Dynamic Optimisation Microsimulation Using FMF: Practical

Dr Andy Evans (http://www.geog.leeds.ac.uk/people/a.evans/)

**1) Install**

First, download the latest Flexible Modelling Framework (FMF) software from:

<https://github.com/MassAtLeeds/software/releases>

and unzip it somewhere.

**2) Locate the handbook**

In the newly unzipped files and directories you should find a handbooks-and-practicals directory. In this there’s a Microsimulation directory. This contains the software handbook: microsimulationHandbook.pdf plus a set of test-data directories.

**3) Read the first part of the handbook**

Work though up to (but not including) Part 4 of the handbook. This will familiarise you with the basics of the software.

**4) Locate the data**

For this practical, we’ll use the larger test-data dataset. If you look in the test-data directory, and then the “large” directory, you should find the following files:

age.csv

sex.csv

crosstab.csv

microdata.csv

Plus the data licence details: DataLicenceReadMe.txt

Here, we’ll assume these are in a directory on your machine c:\FMF\handbooks-and-practicals\Microsimulation\test-data\large\ -- but it will depend where you unzipped the files.

The first three files are census data for Wards in West Yorkshire. The first contains ward IDs, plus the population numbers broken down by age. The second is the same, but by sex. The third is by age and sex cross-tabulated.

The microdata file is individual-level data, crowdsourced and anonymised. It includes data on the individuals’ age and sex, along with the number of cars they own and their NSSEC8 sociodemographic grouping (<http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec--rebased-on-soc2010--user-manual/index.html#7>).

**5) Run a microsimulation using separate variable tables**

Continue reading the handbook, but use the data from the single-variable files above, rather than the test data suggested, as your example dataset. Before you do this though, read down to (6) below.

We’ll start by running through the system once, with c:\FMF\handbooks-and-practicals\Microsimulation\test-data\large\ as our Data Source, and age.csv, sex.csv, and microdata.csv as our registered files.

Our “Population Data” is the microdata.csv, and we’ll link it to sex.csv and age.csv as our linked data (area statistics) files. We’ll use the “Age” and “Sex” columns in our microdata.csv file (but not the “SexAge” column, yet). We’ll create a synthetic population for the areas that distributes individuals from the microdata in the population areas, such that they match the statistics given for the areas in the age and sex files.

The system will build us a synthetic population file, each line of which will be a person ID and the ward ID they live in (the ward ID will actually be the first column, and the person ID second).

Run through the rest of the user guide, utilising these files. First, though, some hints.

**Hints when using the software**

a) To make a new microsimulation (or validation), click on the Microsimulation menu, and choose “New microsimulation model” (or “New validation setup”).

b) Remember when registering the data files, to check the “Headers in first row” checkbox.

c) Remember on the linked table dialog to select the top row and right-click on it to fix it as the zone IDs.

d) Remember to save the links after you’ve set them up – the system will freeze until you do.

e) When you set up a validation setup, you need to drag in and link the same files as before. However, you also need to drag your results population into the box labelled “Population to evaluate <<results table>>”. Once you’ve done this, click on the table in the Data Sources tree to expand the table in the tree (not open it as a window), and then drag the “ZoneID” column from the tree into the box labelled “Zone ID Field <<not set>>” and the “ID” column into “Person ID Field <<not set>>”. If you don’t set up this population to test, the system will freeze waiting for you to do it.

**6) Run through this again, using the SexAge crosstabulation**

Once you’ve run through the user guide, you should find that it generates a very good population list. Validation should suggest a 100% match with the statistics. This is in part because we are matching our two statistics separately. There’s nothing to stop you matching against cross-tabulated data, so the population has to not only match the age and sex statistics of an area, but also make sure that the right aged people are the right sex. You can try this by linking the “SexAge” column in the microdata.csv against the crosstab.csv file.

In actual fact, as there is a wide variety of people in the microdata, and we’re not using very discriminatory variables, the crosstabulated data should still produce a perfect match. Adding more and more single and cross-tabulated datasets will eventually reduce the match as you start to get combinations that can’t be found in the microdata but which are found in the statistics for an area. For example, if we included sociodemographic groupings we might have a slot in a ward that could only be filled by a 28 year old female unemployed person, but we might not have such a person in our microdata. The system would pick either a 28 year old female manager or a 28 year old male unemployed person, and accept a level of error. As a worse example, our microdata includes “Cars” as a variable, and all our microdata individuals have cars. However, there are plenty of people in our wards that don’t have cars. Any microsimulation using our car ownership data as a variable will be wrong to an error level corresponding at least to the level of no-car people in the population.

**7) Finally, construct the full profiles of the individuals**

When we’re happy with our synthetic population, we can combine our synthetic population with the appropriate microdata by linking each microdata ID / person in the synthetic population with their data in the microdata.csv. In doing so, we not only get the individuals and the variables they have that we constrained our model on, but we also have any ancillary variables that are within their microdata profiles, but not in the area statistics. For example, if we use Age and Sex, we also have Car Ownership in the profiles. If there is a relationship between demographics and the ancillary variables, we can expect the ancillary variable statistics in the areas to be useful, even if we haven’t constrained by them (see discussion, below).

We can combine our synthetic population ID file with the original microdata either with standard database software, creating a table link, or with the “Data Combiner” tool on the Microsimulation menu.

**Discussion**

While the synthetic individuals will be a good match on the variables used to constrain the model (the linked tables above), other ancillary variables attached to each microdata individual may also be a good representation of the characteristics of real population. However, this will depend on the relationship between the constraining variables and the ancillary variables carried over with the microdata.

This needs some thought. It may be, for example, that if one has age, socio-economic group, and education, one might make a good stab at predicting car ownership, however, it is equally possible that it would make a poor prediction of marital status. Equally, including ages 20-70 at five-year age gaps might strengthen the predictability of car ownership, whereas including more data but in different classes, ages 0-30 and 60-100 for example, might weaken the relationship as the statistics will mix more non-drivers and drivers in a single statistical age class for each area. A major issue with microsimulation more generally is where geographical factors external to the constraining variables affects the population – for example, clustering around transport networks.

In general, however, because it uses multiple constraints on an individual level, microsimulation will do better at linking known area / constraining variables to additional ancillary and demographically-determined variables than standard regression techniques – where the independent variables are regarded as entirely independent of each other. With cross-tabulation, microsimulation additionally leverages the constraints of the relationship between the two cross-tabulated variables to improve the accuracy of predicting the population, and therefore the third variable. Because of these extra constraints utilised by microsimulation, standard regression between constraining characteristics and the ancillary variables, which assumes independence, will give a lower limit on the quality of the relationship pulled out by microsimulation. The only other option for validating the relationship is to have a set of known data for the specific ancillary variables, for example from one or more of the geographical areas you’re trying to simulate.

However, even if you don’t want to generate ancillary variable predictions, microsimulation used to recreate constraint variables still has the important advantage that the data is in a disaggregated form which can be reaggregated to different areas.

**Useful reading**

**Dimitris Ballas’ book summarises the Dynamic Optimisation methodology used in the Flexible Modelling Framework Microsimulation plugin, and discusses some of the issues. It also lists some useful microdata datasets:**

**Ballas, D.,** Rossiter, D., Thomas, B., Clarke, G.P. and Dorling, D. (2005). Geography matters: simulating the local impacts of national social policies, Joseph Rowntree Foundation contemporary research issues, Joseph Rowntree Foundation, York. ISBN 1859352650. Free online at: http://www.jrf.org.uk/publications/geography-matters-simulating-local-impacts-national-social-policies

There is an alternative microsimulation technique which gets to the same point via a different route, called Iterative Proportional Fitting. This can generate a better fit if you have smaller microdata samples, but doesn’t generate whole individuals. You can find out more about this in Robin Lovelace’s book, which is currently under development, but is in an almost final form:

Lovelace, R. (in prep) Spatial Microsimulation with R. CRC Press. Also available free online at: http://robinlovelace.net/spatial-microsim-book/