Using census data in microsimulation modelling
Mark Birkin
23rd August 2017
i. Preamble

• What is **microsimulation**?
• **Examples** from health, transport and consumption
• Lessons, limitations and **future directions**
1. What is microsimulation?

- **Micro** refers to the small units of analysis (usually individual level) and **simulation** refers to the process of modelling these individual units
  - **Traffic microsimulation** usually refers to modelling the movement of individual vehicles on the transport network
  - **Economic microsimulation** is often used in the context of estimating the distributional impacts of new policies (e.g. EUROMOD, Sutherland and Figari 2013)
  - **Spatial microsimulation** is “The creation, analysis and modelling of individual level data allocated to geographic zones” (Lovelace and Dumont 2016, 3.1).
1. What is microsimulation?
2. Synthetic microdata

- combine the advantages of individual level data with the geographical specificity of spatal data ... to provide 'the best of both worlds' of available data by combining them into a single analysis
- Office for National Statistics ... 1% sample
- ‘facilitating another way of examining Census data, for example through the building of statistical models’ (ONS, 2014)
- sample population contains 17 variables relating to 569,741 individuals with a regional (large geographical unit) identifier ... sampling population of 560,996
- Age, marital status, hours worked, occupation, economic activity and sex – validated against general health
2. Synthetic microdata

- **combinatorial optimisation** to ‘clone’ the individuals in the sampling population
- **Flexible Modelling Framework** (FMF) application, an open source software package developed at the University of Leeds (Harland, 2013).
- Total Absolute Error (TAE) ... The FMF re-iterated the optimisation process until **TAE was equal to zero**
- Generating this synthetic population of 56,075,912 was **computationally intensive**, but was possible using FMF on a desktop PC with 32GB of RAM and took 13 minutes and 27 seconds of processing time.
- This synthetic population is **available for research purposes** via the Consumer Data Research Centre https://data.cdrc.ac.uk/dataset/synthetic-population.
3. UK diet and lifestyle data

• In the UK comprehensive dietary data is collected in the National Diet and Nutrition Survey (NDNS)
• nationally representative, sampling from 799 post code sectors
• 3,450 adults and 3,378 children are now included in the sample.
• stratified (random) sampling ... unable to support local analysis due to a combination of lack of data and privacy restrictions on small respondent numbers
• value exists to understand small area variations in diet.... Using a spatial microsimulation of the England and Wales population – described above – it is possible to estimate small area diet and lifestyle patterns in the UK
3. UK diet and lifestyle data

- Dietary data from the NDNS has been linked to the spatial MSM population using key demographic characteristics which appear in both the NDNS survey and the spatial MSM population.
- The MSOA geographies were aggregated to Region for the purpose of linking via the Region in the NDNS data.
- NDNS = 6,800; MSM = 50,000 approx
- The data matching algorithm for this case study uses a modified Gower distance to exploit the similarity between the survey individuals and the synthetic population.
  - the Gower distance is zero between and pair with a shared state, and 1/n for any pair with distinct states

\[ x_{ij} = \sum_k \frac{1}{n_k} \delta_{ij} \]
3. UK diet and lifestyle data
4. UK holiday market

- Small sample of 16K respondents from a travel questionnaire in the Leeds area
- Gender, age, occupation, income, home postcode, travel companions, travel destination, satisfaction
- Link to spatial MSM using common attributes
- In this case profile households using decision tree to solve ‘many-to-one’ problem
4. UK holiday market
4. UK holiday market

Figure 2: Average Distance Travelled by Segment
4. UK holiday market

- Source: CDRC. Destinations are grouped into 12 regions [ 1 = Cyprus; 2 = Eastern Europe; 3 = Egypt; 4 = Greece; 5 = Indian Ocean; 6 = Italy; 7 = Mexico; 8 = North Africa; 9 = Spain; 10 = Turkey; 11 = USA; 12 = West Indies ]
4. UK holiday market

• Points of note:
  – Segment 3 represents **5.1% of the population**, but only **1.4% of the sample**.
  – Segment 12 is **1.8% of the survey** but only **0.6% of the population**.
  – The main **beneficiaries** from a **demographic shift** towards a more elderly populations are mainland Spain, the Balearics and Canary Islands e.g. Tenerife (4,884 new holiday-makers from Leeds), Malaga (3,623), Palma de Mallorca (3,201), Lanzarote (2,073).
  – The big **losers** are long-haul and exotic destinations such as Enfidha (a loss of 5,452 visitors), Sharm-el-Sheikh (4,584), and Cancun (3,450).
5. Urban mobility

- Use **Iterative Proportional Fitting** to generate household indicators including family size, household attributes, transport mode and distance travelled.
- Add **energy use** and **income** based on link to **Understanding Society** using rules based on the household attributes.
5. Urban mobility
5. Urban mobility

Fig. 3. Vulnerability of commuter patterns in Yorkshire and the Humber according to four metrics: (a) Commuter fuel poverty, (b) individual energetic, (c) zonal energetic, (d) hybrid vulnerability. Bins were allocated by Jenks’ classification of natural breaks.
5. Urban mobility

- Rural zone (is) **more vulnerable** .. Illustrating the importance of income, overall energy use and distance from employment centre
- People in **commuter fuel poverty** ... tend to be older ... (perhaps) due to greater reliance on motorised modes amongst less active older citizens
- Quantifying oil vulnerability could help engage with and **steer the policy debate** towards more long-term conceptions of development
  – (Lovelace and Phillips, 2014)
6. Discussion

- Healthy lifestyles
- Consumer choice
- Urban mobility

- Data matching & linkage
- Data profiling & dynamic projection
- IPF and rule-based imputation

- National Diet and Nutrition Survey
- Travel records & questionnaires
- Understanding Society

Data matching & linkage
Data profiling & dynamic projection
IPF and rule-based imputation
National Diet and Nutrition Survey
Travel records & questionnaires
Understanding Society
6. Discussion

- **Efficient representation** of combinations
  - E.g. holidays – age (6), employment (3), occupation (4), gender (2!), MSOAs (108), destinations (78) = 3.6 million combinations

- **Spatial component**
  - Mediation of behaviour through demographic structure
  - Interdependence of lifestyle and environment?
6. Discussion

- ‘Availability of big data offers the prospect of many more intensive investigations’.
6. Discussion

• Value of the research is limited by data quality (e.g. poor resolution of NDNS, skewed holiday data)
  – Reweighting adds substantial value through normalisation
  – Combining individual records e.g. health and consumption is a future aspiration

• Increasing move from synthetic to real data
  – Constrained by consent, confidentiality, data ownership and research ethics
  – Microsimulation could be the middle ground
6. Discussion

- **Linkage of census data** to ‘lifestyles, recreation and consumption ... (can) ... provide a robust basis for diagnosis, forecasting and predictive analytics’
- ‘The use of microsimulation models to **synthesize activity data** e.g. on lifestyle and consumption with other sources of background population data ranging from the electoral roll to housing market data, satellite imagery or employment statistics could provide one of the more promising approaches to (the absence of such data). More studies of this type can be expected in the future, but well-designed tests of the validity of these methods and the suitability of the underlying data are still urgently needed’.