially, one of the slides I glossed over was we actually, where we envisaged different kinds of users, so I mean basically there are three kinds of, well yeah three kinds of partners. So one is what we call naïve users, so for example if I develop a portlet and I say it's for shall we say healthcare planners in Leeds where we do all the work for them, what we actually say is OK what you can do is you can look at what's happening to 2031, you can ask certain questions about your hospital reorganisation you know new, you know health centres for the elderly or whatever and they can kind of interrogate it and get the answers sort of thing, but they're not playing any part in the active development of it. And I think we're looking for if you like the research users, so people who are maybe interested in the actual, you know the mechanics of the models themselves and you know how they get developed and how they're used but almost in kind of a, you know it's a partnership type context I guess saying you know OK what you need to do with the models is this, this and this. And then the third level is you know what we call the sophisticated or the PARRY users but I mean it's not clear that we, how we actually do this yet, but in principle yes, how you could actually bolt on your own models and you know introduce them alongside our models, but you know clearly in order to do that there are all kinds of questions about you

know the standards that you adopt and the protocols and the way that these things talk to one another and so on and so forth. So you know we recognise that those problems are hard ones, but that's the long term vision I think, that you're actually trying to open this thing up so that everybody can contribute to it and then you can start to explore the power of flexibility of these ideas about you know work flows and recombination of services and components and things, you know so that the whole thing can kind of take off. It's very ambitious I think, but it's you know … And a lot of issues about you know stuff like intellectual property and so forth that other speakers have referred to but you know from an academic point of view you know we don't worry about stuff like that because we're publishing it all and what have you anyway.

END OF RECORDING