ESRC Microsimulation Seminar Series

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

Creating synthetic sub-regional baseline populations

Paul Williamson

[EDITED TRANSCRIPT]

Introduction

[1] The theme of this seminar is 'scaling up; scaling down' and I'm going to be talking about trying to scale down in a spatial sense, which Mark touched on a bit in the last talk. In other words, I'm going to think about modelling at a sub regional level. In his talk Mark showed some preliminary results from MOSES, and Belinda will be talking more about MOSES, which is a microsimulation model which operates at a city level. In this talk I want to look at one particular aspect of modelling at sub-regional level, which is trying to create some baseline population microdata which look like they are plausibly representative of sub regional populations. This is a topic I've been working on for a number of years. The particular results I'm going to present today stem from recent collaborations with Rob Tanton from NATSEM (University of Canberra) and Ludi Simpson from CCSR (University of Manchester).

1. CONTEXT

(a) What do we want?

[2] So what's the context? Well, first of all what do we want? What we really want is to try and look at problems at a sub national level or a sub regional level, so we want access to some microdata which are coded by sub region and include a range of relevant population attributes. So, for example, if our interest lay in health, we might be interested in where smokers and non-smokers live and how the interaction between smoking and income varies between different areas of the country. Or we might be interested in links with alcohol consumption or obesity and so on. In the toy example presented in this slide, in one sub-region everybody smokes regardless of income, whereas in the second sub-region nobody smokes regardless of income, reflecting differing interactions between smoking and income depending on where you live. This is, of course, not a real-world example, but it has been designed to illustrate the idea that there are some interactions between population attributes that will vary spatially.

(b) What have we got?

What we want is essentially the census on our desk. In fact, we don't even want the census on our desk. What we really want is a 2009 census on our desk that includes not only the standard census measures, but also all of the key variables of relevance to our particular study, such as smoking behaviour or obesity. We know, of course, we know doesn't exist. So, what have we got?

[3] Well I always like to show this slide to be reassuring (slide of blank screen). The point I'm trying to make here is that, from a starting point of nothing, any estimate, however crude, is probably going to be an improvement, provided the estimate isn't demonstrably biased or misleading in some way. In fact what we actually what we do have readily available to us are a set of anonymised government survey datasets, typically anonymised in part by the removal of detailed spatial coding. For example, from the 1991 Census an anonymised 2% sample of respondents has been made available, spatially coded to large local authority district. From the 2001 Census a 5% anonymised sample has been made available, also spatially coded to local authority level.

[4] illustrates just one extract from the 1991 Census 2% sample. The point of this illustration to highlight some of the issues raised by the types of estimates such survey data make possible. First, it should be noted that this is a very large survey, covering 2% of the whole population of Britain and this is the result for one area, for Leeds, a district with a population of around 740,000 people. In fact, Leeds is the second largest district in the 1991 Sample of Anonymised Records. So this is a large sample from a large district.

The table picks out those people in Leeds who are adults and presents a bivariate tabulation of sex by employment status, which is then converted into a conditional probability of being a woman given employment status. Adjacent to these conditional probabilities are confidence intervals for these estimates. Notice both that these confidence intervals are relatively wide, and that the numbers of survey respondents falling in each table cells is getting small (e.g. only 31 women on a government scheme), even though this is a large survey, for a large district, and the tabulation itself is very simple. If you start thinking about analysing minorities, such as ethnic minorities, the numbers in this tabulation would rapidly become even smaller still. If you then start to think about proceeding to a more multivariate analysis, such as employment status by illness by sex, or anything like that, the frequency counts begin to fall even further. And this, is the case even for a large (2%) sample. The same principle applies even if you move on to use the 5% Sample of Anonymised Records from the 2001 Census. These results push us into thinking that using a sample survey, of whatever kind, to estimate unknown local area interactions, is never going to prove satisfactory.

Am I over exaggerating the problem? No. In fact it is arguable that I'm understating the problem by sticking to large surveys for large areas using minimally multivariate distributions of common population sub-groups gathered only infrequently (once a decade). In short, surveys alone aren't going to provide the sub regional estimates that we need. One alternative solution is to create what Mark has alluded to as 'synthetic population microdata'.

[5] So what's the solution? I'm going to argue that the solution is to take some survey data and re-weight it to fit known facts about the local area. For example, in the case illustrated a national survey might provide reliable information on the interaction between smoking and income, whilst a local survey (or census) might provide reliable information on the local univariate distributions of income and smoking. This is because survey estimates of univariate counts can remain relatively reliable for small areas even though the number of survey respondents is too small to provide reliable bivariate estimates. These local 'margins' or 'constraints' can then be used to calculate weights to apply to the national survey. This is directly comparable to the re-weighting of survey data to national constraints with which, as microsimulation practitioners, most of us are well acquainted. For example, many of us will have used re-weighting as a technique for uprating the latest survey population to fit the latest population mid-year estimates. In the present example the principle is the same, but instead of weighting the survey data up to the latest population tables we're weighting them down. As a result the the 10,000 or 100,000 people in your survey are weighted down to represent however many people there are in your local area - perhaps a few hundred or so. And the weighted survey data then become your synthetic microdata.

2. IPF (RAKING)

[6] In this presentation I want to look at three approaches to reweighting survey data to local constraints. The first approach is perhaps most commonly known as raking, which is really Iterative Proportional Fitting (IPF) in disguise.

Understanding IPF

So, how does IPF work? Well, let's assume that you have a tabulation from a national survey that captures the interaction between age and sex. And that you also have some reliable survey-based univariate distributions for the local area. For example, in this illustration, local survey data tells us that there should be 50 males and 50 females, whereas the national survey tabulation has captured, respectively, only 3 and 5. In this case we can simply re-scale by sex to get the right number of males and the right number of females. The scaling factor for males = 50/3 = 16.67; for females it is 10 (50/5). The rescaled data now sum to 50 for both males and females, but the rescaled estimate still doesn't agree with the local age constraints of 20 young and 80 old. A second step, therefore, is to re-scale by age to get the right number of young and old. But this re-scaling misaligns the scaled totals of males and females. The solution is to keep going, iteratively, rescaling first to meet local sex totals, then to meet local age totals, until, eventually, you find a set of scaled counts that sum to simultaneously satisfy the local sex and age constraints. In practice few iterations are normally needed – in this case only 5.

For a convergent solution you can think of the scaled table counts as weights which, when applied to a set of survey microdata, will make your weighted microdata representative of the local distribution of young and old people, males and females. If you undertake this rescaling process via tabular data, as illustrated, before converting to weights, this is known as Iterative Proportional Fitting. (Outside of population studies this same technique goes by a variety of other names.) If you apply the rescaling approach directly to the microdata, so that the record-level survey weights are directly rescaled on

an iterative basis, then this approach is perhaps most commonly known as raking. Both work on exactly the same principle.

What is IPF/Raking doing?

So what is IPF or raking doing? Well, to get technical for a moment, it is preserving the odds ratios.

[7] This illustration has been generously provided rom a draft paper on IPF that a PhD student of mine, Maja Zaloznik, has written. The lower left-hand table is a version of the upper left-hand table that has been rescaled, via IPF, to agree to a set of marginal table constraints (75 male; 25 female). The original male counts are all rescaled by 1.5 (75/50); all female counts by 0.5 (50/25). The central and right hand columns of the illustration demonstrate that, despite this rescaling, the odds ratio – that is, the relative chances of being rich for a male compared to a female – remain unchanged. In other words, the interaction, between the variables in the tabulation remains preserved even though the margins have changed. At first hearing this might sound a little counter-intuitive but re-working through this example at your leisure should persuade you of the truth of this assertion. The chances of being rich for a male haven't changed even though the data have been reweighted to fit some local constraints.

Variation independence

[8] And this raises the important problem of what in the literature is known as variation independence. In another illustration borrowed from Maja's paper, three tabulations of income by sex are presented. All have different margins, but all have exactly the same odds ratio. This perfectly illustrates the property of variation independence. If you know the margins and you know the interaction then you know the table, but these two elements of the table are de-coupled (independent) of each other. So with IPF, although you change the margins to fit the local known conditions, you are not changing the local interaction, which is a bit of a problem (to put it mildly) if this is what you are trying to estimate.

3. COMBINATORIAL OPTIMSATION

[9] Another approach to reweighting survey data to local area constraints is known as combinatorial optimisation.

How does this work? Once again we have a target local area cross-tabulation, in this case age by sex. For this target we have some survey data which captures the interaction of interest, but for a higher level (e.g. national) geography. We also have some local univariate (marginal) constraints, drawn from census or other data for the sub-region of interest. Under Combinatorial Optimsation the reweighting process works as follows. First, if you know there should be 10 people in this local area you initially select at randomly 10 people from the survey and assign them a weight of 1. The remainder of individuals in the survey are assigned an initial weight of zero. Following on from this first step, you aggregate the weighted data survey data to provide an estimate of the target cross-tabulation. In this case the initial set of random weights generate an estimate that of 5 males and 5 females, all young, which obviously fails to fit the known local constraints. The next step, therefore, is to randomly reassign one of the existing weights (i.e. randomly move a weight of 1 from one survey respondent to another), and reaggregate. If the resulting estimate provides a better fit, the random change in weight is kept; if not it is reversed.This process of randomly changing weights continues until, eventually, a set of survey weights are arrived at which, when used to provide a weighted estimate of aggregate age/sex totals, fully satisfies the known local area constraints, as in the illustrated example.

This is a broad brush description of combinatorial optimisation. In reality I'm glossing over some relatively sophisticated algorithms that underpin the so-called random process of weight changes, which include the use of simulated annealing to allow the occasional acceptance of weight changes leading to deteriorations in estimate fit, in order to avoid getting stuck with a sub-optimal set of weights. The main point, however, is that you end up with a set of integer weights which, when used to create weighted estimates, satisfy the known local area constraints.

4. IPF/RAKING V CO

Dealing with inconsistencies arising from disclosure control

[10] Arguably the most fundamental disadvantage of IPF relative to Combinatorial optimisation is that iterative

proportional fitting (raking) requires local area constraints (table margins) to be consistent with one other. If the margins are non-overlapping – for example counts of young or old on one margin and counts of male and female on another, the only requirement is that they both add up the same total. Unfortunately, in their collective wisdom a number of national statistical agencies, including the Office for National Statistics, apply disclosure control to all of their outputs. This means that in the UK, for example, all small area statistics are rounded to the nearest 3 for small counts. These perturbed counts are then used to derive table totals. The means that the local area marginal distribution of sex and might record a total of 11 males and females, whilst the local marginal distribution of age records a total of 9 young and old people. IPF can't cope with inconsistencies in margins, as this means that it won't be able to converge on a stable solution If you want to use IPF/raking, therefore, you have to put a lot of effort into trying to make your margins consistent with each other before you can even start. In comparison combinatorial optimisation just tries to find the set of weights which best fit the margins (i.e. splits the difference). For example, if there should be 11 males and females, 9 young and old, the best fitting estimate will assume a local population total of 10. This means there will be a small error in the fit to the total males and females, and a similar but opposite error in fit to the total number of young and old. In other words, combinatorial optimisation automatically averages out the inconsistencies caused by disclosure control without any need for user intervention.

Relative performance

If we compare these two approaches how well do they work relative to each other? In the following example, I've applied combinatorial optimisation to take some of the Sample of Anonymised Records from the 1991 Census and re-weight them to published enumeration district (street block) margins. This table lists all of the margins that were used as constraints in this reweighting process. Note that these 'margins' don't comprise simple univariate margins. Rather they comprise a whole set of multivariate margins, meaning that the weights fit the known local area interaction between age, sex and marital status distribution, the local household composition by tenure distribution, and so on. In fact the local area constraints being fitted by the reweighting process comprise a total of 814 different census counts spread across 15 different variables, and their interactions across a range of 14 different univariate, bivariate and multivariate tabulations. Despite the number and complexity of these local area constraints, it turns out that combinatorial optimisation can provide sets of survey weights that satisfy them reasonably well, in the sense that there are few observable statistically significant differences between the estimated and known local area counts.

Comparison for *margin-constrained* tables

[11] In comparing the results of the CO-based estimate with IPF, what I've done is to use both methods to estimate the district-level age by sex by tenure by economic position interaction, for a set of 17 districts, and compared the two sets of estimates. I've also compared both sets of estimates to that derived from a large-scale local survey.

The results are as follows:

If we use the 2% Sample of Anomymised Records to estimate the district-level distribution of age by sex by tenure by economic position, then 32% of the estimated cells differ in a statistically significant way (NFC = 'non-fitting cell') from the known distribution, as recorded in published census outputs. This poor performance is clearly attributable to the size of the conference intervals associated with sample surveys that I noted earlier.

For IPF, I experimented with two alternative estimation strategies. The difference revolves around the source used to provide the initial estimate of the local area 'interaction'. The first experiment assumed that there was no interaction (i.e. all variables are independent of each other.) This was implemented by staring off with a uniform distribution of counts across cells (IPF_U). The second experiment involved starting off an initial estimate based upon the interactions captured in the Sample of Anonymised Records at a national level (IPF_N). Unsurprisingly, if the starting point is an assumption of uniform (no) interactions, the results IPF-based estimates (which preserve the odds ratios of this initial estimate) are very poor (37% of estimated cells 'non-fitting'). In comparison, seeding (starting) IPF with a national sample distribution does better (22% of estimated cells 'non-fitting').

The best estimate fit (20% less error than IPF_N) was found to be that produced by Combinatorial Optimisation, although it should be noted that the estimate produced by CO itself was associated with 18% of estimated counts 'not fitting' the true census distribution.

Simpson & Tranmer

[12] Another test comparing CO and IPF, and a slightly more challenging test in some ways builds upon work by Ludi Simpson and Mark Tranmer, who a few years ago tried to push IPF to the limits in terms of making it as sophisticated as they could.

The two central columns in this table summarise the results from a paper they published. Subsequently I have replicated their analysis for 816 wards for which I happen to have synthetic populations generated via combinatorial optimisation. The two right-hand columns of the table show that, by chance, these 816 wards appear to be slightly easier to fit than the full national set of wards. As a result the levels of fit reported for the various IPF-based estimates (1-4a) are better than those reported by Simpson and Tranmer for the full set of 9363 wards. The various approaches to IPF reflect changes in the source of the starting distribution fed into IPF, ranging from an assumption of independence (1) through seeding an IPF with a national distribution (2), seeding an IPF with district level interaction (3), seeding an IPF with district level interaction plus a sophisticated multivariate model to try and adjust the local area interaction (3a), seeding an IPF with a distribution drawn from a 1% sample of respondents living in wards of similar geodemographic type (4) to a seed based on sampling from similar geodemographic areas, modified via multilevel modelling to borrow strength from other sources of information (4a).

However, despite the sophistication of these approaches to IPF, the results reveal that estimates based upon combinatorial optimisation, whether estimates of the counts in the local area distribution, or of some of the conditional probabilities that can be derived from these counts, outperform all of these IPF-based approaches. To be fair, Ludi has argued that this comparison is unfair because, as I showed you, I fitted the weighted survey data to 814 census counts, whereas Ludi IPF'd only the interaction between car ownership and tenure. However, as Ludi has also acknowledged in a subsequent paper, scaling up IPF to cope with a more complex set of marginal distributions is highly problematic due to the impact of disclosure control on census outputs.

5. GREGWT

[13] So far we have compared two methods of reweighting survey data: IPF and CO. A third and final method that I am going to consider today is GREGWT. GREGWT is a Generalised Regression algorithm used by the Australian Bureau of Statistics, based on a similar algorithm called Calamar which is used by, amongst others the Office of National Statistics and INSEE in France to reweight survey data.

Understanding GREGWT

[14] GREGWT, like IPF, is a deterministic algorithm, in the sense that it produces the same result each time it runs (unlike CO, which involves a stochastic element). In theory it is not only deterministic, it also can provide a one-pass solution (i.e. solve a set of simultaneous equations in one pass). In practice, imposing a constraint on GREGWT that no weight can be negative causes it to move into iterative mode, using a Newton-Raphson type algorithm to find the set of weights that satisfy a given set of constraints.

It also turns out that GREGWT cannot always solve the problem of finding a set of weights that simultaneously satisfy a large set of local area constraints. For the technically minded, this 'non-convergence' of the alogirthm is triggered by matrix singularity, making the matrix inversion required for a numerical solution impossible. In addition, GREGWT is sensitive to the consistence of the local area counts used as constraints, in the same way as IPF. Inconsistent constraints will also trigger non-convergence. So if you've got problems of small cell adjustment, which they do in Australian census outputs, you first of all have to go to some lengths to make your local area constraints consistent with each other; and even GREGWT might not converge on a solution. And as we'll see, that can cause a bit of a problem.

6. GREGWT v CO

[15] In order to compare GREGWT to combinatorial optimisation, I'll be reporting on some work I undertook in collaboration with Rob Tanton from NATSEM.

[16] The following table compares the fit of survey data to small area census constraints after reweighting using, respectively, GREGWT and CO. There are, of course, many different ways of measuring estimate performance. The measures in this table record (i) OTAE - overall total absolute error (the sum of the differences between the weighted data and their local area constraints); (ii) OTAE/HH – total absolute error per household being estimated; (iii) OTAPE – overall total absolute proportional error (the sum of differences between the weighted data and their local area constraints when expressed as proportions of table total); and (iv) ORΣZ^2 - the sum of the differences between estimated (reweighted) and constraint (census) cell counts measured using a statistical measure called the Z Score, which captures a mixture of both proportional and absolute error.

For all of these measures combinatorial optimisation does better, particularly it does better for the proportional error but also the absolute error. And this is for areas where GREGWT actually converged; in other words those areas for which it produced a solution. There are some areas where GREGWT didn't converge and, as the table shows, for those areas the GREGWT-based estimates are, unsurprisingly, a complete disaster, whereas CO still produce estimates, albeit significantly worse than the CO estimates for areas for which GREGWT converges - because they're harder to fit areas - but they're still reasonable estimates.

[17] Proving that reweighted survey data fit the known local area constraints is kind of boring because, although this shows that we can reproduce the local area constraints, we already know these. More interesting is to estimate something you didn't know – or at least didn't use as an input to the reweighting process – and that's what we tried to do next. The Australian Bureau of Statistics (ABS) produced a special census tabulation of household income by mortgage and the level of mortgage or rent payments they were making. NATSEM converted these data into census-based estimates of households in housing stress (mortgage/rent too high a proportion of overall income). We then produced equivalent estimates of housing unaffordability using the survey data reweighted to local area constraints produced using first GREGWT and then CO. Now, there are problems with all of these estimates. In theory the ABS estimate is the one we're trying to agree with, but in practice this is impossible because there are some inherent definitional differences between the data used in the ABS estimates and in the ABS local area reweighting constraints.However it is still the target we're trying to hit, and, as the results show, CO did marginally better than GREGWT in estimating the unknown interaction between income and rent payments for a variety of states.

7. VARIATION INDEPENDENCE (AGAIN…)

[18] OK so that's a quick romp through three different ways of trying to re-weight survey data down to some local area constraints to produce spatially detailed synthetic microdata. To try and start drawing all of that together, if I can, we need to briefly revisit the problem of variation independence.

To recap, the margins of the table and the interactions found in the table are independent of each other, at least to an extent. IPF can't solve this problem, as it preserves the known odds ratios. Combinatorial optimisation can perhaps go a little way to addressing it. The reason for this is that the more constraints you use in the process of re-weighting survey data, then the more strength you are potentially 'borrowing' from the interactions between constraints already captured by the survey records. For example, if you are weighting a household survey file rather than a personal level survey file, and you choose a household because the head of household is unemployed then you'll get 'for free' the characteristics of the other people who live in the type of household where the head of the household is unemployed.

I want to just show two examples of some results which show a little bit of that kind of thing.

[19] First of all a relatively simple approach. In this example, instead of attempting to satisfy a complex range of bivariate and multivariate constraints, we're simply trying to reweight survey data to satisfy a set of univariate counts for a local area. **[20]** If you do that then you get a nice graph like this, where the horizontal axis plots the census (constraint) based Townsend Score, whilst the vertical axis plots CO-based Townsend Score. As you can see, the two estimates are practically identical, for all except a very few areas, although this is, perhaps, kind of unsurprising. If you re-weight survey data to constraints based on the right number of unemployed people, and Townsend Score's is based on the number of unemployed number and the number of people with no cars and so on, then you would be surprised if you didn't manage to do that quite well. So that's maybe not a very strong test.

[21] A stronger test is if we go back to a different example, where I re-weighted survey data to local constraints comprising a set of bivariate tables covering 586 local area counts.

[22] This graph tries to give some context against which to look at the next two slides. If you look at the Bs, B1 representing 'middle England' and so on, this is a graph of how far each enumeration district (local area) falls from the national average. So if you've got incredibly high levels of ethnic minority population you'll be up here somewhere (far right hand-side of graph), and similarly for very high levels of unemployment, very high levels of professionals and so on. But, if you've got a near national average level, the enumeration district would be locatrd here (bottom left-hand corner of graph). So middle England, B1 lies right near the average, B2 'rural' is also close to the national average, but B3 'Deprived industrial is starting to slip away from the national average, whilst over here (right-hand side of graph) enumeration district B4 is right out on the tail, representing a really unusual area, comprising a deprived urban area containing lots of council flats. So those are the four areas.

[23] If we take some survey data and re-weight it using combinatorial optimisation, because it's a random process, each time you run it you generate a slightly different set of result. So if we run CO 100-times and produce 100 sets of weights, we can use those weights to estimate local area interactions which haven't been used as reweighting constraints, and compare them to the equivalent census-based tabulations. In each case presented in this table the margins of these interactions have been constrained. For example, the reweighted data have been constrained on socio economic group and household composition, but not on the interaction between socio-economic group and household composition. In other words, we've constrained to fit the local distributions of socio-economic group and household composition, but we haven't constrained to fit the local interaction between socio-economic group and household composition. In other words, we've reweighted to the margins, but haven't constrained the reweighting on the interaction. Even so, as the results show, regardless of area type, rural areas, middle England, deprived industrial areas and deprived urban areas, for all of areas types the margin-constrained estimates of local interactions all 'fit' their census-based equivalents. The only exception is the distribution of sex by marital status by tenure where, for rural areas, up to 16% of the estimates don't fit. Otherwise pretty much everything fits.

So maybe this provides some evidence that combinatorial optimisation is starting to overcome this problem of variation independence; in other words that we are capturing something about the locally varying nature of interactions, which would help to explain why there's lots of zeros in this results table

[24] However, and here's the big note of caution, although the survey data includes attributes recording both migration behaviour and age, for example, we have constrained on age, but not on migration. The interactions between these two attributes, therefore, might reasonably be described as 'unconstrained'. As this second results table shows, where there are unconstrained margins in the tables being estimated, as well as unconstrained interactions, then suddenly the numbers blow up and nothing fits. So the message from that is if you don't constrain on the margins, when you produce the tabulations involving these unconstrained margins your estimates are likely to be highly unreliable. This is perhaps surprising because we often think of taking survey data, re-weighting it by age and sex and then doing some analysis of, for example, age by sex by illness. But as these results confirm, and the property of variation independence suggests, if you haven't actually constrained on illness as one of the sets of constraints then your estimates are likely to be off, certainly at the small area level which has been the focus of this estimation exercise.

8. CONCLUSION

(a) Accuracy of estimates

[25] Moving on to a conclusion, how accurately can we produce synthetic microdata for small areas? Well, harking back to the beginning of this talk, better than a blank piece of paper; better also than IPF, which is a technique which is widely used to produce small area estimates for one particular tabulation of interest (rather than a whole set of microdata); and maybe that's encouraging enough. A question to all of us really and a question that I would like to know the answer to is what do we need to do - what do I need to do - to convince the wider scientific and policy community of how good this is as an approach? What other evidence needs to be produced?

We can also note, again, that if you've constrained on the margins the estimates are likely to be plausible; but not if you haven't constrained on the margins. So if smoking was one of the variables in the data set and I didn't constrain on that,

and I then produced a table of age by sex by smoking, the chances are it will just look like the national average really, and won't reflect anything like the local interactions.

(b) Unanswered questions

I've shown you all sorts of examples of different numbers of constraints, 816 counts, 500 counts and so on. You can obviously use fewer constraints than this, or more; you can also use univariate constraints, bivariate constraints, multivariate constraints, but what's optimal? I have yet to prove this, but I intuit that there must be a kind of inverse U curve: very few constraints and you'll fit the constraints but you won't have actually pinned anything else down, so the resulting estimate based on reweighted survey data won't be very reflective of local conditions; too many constraints and you'll never fit them as to so will prove to be too challenging a problem; and somewhere in the middle must be a optimal number of constraints. But how on earth do we work that out?

A second unanswered question is whether we should weight household surveys or person surveys? Does it make a difference? What value-added benefits you get from weighting households rather than persons? For example, if you select a household comprising a set of unemployed people, a range of other associated attributes, such as housing tenure, car ownership etc. are likely to come in tow. If so, how much extra strength does that give to our synthetic data compared to just weighting a person level file?

(c) Applications in the real world

And then there applications in the real world. As Mark's has already shown us examples of these in his talk, I'll stop here. There are clearly many applications for synthetic data, for small areas, once you've produced it and persuaded other that it's plausible.

QUESTIONS

Howard Redway: Are there alternative algorithms for implementing Combinatorial Optimsation?

Speaker: Oh there are many algorithms yes, yes there are.

Howard Redway: Can you recommend one? Particularly if you wanted to play around with something like this and get your hands dirty.

Speaker: Mark and I wrote a paper a number of years ago exploring alternative algorithms, and ended up opting for one based upon simulated annealing because it stops the algorithm getting stuck in a local sub-optima.

Howard Redway: Is there an 'off-the-shelf' software package that can implement this algorithm?

Speaker: Sort of. The combinatorial optimisation algorithm I've written up as a piece of software which you can download and run. I know at least one person who has downloaded it and used it, so it's do-able without any help from me, but it's probably not just an hour's task, which is why I'm hesitating. Mark may be aware of some more readily applicable off the shelf implementations.

Mark Birkin: There is a guy called Kirk Harland who is putting something together at the moment into a similar format, but it would require intervention from him in order for you to be able to do something. But Paul is being too modest in recommending his own software package.

Speaker: OK, well in that case, yes, you can download mine and the documentation that goes with it and away you go.

Holly Sutherland: I thought that was really interesting. I have no experience of estimating small areas but using these methods to change of information and the two points that you made or questions that you raised at the end. One is the optimal constraints and I'm not sure that's the right question to ask because I think it depends on what you're trying to

do. I think once, you know if you can kind of characterise what you're trying to do then theyre maybe an optimum but otherwise I think it's kind of up to you isn't it?

Speaker: Mm, yeah.

Olly: I mean I used CALMAR and going back to the question that was asked first, I used it because it's there and I understand French - it's coded in French. And all I can do then is to experiment and to look at the effects on the target variable, you know things I'm actually interested in, the reason why I'm doing this in the first place. But there's a, it's a matter of judgement, it's not science.

Speaker: Indeed yes.

Holly Sutherland: The other point you made about whether to do things at a household level or a personal level, this may not apply if you're using the census. I think there's a lot to be gained by using both at once and that's like adding a lot of sort of implicit constraints, so you then end up adding fewer than you would otherwise if you were at one level or the other. But you're sort of implicitly picking up, as you said, a lot of special information about the types of households that unemployed people are in, but also the types of people that are in rented houses, so it works both ways. So again, simply in a crafts-person viewpoint, rather than applying science, my experience is that using both is good. In fact for personal tax benefit modelling it's also good to be able to operate at the benefit unit level. But it would be really nice if somebody somewhere is doing something more scientifically rigorous rather than intuitive on this.

Mark Birkin: You were talking about combinatorial optimisation being an improvement on some, some variation on geodemographics, IPF or geodemographics or something. I just wondered if you'd ever considered combining kind combinatorial optimisation and geodemographics in some way, like when you look at car ownership, you know presumably all those variations are related to comparing inner city areas with suburban areas and such like. Have you done it and it didn't work?

Speaker: No. On my next to do list is a whole range of algorithmic tweaks and that's one of them. If you've got the whole survey, do you re-weight the whole survey or do you only take the regional sub sample and re-weight that or the district sub sample, or households from the same type of area. Such evidence as I've got so far tends to suggest that the more households you've got to pick from then the better your end result would be, but I haven't actually tried the geodemographic banding. For regional banding, households from, say North Yorkshire are so different from households in inner city Leeds, that just because they're in Yorkshire and Humberside we have found that not to be a very good strategy we found; but this is something to be looked at more.

Belinda Wu: If you start with a small sample to reweight, is there a danger of ending up with too many identical households, leading to the same survey household having to represent 10 or 20 households?

Speaker: Yes, that's another issue, I kind of danced over that, but all the work I have done has really been re-weighting down to output area or enumeration district level, so street blocks and then aggregating back up to wards or districts or whatever you're actually interested in, on the grounds that the small areas are very distinctive, you get concentrations of people of different sorts, so if you fit that then hopefully you're going to tease out more of the spatial heterogenerity and then add it back together, whereas if you just fit to district totals you kind of get a bland district average and that seems to have worked so far. But I mean in doing that if you're fitting a very unusual area like a student area, there are only, I don't know, 100 student households in the sample of anonymised records and you need 200 students, with some of these households you're going to get some fairly high weightings because they are the only ones available to pick, but then it's probably still better to have a bunch of student households and a set of households which have got lower weights but are more like a nationally representative set of households.

Questioner: I was just wondering about combinatorial optimisation is this optimising for entropy?

Speaker: Yeah well IPF produces a maximum likelihood estimate or maximum entropy, which are the same thing in this context. Combinatorial optimisation wouldn't claim to do anything similar, and in fact the word optimisation could be objected to, as the result is just the set of weights found to best fit the constraints in the time available, which isn't quite the same as optimal.

Questioner: We were trying to estimate origin destination migration tables and we had the margins from NHS data and from the Census we had interaction data and generated, it was almost like a [?] from a statistical stance trying to constrain it that way seems like… or I was thinking maybe optimisation of entropy was a good statistical analogy?

END OF RECORDING