

Dynamic Social Simulation Models Enabled by e-Research

Mark Birkin, Belinda Wu

School of Geography, University of Leeds

Email address of corresponding author: m.h.birkin@leeds.ac.uk

Abstract. A synthetic representation of the entire population of the city of Leeds has been generated from publicly available datasets. Using an event driven model which simulates discrete demographic processes, the population has been projected 25 years into the future. Whilst the approach is grounded in the methods of microsimulation, concepts from spatial interaction modelling and agent-based systems are incorporated in an innovative way. Following the introduction of appropriate simplifying assumptions, the model still incorporates a great many parameters and assumptions. A brief sketch of outputs and potential applications of the model is provided.

1 Introduction

Desktop simulation games of urban areas have been phenomenally popular in recent years. The underlying aim of the research reported in this paper is to translate such games into real world policy environments. If planners were equipped with the means (through simulation) to understand social and demographic changes in response to shifts in policy, such a device would have valuable practical applications as both a ‘decision support system’, and as a pedagogic tool for understanding how cities work. As an academic and intellectual challenge, the ability to reproduce and predict the behaviour of real city systems constitutes the ultimate demonstration of a deep understanding of such systems.

The paper describes a research project – Moses – which aims to produce a simulation model of the UK population, as it now is and as it can be expected to develop over a twenty-five year time horizon. As a fundamental basis for the approach, the technique of microsimulation has been adopted. Microsimulation has a fifty year history as a technique within economic analysis (Orcutt, 1957), while more recent applications of spatial microsimulation have embraced problems such as transportation, healthcare and housing. The benefits of microsimulation (in contrast to macroscopic modelling approaches with similar objectives) within a demographic modelling context have been argued persuasively and eloquently by van Imhoff and Post (1998). In particular, these authors demonstrate the richness of microsimulation as a device for the representation of both relationships between members of a population, and of the transitions between states within a population. In this paper, a microsimulation model of the population and its dynamics will be described. The application of the model to the urban area of Leeds, a city of 730,000 people in the north of England, will be introduced. The Leeds area is used for illustrative purposes throughout this paper, but is completely generalisable between local areas across the United Kingdom.

In the remainder of the paper, the next section describes the establishment of a baseline model, and Section 3 moves on to the main features of the dynamic simulation. Section 4 of the paper includes a review of some results from the current implementation. Brief concluding remarks are offered at the end of the paper.

2 Population initialisation

The base year simulation was established by reweighting the Household Sample of Anonymised Records (HSAR) to individual census wards in Leeds Metropolitan District. HSAR comprises a 1% sample of households from the UK Census of 2001 in which the census questionnaires are completely enumerated for households and their constituent individuals. HSAR comprises a portfolio of more than 50 characteristics for every individual in a household, ranging from social and demographic variables to property type, employment and education. For a full list, refer to CCSR (2005). The essential mechanism for protection of the anonymity of individuals is through reduction in the geographical resolution of data. Thus each household can only be identified to the level of a Standard Region (i.e. South-West, Yorkshire and Humberside etc), which limits the value of the source data for the purpose of spatial microsimulation.

The approach adopted here is to synthesise records from the HSAR in accordance with the structure of individual wards. Each ward population is therefore a unique extract or ‘re-sampling’ from the HSAR in accordance with the local geography. For example, in areas like Headingley there is a high probability that students will be selected from the HSAR, while in Wetherby the selected individuals are much more likely to be prosperous and more elderly. Households are sampled ‘without replacement’ from the parent distribution and are therefore unique within an area, although it is not only possible but necessary that some households will be duplicated in the Leeds area¹.

The reweighting procedure is based on successive proportional sampling from the HSAR. Initially, the probability of selecting a candidate from the HSAR is random. From this random selection, the composition of the sample is compared to key population distributions from the 2001 Census Small Area Statistics (SAS). The weights are adjusted to increase the likelihood of selection for population sub-groups which are under-represented, and vice versa. This adjustment process continues until the observed and predicted population distributions are similar within a specified tolerance. For the work reported in this paper, the population was reweighted based on two household characteristics – the age and social class of the head of household – and two individual characteristics – ethnicity and health status. A detailed description of the process is provided by Birkin et al (2006).

The simulation model is anchored in a base year population for the year 2001. This is necessary in order to exploit the richness of the UK census of population and households for the purpose of the simulation. There are no appropriate data sources to facilitate the preparation of a more up-to-date base population.

The method of ‘iterative proportional sampling’ is similar to the procedure by which realistic populations have been generated for social simulation experiments using the American city of Portland (Barrett et al, 2005; Beckmann et al, 1996). An alternative approach to the problem of reweighting based on simulated annealing has been deployed with good results elsewhere

¹ Leeds accounting for more than 1% of the UK population, there are more households in the city than in the HSAR.

(Ballas et al, 2004; Kavroudakis et al, 2007), while the present authors have also experimented with genetic algorithms for the same purpose. In early research, the regeneration of synthetic individuals and households from aggregate distributions was advocated (Birkin and Clarke, 1988).

An advantage of the synthetic estimation approach is that any chance of breaches of individual confidentiality is automatically precluded. Basic arithmetic suggests that as many as 10 synthetic households in Leeds could be original HSAR respondents located in their actual ward of residence, but given that these are in effect hidden at random within the entire population the chances of them being identified are remote, even if complete household profiles were to be disclosed. Such disclosure is not considered to be a useful or necessary step within the research without the benefit of secondary aggregation of the data. Confidentiality of the original data is therefore not viewed as a primary concern within this project.

In contrast to the Household SAR, the British Household Panel Survey (BHPS) has been proposed as an alternative source of household data for the synthetic reconstruction of small area populations in the SimBritain project (Ballas et al, 2004). BHPS has the advantage of a much more substantial list of attributes, including lifestyles, attitudes, and detailed household income and expenditure profiles. BHPS is also longitudinal, providing a view of household change through time, from 1991 up until the present day. However BHPS is based on a limited sample of 16,000 households and is unlikely to be as representative of the full range of household types as the HSAR. In this research, the HSAR is therefore preferred as the raw material for the population reconstruction model, while BHPS is used extensively in the extraction of relationships for the dynamic model (see below). The possibility of data linkage models between the BHPS and HSAR has also been explored (Zuo, 2007, see Section 4 below).

The selection of the appropriate constraining variables is also problematic to the extent that there are no guarantees as to the accuracy of other distributions. The prediction of secondary distributions such as tenure and car ownership is not unreasonable but lacks the accuracy of those estimates which are explicitly included within the modelling process (i.e. age, social class, health, ethnicity). This mirrors the implied view of authors who have favoured the use of domain-specific constraints in the context of applications such as health (Smith et al, 2006) and education (Kavroudakis et al, 2007). It remains an open question whether there is a set of key attributes which would allow the production of a general purpose base simulation, or alternatively which are the key variables for particular domain applications.

3 Dynamic modelling

The objective of the dynamic modelling is simply to project the population forwards in time. For such purposes, the most commonly used approaches have been cohort-based 'macrosimulation', in which the population is divided into categories, and multipliers – such as survival probabilities or birth rates - are applied to those individual categories. However these methods are problematic whenever a relatively rich set of population attributes is involved, as the number of categories begins to grow exponentially. Van Imhoff and Post (1998) present an example in which the population of France is represented more efficiently through an individual microsimulation model even though age/ sex, marital status, parity and place of residence are the only variables.

A number of projects have therefore attempted to build dynamic microsimulation models, particularly for economic applications (e.g. Rephann and Holm, 2004) but also for social and anthropological applications to problems of kinship and community (Murphy, 2004; van Imhoff and Post, 1998). However the only examples of demographic projection with spatial microsimulation use the technique of 'static ageing', in which a base population (the HSAR or BHPS in our previous example) is resampled in the context of independent estimates of future population change (e.g. Ballas et al, 2004). Therefore this method provides no means to monitor the dynamics of change within a population, and ignores the benefits of dynamic microsimulation as a means for the projection itself.

Some authors have seen fit to make a distinction between policy and pedagogic applications of microsimulation (van Imhoff and Post, 1998). In this context, it is argued that the accuracy of models with a policy orientation need to be validated in a real world context, whereas pedagogic models need only to reproduce local interactions within the population. The present authors remain sceptical of such a distinction, and view the process of validation as essential if robust conclusions are to be drawn which is independent of the model as artifice. In the following sections we therefore describe the components of the Moses dynamic microsimulation model and the means for estimation of its constituent parts and processes.

3.1 Mortality

Survival rates in the model are derived separately for each of 33 census wards, both genders, and 101 individual age groups, giving a total of 6,666 individual parameters. Age and gender specific rates are derived from national data (for the year 2001) in which the number of deaths for each group can be normalised by a population at risk. These national rates are then applied to the population-at-risk within each census ward in the Leeds area to yield the expected number of deaths. This prediction is compared to known death totals for each ward from ONS Vital Statistics (again for base year 2001) and an appropriate multiplier is applied within each ward in order to balance observed and predicted deaths. At each time period, a survival probability is applied to each individual on the basis of age, gender and location. The model is run in annual time increments, and therefore the ageing rule for all survivors is that they become a single year older in each time step.

It can be seen that for this part of the model alone, projection of mortality over a 30 year time interval involves the estimation of roughly 200,000 individual mortality rates. In order to make this problem manageable, the two options which have so far been considered have been the adoption of constant mortality rates across the projection period; and a fixed percentage reduction in mortality rates across all age-gender and location combinations at each time step in the simulation. These options are discussed further at Section 4 below.

3.2 Fertility

Ward-specific fertility rates are derived in a rather similar way to the mortality rates as described at section 3.1 above. In this case the population at risk, whilst exclusively female, is segmented into married or unmarried, and the rates are partitioned into five year cohorts between the ages of 15 and 49 giving seven groups in total. National rates are again localised in accordance with ONS Vital Statistics. Thus a total of 462 different fertility rates is estimated for Leeds in the base year. Although this is less extensive than for mortality, nevertheless projection across a 30 year time frame would still require considerable feats of calculation. For this reason, constant fertility rates are again adopted as a reference point for

the simulation. Since fertility rates are known to be variable, possibly even cyclical over time, other structured variations in rates are also explored later.

3.3 Health status

Individual health states are recorded in the HSAR and SAS within five categories ranging from very poor to very good, but for both convenience and robust estimation these are reduced to three categories of 'poor', 'medium' and 'good' within the simulation. For each individual, the probability of a change in health status is assumed to be dependent on current health status, age and gender. These rates of change are derived from the BHPS, and are applied evenly in all geographical areas, as we assume that spatial variations are captured by the initial healthy states in the population². The base simulations also assume no change in these rates over time.

3.4 Household formation

Changes in household composition are determined by four processes in the model. These are the formation of new unions (including marriage), the dissolution of existing unions, movements in which one or more persons leaves a household, and the death of a household member. Again, each of these processes is calibrated using data from the BHPS. In order to increase the validity of the sample, data is combined from five successive waves between 2000 and 2004.

The household formation mechanism includes a number of important simplifications which enhance the computability of the model. Regarding household dissolution, the data shows us that the vast majority of changes involve the loss of a single household member, and with this justification we choose to ignore any splits which involve two or more members in both the stayer and the leaver group. Leavers, whether single or multiple, are subject to relocation in accordance with the migration model described at section 3.5 below. Dissolution of a union is always associated with the dissolution of a household, although the reverse is not true. In other words, one partner always leaves home when a relationship breaks down, but there are other reasons for leaving home (e.g. start a new job or go to university). The formation of new unions involves the connection between two unmarried people in a given area. It is assumed that all unions involve one man and one woman, although dependent children from previous relationships may also be involved. Although we recognise that household transitions due to a change in health status, specifically the movement of elderly and infirm people into residential care, is an extremely important and interesting transition, at this stage this process is ignored by the model.

3.5 Migration

It is the difficulty of modelling migration flows between small areas which present the greatest problems in the dynamic spatial microsimulation. The main problem is that it is difficult to represent the movement of individuals and households at the purely organic level, and without reference to market level processes which govern factors such as the availability, price, and attractiveness of households within different neighbourhoods. In order to accommodate migration, the model has two important features:

² Of course this assumption is essentially illogical since it implies that over time spatial variations in health patterns will be gradually eliminated. Yet this also illustrates some of the difficulties in the model estimation, since to allow spatial variations for 101 age groups would now create 9,999 parameter estimates for the city of Leeds. Robust estimates could not be supported from 16,000 BHPS records with little spatial definition. Richer hypotheses and/or richer data sources are required to support further disaggregation in this module.

- i. There exists a stock of houses which are independent from the households which occupy them. This includes key features such as housing type (e.g. detached, terraced etc), amenities (e.g. central heating), and size (e.g. number of rooms). At any time, some housing in any area will be vacant, and new vacancies will be created by households which are themselves movers, and can be added to by new house building. Housing vacancies are reduced by the occupation of vacant housing and by the demolition of existing properties.
- ii. There is a location search process which is mediated through an aggregate spatial interaction model (SIM) of migration. The model recognises that different households – according to their size, composition and age – have different housing preferences and search horizons. These processes are calibrated against data drawn from the 2001 Census Special Migration Statistics (SMS). Migration is represented as a process in three stages. In the first stage, migrants are selected using movement probabilities drawn from the BHPS and reweighted to observed movement levels in small geographical areas. Then preferences by housing type and location are derived from the meso-level SIM, and these preferences or probabilities are sampled in the usual way. Finally, the mover household is matched with an available house of the required type at the destination location.

In addition to migration between areas within Leeds, external migration flows in and out of the city need to be considered. National migration patterns can again be detected from the SMS; international migration patterns are assessed from the International Passenger Survey. Outmigrants are selected at random in accordance with household characteristics (age, ethnicity, composition) while immigrants are selected from members of the HSAR who are themselves recent migrants.

3.6 Module sequence

A final question concerns the order in which the various demographic modules are implemented. Van Imhoff and Post (1998, 116-117) note four alternative procedures involving simultaneous implementation, versus sequential implementation with a fixed order of events, a random order of events, or in which the most likely event is itself simulated for each individual. In this application, we have adopted a fixed order of events in the sequence – fertility, health change, mortality, migration, household formation. Fertility and health change both appear before mortality, since there are still risks of infant mortality, and although death rates in our model are independent of health status there is a logical connection between mortality and deteriorating health. Since many new households are formed in association with the migration process, it makes sense to consolidate household structures once a move has taken place, while recognising that a more desirable option would be to consider these two processes simultaneously.

3.7 Student migration

Students living away from the parental home provide special problems in demographic simulation. In wards where student migration has a great impact, any cohort-based model will struggle to reproduce the student population renewal and they grow old in the areas as other people do. It is known however that students tend to only stay in such areas during the period of their study and then leave while other new students move in. Due to the replenishment of student population each year, the population in such wards stays younger than that in other

wards. A hybrid approach combining ABM techniques is therefore adopted to strengthen the modeling of such subtlety of the local migration patterns and the behavior modelling of student migrants. A brief description is provided here: for more details, see Wu et al (2007).

To model the student migration, we recognise the following groups of the students: First year undergraduates; Other undergraduates; Master students; and Doctoral students. From analysis of student records, we found in reality that the first year students tend to stay in university accommodations that are really close to the university where they study. From the second year on, as they get familiar with the area, they move out and find private rented accommodation, often with their fellow students in the same area. Based on such assumptions, we then apply following general rules to the “student agents”:

- Each group is allowed set years to stay in an area
- Students stay close to their university of study, subject to housing availability and
- They don't marry or have children

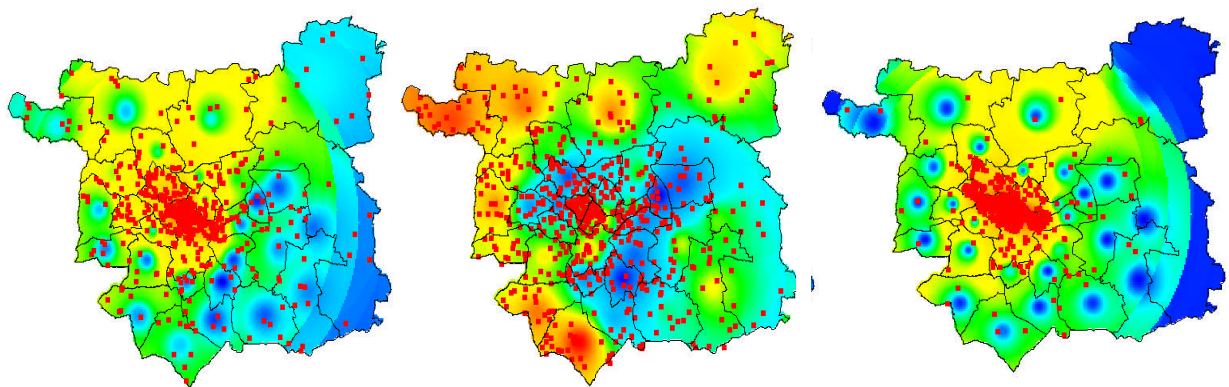
By applying different rules to individual “student agents”, the hybrid model presents a better reflection of the observed student population in wards of Leeds. Below are the comparison between the currently observed student population and the simulation results 3 years later using a pure MSM approach and a hybrid MSM and ABM approach. In the maps, each red dot represents 100 students. The decline of the density of the student population in area is indicated by the color of red, yellow, green, pale blue and blue in the map. From the maps we can observe that instead of pure MSM projecting students scattering around the whole city and even resulting in some over-representation of student population in suburban areas, the hybrid model projection indicates that students tend to gather around the universities in the city center and they normally do not live in suburban areas (Figure 1).

Figure 1 Leeds students: simulation results

a) Observed

b) MSM

c) ABM



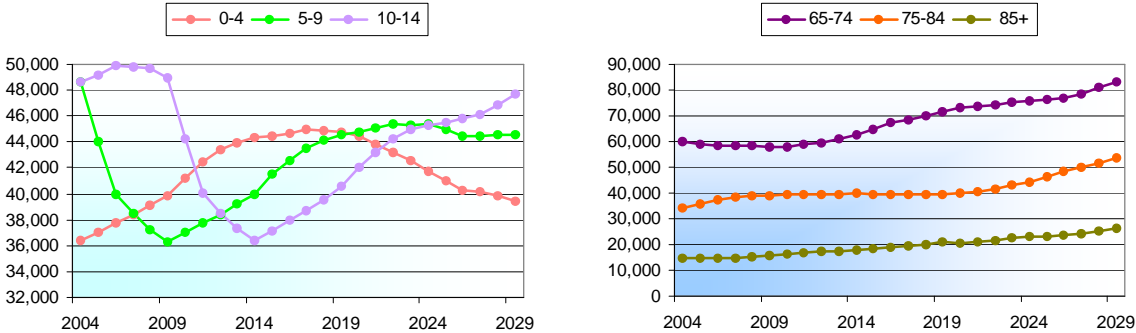
4 Indicative results

Some key results from the current implementation of the Moses dynamic spatial microsimulation model are described in this section. Results are presented over a 25 year

projection horizon from 2006 to 2031. In relation to our previous discussion about assumptions, survival rates are expected to improve by 1% in each age-gender-location combination in each year of the simulation. Fertility rates are expected to increase by five percentage points for each demographic sub-group in each year to 2011 before stabilizing.

Firstly, we note some basic demographic trends, shown in Figure 2. Regarding the age composition of the population, we find a surge of 20% or more in each of the elderly cohorts 65-74, 75-84 and 85+. The school age cohorts all end up reasonably close to where they started, albeit by rather different trajectories. Current fertility levels in Leeds as elsewhere are close to their historical low, but beginning to rebound. An assumption of increased fertility levels is offset by lower numbers moving into the major child-bearing age groups. These results are very much in line with the Office for National Statistics expectations for the Leeds area (ONS, 2007).

Figure 2. Projections by age cohort: a) School age; b) Elderly



Ethnic group projections show substantial growth in all minority groups. These are the product of two effects – ongoing net migration of minorities into Leeds and a demographic bulge in the younger, more fecund age groups. Note that fertility rates are uniform between ethnic groups with the same age and marital status profiles. These projections are again in line with estimates produced on behalf of the local development agency (Rees et al, 2007).

Table 1. Ethnic Minority Projections

| | 2006 | 2011 | 2016 | 2021 | 2026 | 2031 |
|---|--------|--------|--------|--------|--------|--------|
| UK | 671334 | 675994 | 687030 | 697905 | 702397 | 702929 |
| New Commonwealth - Africa and Caribbean | 13464 | 14321 | 15815 | 17786 | 19489 | 20802 |
| New Commonwealth - Asia | 32515 | 34987 | 39457 | 44709 | 49551 | 54118 |
| Others | 21245 | 24616 | 29162 | 34236 | 39900 | 46556 |

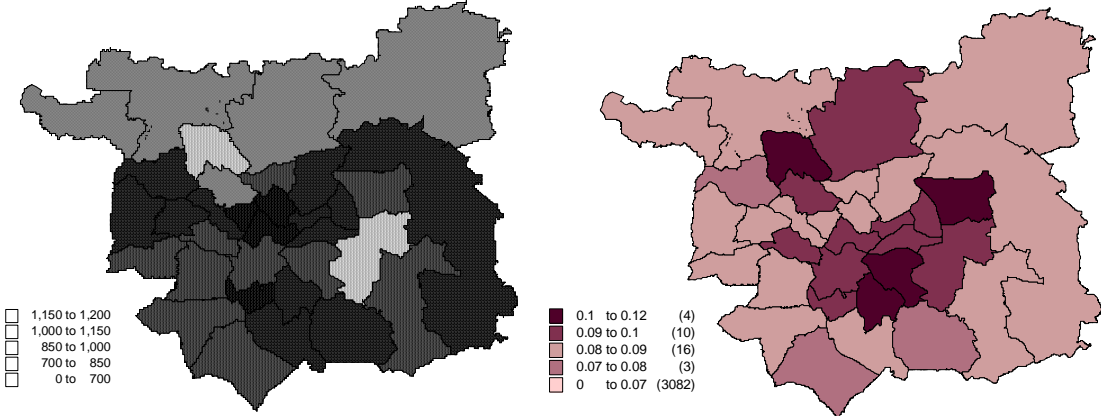
Spatial disaggregation in the growth of elderly populations is illustrated in Figure 3. Here the notable trends are an accumulation in areas where the older age groups are most heavily concentrated already, notably the Wharfe valley across the northern edge of the city, Cookridge and Halton wards. However there is significant growth in all areas through ageing in situ: the migration process does not simply relocate all the pensionable age groups into a small number of retirement areas.

More detailed effects can now be explored through secondary modeling. In the next illustration, we have used a data linkage model (Zuo, 2007) in order to match records from the BHPS with individual HSAR records, and from this we have extracted disability rates. The total individuals with a disability is then summed back to census wards and plotted as

Figure 4. Here we can see some relationship between disability and old age in the wards of North and Cookridge, but the dominant pattern is one of centralization. The poorer socio-economic groups, often associated with low economic activity rates and high unemployment, in the central areas clearly experience much higher like-for-like disability rates than their more affluent suburban counterparts.

Figure 3. Growth in population 85+ to 2031

Figure 4. Disability in Leeds, 2006



Rates of disability are projected into the future in two ways. First, the projected individuals from the dynamic microsimulation are again matched with individuals from the BHPS 2004 to provide an estimate of disability with current health patterns. In this scenario, the proportion of disabled people in Leeds rises from 9.1% to 14.1% in a 25 year period. In the second estimate, we assume that for populations aged 40 and over, individual health improves steadily so that in 25 years time everyone is ‘five health years’ younger than at present – for example, a typical person aged 65 in 2031 has the average health characteristics of a 60 year old in 2006. Given current concerns over conditions such as obesity and diabetes, such an assumption may well be optimistic, but even in this scenario the disabled population of Leeds grows from 51,600 to 70,400 – a 36% increase. On this metric it is clear that social services in Leeds will be under acute pressure over the next 25 years, and that the spatial consequences of increasing need will be uneven.

5 Conclusion

A synthetic representation of the entire Leeds population has been generated from publicly available datasets. Using an events driven model which simulates discrete demographic processes, the population has been projected 25 years into the future. Whilst the approach is grounded in the methods of microsimulation, concepts from spatial interaction modeling and agent-based systems are incorporated in an innovative way.

Although appropriate simplifying assumptions have been introduced, the model still incorporates a great many parameters and assumptions. A brief sketch of outputs and potential applications of the model has been provided.

References

- Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., and Rossiter, D. (2004) 'SimBritain: A Spatial Microsimulation Approach to Population Dynamics', *Population, Space and Place*, 11, 13-34.
- Barrett S, Eubank S, Smith J (2005) 'If smallpox strikes Portland', *Scientific American*, 292(3), 54-62.
- Beckman, R., Baggerly, K. and McKay, M. (1996). 'Creating Synthetic Baseline Populations'. *Transportation Research-A*, 30(6), 415-429.
- Birkin M and Clarke M. 1988. 'SYNTHESIS – a synthetic spatial information system for urban and regional analysis'. *Environment and Planning A* 20: 1645–1671.
- Birkin, M., Turner, A., and Wu, B. (2006) 'A Synthetic Demographic Model of the UK Population: Methods, Progress and Problems', *Proceedings of the Second International Conference on e-Social Science*, NCESS, Manchester.
- Centre for Census and Survey Research (2005) *2001 Household SAR (Licensed) Codebook*, CCSR, University of Manchester.
- van Imhoff, E. & Post, W. 1998. 'Microsimulation methods for population projection.', *Population: An English Selection*, 10: 97-138.
- Kavroudakis, D., Ballas, D. and Birkin, M. (2007) 'A spatial microsimulation approach to the analysis of social and spatial inequalities in educational attainment', 54th Annual North American Meetings of the Regional Science Association, Savannah, GA, USA
- Murphy, M (2004): 'Tracing very long-term kinship networks using SOCSIM', *Demographic Research*, 10, 171-196.
- Office for National Statistics (2007) *Sub-national population projections, 2004-2029*, HMSO, London.
- Orcutt, G., (1957) 'A new type of socio-economic system', *Review of Economics & Statistics*, 58, pp. 773-797.
- Rees, P., Stillwell, J. and Boden, P. (2007) *Ethnic projections for West Yorkshire*, School of Geography, University of Leeds for Yorkshire Forward.
- Rephann, T. J. and Holm, E. (2004): 'Economic-Demographic Effects of Immigration', *International Regional Science Review*, 27, 379-410.
- Smith, D.M., Clarke, G.P., Ransley, J. & Cade, J. (2006) 'Food access & health: a microsimulation framework for analysis', *Studies in Regional Science*, 35(4), 909-927
- Wu, B., Birkin, M., and Rees, P. (2007) 'A spatial microsimulation model with an ABM insight', *GeoComputation*, NUI Maynooth.
- Zuo, C. (2007) *A model of house prices in the Leeds area*, unpublished MSc thesis, School of Geography, University of Leeds.