ESRC Microsimulation Seminar Series

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

Microsimulation in Canada: Lessons from Statistics Canada

Chantal Hicks

[EDITED TRANSCRIPT]

So I want to start by saying that although I talk to a lot of people, the opinions of lessons for success and stuff are very much my own and not necessarily those of all at Statistics Canada. This presentation will have a slight bit of emphasis on the models with which I've been involved.

I've been at Stats Can for over, about 15 years now and I've been involved in I would say the majority of the models there other than our personnel models. So I will have covered at least all of them slightly.

So before we go about us, I was asked to talk about Lessons from Canada. Maybe I could just talk just briefly about micros and the rest of Canada. By the way, I'm so sorry the whites aren't showing on those slides. In terms of Govt of Canada there are policy groups who have their own models, you know we have static tax/transfer models both at HRSDC, our human resources department, the department of finance has their own, there are dynamic pension models, the Health and Public Agency in Canada have a variety of models of their own.

I come from Statistics Canada. We're a statistical agency, so we do the official stats, we do the surveys, the censuses etc. And overlap exists amongst the microsimulation models, the same as here [UK], so basically we have at least 3 different models that calculate taxes and transfers starting from the same survey. And there is talk about us sort of perhaps combining some of the efforts because a lot of the time is actually spent, I'd say more time is spent touching the survey than touching the modelling bit and perhaps some of this could be combined. We've never been successful, for different reasons.

In terms of Statistics Canada we are one of the larger microsimulation groups in the Govt, in part because we have all the data, a point I missed off my slide - but it's really a key part of the success and why we do this, is that we have easy access to data. But even within our group it's all split up in terms of we have groups that do more health and disease models, some do more demographic projection models, there is a personnel modelling group and then there is a modelling division and we do the modelling technologies, most of the socio economic models, and we also do the infectious disease models. And the groups are all in close collaboration, with quite a migration of staff between the groups.

This slide shows just some of the major microsimulation events since 1984 and I'll go through some of these, not all of them, and this is only major, there's many, many things missing and the colours don't show up right which is a bit sad.

Our first model is something called the Social Policy Simulation Database Model (SPSDM). It was the first model developed at Stats Can. Like the ones this morning it's a static tax transfer model. It starts with survey data, but as this it wasn't enough we created a synthetic database to do all tax transfer issues. And it was originally written in C, then imported to C++. This model was always run on microcomputers and the visual front end was released in 2006. The model basically starts off with Michael Wolfson, who leads microsimulation at Stats Can, coming in 1984 with the concept and the funds. It first saw public release in 1988. So this model with a front end is so old basically to the public, our Govt departments use it, policy groups use it, provincial governments use it, but most analysis is done outside of Statistics Canada. And it basically stayed the same; that course hasn't really hasn't really changed in 20 years, it hasn't had to. Finding people couldn't cope with the DOS interface so we gave them a pretty interface in 2006 but it's still running the same model.

Now in the early 1990s there was a project that was trying to do a hybrid of static and dynamic models and it included FADEP which was a dynamic model which projected family structure. It had worked on a closed population, was written in C++ and it came in around about 1992.

With the population health model called POHEM, this is a dynamic model of health and disease looking at the cost of different healthcare interventions. It's a cohort model, which started off as a single cohort model, although it's now an overlapping cohort model. It has had multiple versions throughout the years, some have used synthetic data, some have started from a survey. The very first version of this was started in APL, so the demographic stuff actually was 1983, and it was basically redesigned in APL 1988.(05.04)

And so by 1993 Stats Can had gotten a few lessons learnt. The SPSD/M was very successful. The dynamic models, and by that we mean POHEM and FADEP, we were finding were very hard to maintain and they were hard to modify and we found that the coding of them was just really very time consuming and error prone and just taking a huge amount of time. So we decided we wanted just a new interface to make this easier because we were going to get money to redevelop POHEM and we really wanted something with a flexible design, we wanted it to be modular, we were finding that writing to disk was just slowing everything down so we wanted everything in memory as much as possible for our models to run quickly. And though we still have both dynamic models we found that they were making just such high demands on memory or if you had a smaller sample there was a high sampling variability, that we decided if we don't need a closed model particularly for the questions we want to, why bother? And so we have a preference for overlapping cohort models but we do, for some purposes still used closed models.

And so we developed MODGEN – a generic tool used to create dynamic microsimulation models. It's basically a C++ compiler pre-processor and I don't know quite what that means [LAUGHING] but it's C++! Like from my perspective I code things, it becomes C++ and compiles. It basically has some handy features; it creates and stacks event queues, so you know as time advances it keeps track of everything for me; it gives us this nice graphical user interface; it derives states that you don't have to code that in yourself, everything is maintained for you; it contains tabular language, we do on-the-fly- tabulations, we don't write to disk; you can do continuous or discrete time; you can do cohorts or closed populations, it does, it's multilingual, it gives documentation, it's a nice piece of software.

And the first working version arrived in about 1993/94. POHEM was actually the first model that was redesigned using MODGEN, and the POHEM redesign actually funded the creation of the language to some great extent.

FADEP – basically we dropped this model and we replaced it with LifePaths. It's a dynamic model, individuals and families, it's designed to analyse Govt programmes plus demography and family structures. It's an overlapping cohort model, so we dropped the closed population. It gives you a full cross sectional population by 1972 so we started creating things, we started creating people in 1872 and growing them up. The data for the you know early, you know late 18th century, 1800s, early 1900s there's shall we say has gaps! But it's the best we know. And we added pensions from the get-go, so the data gets better as people get older, the cohorts get older.

The first really big project was student loan reform, but it's been used like at pension policy, time allocation, intra-generational issues and it was the other, it was the first big model written in MODGEN.

And POHEM and LifePaths still exist to this day, I mean we still use them, they've had things that have changed over the years, but they still exist.

In terms of the next really big new thing was DEMOSIM and that's actually happened in the past 5 years. It's a demographic projection model. It projects the population by visible minority status. It starts by reading the entire census. The first version was 2001; as you will know it came out before the 2006 census. It was called POPSIM at the time, but there's a new version now, DEMOSIM, that is going to read the 2006 census. And as we're using MODGEN it's easy to write. So that happens, the first version came in 2004, we did this as part of funding from the Department of Heritage because they were very much interested in projecting visible minority status and cultural status in terms of the future.

CAREMOD – which is not the official name of it, it's our temporary place holder – is a cancer model, focusing on the cost effectiveness of cancer treatment and prevention. It's based on POHEM. Basically this project has about a 6 to 8 month turnaround time, meaning we got the money in March and we're releasing at the end of October and it's going to be web ready. We could not do that if it wasn't based on POHEM; you have to have a starting point to do something that quickly. And a huge part of this contract is actually putting it on the web. So we're basically doing this global interface, so all MODGEN models, inputs/outputs will be web ready in the future. And that's there.

The personnel model was done by a completely separate group, looking at hiring, retirements, maintaining personnel needs. And they resisted for years, but in 2008 we convinced them that really the problems with maintaining the model would be solved by moving to the MODGEN platform, and so we have a prototype which was created in a few weeks for them to show how easy it would be. They now want funding to just migrate and they're looking but they've been convinced that this is the way their future. So that's been exciting. But they started way back in 1992 and moved.

The green names on the slide were MODGEN type models. There were others which we are going to talk about, such as PopModM, RiskPaths. We also have, and these are supposed to be in a different colour yet again, those are all, all those green MODGEN models were basically cohort models or case based models. We are also building interacting time models so we have experimental, economic model there. In 2002 IDMM became our first large scale interacting model, it was for infectious diseases. We then collaborated with the World Health Organisation and academics in South Africa to do HIV tuberculosis model called HIVMM, London School of Hygiene and Tropical Medicine hired us to help work on a child vaccine model, CVMM etc.

So all these models we worked together and there are some key issues that are common to all of them. Well first funding, now we are part of the statistical agency so we have core funding which is lovely, for those of you have to find funding all the time, it's nice to have core funding! Having said that, it's not enough, we can't, we don't have enough to actually build all these models, so we still have to find other funding mechanisms for other ones, and different projects have used different funding mechanisms throughout the year. I'm going to sort of talk about some of the main ones.

Core funding has been really important for all, except for the first parts of DEMOSIM which was mainly done completely on a contract. SPSD/M is the only one that sells to clients directly and the others often work in collaboration, whereas clients will give us money to do research or build enhancements and we may do runs for them. For example, for othe LifePaths model, HRSDC has been the major source of funding throughout the years and it's ended up being replaced with something called GAPS, which was a project to give funds to Statistics Canada for surveys mainly, to fill the information needs. So government departments got together and said you know we want this survey on a new time use survey or we want a survey on pensions. But actually they ended up also giving LifePaths some money in terms of this; integrating all the surveys together in one model was important, we got money through that. Now those funds ended March and we have yet to find a replacement for them.

The SPSD/M has core funding in sales to clients but we can't survive, the sales aren't enough to maintain the model. We were, asked in 2004 to become self-sufficient and studies show that we would fail; clients weren't willing to pay enough money to make that completely cost recovery and we were allowed to survive without it, in great part due to the Department of Finance who had just given Statistics Canada and were quite upset they were going to cut them all, or were thinking of cutting them all. So they actually provoked the survival of them all.

And for MODGEN, core funding is important but development of other projects often involves basically money to maintain MODGEN, else how else do you maintain these other parts? So POHEM started the, funded the creation of the model, the WHO project funded improvements to interactive populations and CAREMOD, the cancer model was funding the web interface.

Now we are the statistical agency and we do some research, we have a small research area within the agency but the agency's main focus is to build surveys and census etc. But nonetheless some of our models were mainly used by us to do research on our own as part of our mandate to do research for Canadians. But other models are used by the public and most models are run by Statistics Canada for projects funded by external users. SPSD/M is released completely free to the public, LifePaths is on the web for free and MODGEN is also on the web for free. POHEM, DEMOSIM and PERSIM aren't actually publicly available at this point. Having said that, CAREMOD will be, DEMOSIM is going to have a public version in 2010, although it's not clear to me that all of the data underpinning DEMOSIM can be made publicly available, so it may still be in locked research centres because it relies upon the full census which is not publicly available. And the personnel data will always be confidential, so there's nothing much we can do about that. LifePaths is completely open to the public though we don't give all tables out – people will have to print their own tables.

In terms of major users, well there are external users here and there for the publicly available models, Statistics Canada also operates some of the models on a clients behalf. Having said that, for tax transfer model, it's the only model where external users are the majority. Indeed though MODGEN is close behind, it's increasing as modellers, researchers, academics will do models on their own behalf using that tool, it's not a model but a tool. And LifePaths, those available on the web has had very very few external users, in part because it's much more complicated to use. In our tax models the one that has the least amount of use in house and that's because we can't analyse policy, so it's hard to do a lot of analysis on a tax benefit model when your organisation isn't allowed to just comment on policy. So it makes for, no it does give us a nice hands off thing, it's like yes here it is, here's the tool, not used.

Now external users are, they're all Government departments, some of whom have their own model, so it's used as a compare and contrast; provincial Governments who often don't have the funds to create their models, but we also have academics who give them out for free as well as policy groups. And it's always a good day for me when I see someone from the very left and someone from the very right, using my model. So it's a flexible tool and the question you ask gives you very different answers, and if you want to grow a recession or not, you can and it will give you very different impacts.

In terms of staffing, we've had different groups. One of our key distinctions though is who touches the code? Do you programme your own model? Do you have one programmer and you have different people give equations to the programmer or does everyone programme their own stuff? In terms of SPSD/M and LifePaths we're in the same group and we really believe that you should touch the code. It's hard to understand the model if you haven't touched the code. Other models, you know POHEM started off with one person programming the whole model, but it's moving to multiple. DEMOSIM has gone a bit of both, although mainly they'll have one person doing the bulk of the coding. In terms of speed if no one knows the thing it's easier to have one person become an expert but if you think in the long term it's really useful to understand what you're doing.

And so from all this history, I'll draw 10 key lessons. And the first lesson – or perhaps I should say keys to success rather than lessons is that strong leadership has been really key. Michael Wolfson led, came to Statistics Canada in 1994 and basically led the microsimulation since that time, although at the same time we had Steve Gribble who has been basically designing the architecture of all our models and doing the leadership and Geoff Rowe who spearheaded research and equation estimation. And they've all been there since 1984. In terms of future challenge, Michael Wolfson has announced his retirement and everyone else could leave too. But still it's been key, it's been key to how this long term leadership comes to get funds and have believability behind them.

Lesson number 2 – or reason for success number 2 is you keep staffing. Now it takes time to learn it, like yes we also have recruits for every 8 months and you know come and go and we lose staff, but to become a really, an expert in the thing takes time and it's really much easier if people can stay for more than one year or 2 years. Now this, we found, is easier if the team is big enough that people can migrate, so at least you can keep the skill working on different projects so it still feels new and exciting as opposed to being I will do the same exact thing for 20 years. So we find that you know you can keep the skills and work on other projects within the team, and it also means in crunch times we're looking to pull back and pull off. So using the same technology for most of our projects makes this a lot easier. And again, at least in our group, I'm in the modelling division, we find it useful to create your own code, so that also means there's not one person who has the code and they leave and you have to learn it, everyone knows their own code and you move around a fair bit. So that's been very, a key to our success.

The third one is MODGEN, investment technology. The fact that we had basically this dynamic microsimulation modelling language has meant that we can develop new models really rapidly, we can do prototypes really, really quickly which means that what's left is actually the data. So you can sort of give to a potential client this is what it would look like, now the hard part is actually doing the equation estimation and putting it in, but you can get an idea of what the inputs and outputs would look like with fake data and it just, it has huge marketing appeal. It's like oh, you know, then he says oh it's actually 8 months to do the data analysis behind it but we think that's where the time should be spent, is doing the data analysis, not actually creating the model per se. It's also made like a really flexible team that move from different projects and yes, the data analysis is where the time goes which you know, as a data agency of course we believe in data analysis a fair bit!

Having said that, and this is really where the things that might appeal(?) perhaps more than Stats Can is the whole, is that if it ain't broke don't fix it. SPSD/M was written before MODGEN and it has been stable for over 20 years and I must admit I currently manage that team in addition to working on some of the longitudinal models and I've be resistant to change, it's been stable, our clients see the code, they read the code, changing that just for change's sake means everyone has to re-learn the code and stuff, it's really useful for them to see it. It's change, changes in operating system, we update the tax code twice a year and we have external users from across the country who use it, so it's been incredibly stable. So I've been, we've been, resistant to change.

The visual interface though, we were flexible on that and it was created in 2006. There was talk to migrate the whole thing to MODGEN at that time. This is still debated, in fact probably the hotly contested issue within our group is whether not going over MODGEN was a good or bad idea. We didn't go to MODGEN because it would have been about 6 months of MODGEN redesign work and we didn't have the staff – , you know it was like our priority is we release on schedule. So we've kept to this new interface, still debatable.

Another lesson is to survive through lean times, as policy needs ebb and flow. The tax model is 20 years old, has never been as popular as it is right now, in part because we've had a string of minority Governments who are interested in changing sales tax versus income taxes, it's one of the big debates, who knew? This came right after a big funding crisis and also became popular. You know so it was useful. Core funding to maintain us through those low times has been really critical. The more diverse your client base is, the more the ebb and flow ranges of people caring about different policy dimensions means you can survive. And loyal clients can save the day, as they did for our tax model.

The reason you want to survive is we really want to use strategic opportunism, as my boss Steve Gribble likes to say. MODGEN's useful, as we can do new models really quickly, so it's been very key, but again the data analysis takes time. So the trick for us has been to build on existing models, you want to build an enhancement to LifePaths or you want to use POHEM as a core and build stuff on top of it. But these models take time to build and do the data analysis and if you let it run fallow for 5 years everything is out of date and it takes another year to build up. So you want to be able to keep maintaining these things even if the use isn't key right away, if at all possible, and we've been able to do that. And you know we currently have this big project with CAREMOD and LifePaths has been used to study pension reform and it's been very key.

Having said that, these big complex models have a cost. LifePaths, and I'm going to talk about LifePaths because I know it the best of these launch models since I've worked on it for many years. They're hard to keep up to date and they have a large overhead, and when they get the complex models it's hard to understand what's causing your change, because everything's interacting. Complexity also means that it's been hard to track user's LifePaths as opposed to some of our other models in terms of you know we do projects with other people but sometimes it can be so complex that we have found it to be a marketing challenge compared to some of the other models.

Having said that simplicity doesn't always work, I mean if it doesn't answer the question, you know I'm sorry these longitudinal you know pension issues are complex questions, people's circumstances change and we model them. And even something right now, the Social Policy Solution Database, wide user base, well loved, cannot answer the biggest question of the day in Canada which is employment insurance reform. They're trying to expand the programme and we're based in part on administrative data. They're not there, it's hard to do and it's behavioural and longitudinal and not calendar year based, and yet everyone wants to use a static calendar year model to do it.

Links to clients and external groups are essential. We've had strong links for most of the models. LifePaths I would argue has had really strong external relationships and sometimes a little less strong. And it's been one of our bigger challenges to keep that going. And our links to the microsimulation community, we're thinking in terms of keep, if we could at least get people to go from one model to the other, I suppose go into policy would be useful to us or trying to promote those ties as much as possible.

And this leads me on to my last lesson, and this one is dear to my heart, although arguably this is the least one. Pick a good name! Now we have outreach in terms of our static model, we offer courses twice a year, well four times a year, twice in French, twice in English, we have hotlines, if you phone up there's a line that goes on all our phones, it's exciting. So every single day pretty much, well I mean not every day, twice a week I have to pick up the phone and do this, SPSD/M hotline, it's hello, bonjour, put vowels, these are horrible, horrible names! They have no vowels, it's a pain to say. I mean it's shocking but actually SPSD/M in Canada is a brand name, because it's sometimes called SPSD, SPSD/M, SPSM, it's actually recognised which is a key to our success because branding wise, and it's different in the other language.

QUESTIONS

Questioner 1 – Very interesting talk. Could you say a bit about, my question is sorry that Statistics Canada has, it's a big thing in Canada in terms of microsimulation but also kind of general positive analysis. Is that the correct impression, so in other words comparing you to both Government and the user community, would you say you were the dominant force?

Chantal Hicks – No, I mean in terms of data like, data yes obviously.

Questioner 1 – But in terms of ?

Chantal Hicks – The dominance, well we probably are the biggest in terms of number of people in models for microsimulation, yes. Though other Government departments have their own, sometimes they use ours in addition and there's no NATSEM in terms of centre for microsimulation in Canada, so in the academic world, so there are different academic users for microsimulation. So in terms of microsimulation, yes, in terms of data analysis, no, I think academics would be wider range, like we have a small analysis department and it's always, well in terms of strong leadership we have the same chief statistician for I think 30 years, he just retired at the age of 75 or something, like so we had, and he believed in analysis, so our analysis got stronger. But it's always problematic because we can't analyse policy. So that's very much outside of our range, so you can analyse you know demographic and family trends etc, so no, I would say in terms of research academics do more research and the policy departments also. We do have a small research group and the microsimulation fits into that.

Paul Williamson – Could you explain a bit more about the problems of overlap, not even the problems, but overlaps between models and between different Government departments and Stats Canada?

Chantal Hicks – OK, well partly, the SPSD/M was built after some of the Department of Finance had their own model, so we were not the first, we were about, it was similar timeframes, but our goal was to get outside Government so people could use it. And at first I think some of the other Government departments weren't that keen and now you know we're, survive long enough and things change. Some of the other models have died and they've used ours. In other cases, the Department of Finance's model is built on income tax data, administrative data, so it's stronger in terms of population size, but weaker in terms family structure, because the family structure is poorly captured in admin data. Whereas the strength of basing it from survey data is our family structure is very strong but we have to impute some of the tax information in terms of deductions are very big in Canada so we don't ask things like how much do you give to charity in our surveys necessarily. So they can use it for benchmarking which is useful. In terms of models, having said that a lot of these models are built on exactly the same survey, our surveys at Stats Canada for privacy reasons, the ones that are released to the public do things like, they round all income data which is really awkward for microsimulation model. They also do things like set odd looking people to 'don't know'. So it's hard to do a model of taxes, in our cases our taxes are provincial, within the provinces I don't know. Like we know full well, you know when we got this survey, went to their house, we know where they live, but in terms of things released to the public … And so a lot of imputations happen and I mean we spearhead the effort to try and have links and sometimes they work, sometimes they won't. People want to know what they do and there is a thing that you know we are believers in our own things, of course buying a model when you can't programme it means you lose that, so I can understand why you'd want to do your own tax model. What I think the best link would be in terms of reducing that overlap is at least the data clean up because it's just very, very time consuming, and I speak from the data agencies, so I'm basically undoing stuff my colleagues have done to the surveys because I have access to the real survey, my problem is my model has to go out so I can't use it. So I can use it to inform imputations but I have to go to a committee every year and show how my imputations will not, you know will not breach confidentiality of our survey respondents. It's a huge part of the work we do and I think we're placed in the best position to do that since we do have access to more data than the other people. And that we'd be willing to discuss with other groups how they'd like to do it. But we, you know you get nibbles but after a while you're there, you're set up. One of those models still survives on the main frame, and at a certain point, I think it's one of the only things that's left on their main frame, you'd think at some point they'd want to move off but we've been unsuccessful. So I think in part because people want to do their own models, which I understand. But we give out the code which is a flexibility thing, so in terms, you know like how we've done something, the assumption, feel free, it's there, it's documented, academics can get all their ? for free on the database, what's missing is the data imputations.

Questioner 3 – Sorry just before you, where would you say the academic centres who do use microsimulation are across Canada, if there are not too many to give.

Chantal Hicks – Our links to them are weak. BC are the ones we work with the most often, University of British Columbia and Simon Frasier, sorry I'll take both of those together. Actually SPSM has recently been challenged by a SDATA data programmes that's been given out and in their advertisements they were saying it's like SPSM but free. Our stuff is actually free to academics. And the work involved really is a database creation, it's like yeah but the data has no deductions, this is, your taxes will be very limited. So …

Questioner 3 – And that's mostly tax benefit modelling that they've been doing.

Chantal Hicks – There's a lot of infectious disease modelling over there too. In terms of other, I don't know of anyone else who does tax benefit modelling in the academic community, it's not actually a big, we have research groups that are more policy oriented, they tend to do more of that, less so than academics. There are different parts who use things here and there. The academics are also involved in collaboration with, sometimes with us to do some of the health models, the population health model as part of its team has some different academics as well as right now CAREMOD has a bunch of oncologists on that team. So they then tend to publish, and I cannot remember where Dr Evans works, sorry.

END OF RECORDING