The performance of Iterative Proportional Fitting for spatial microsimulation: applying new tests to an old technique

Robin Lovelace

February 8, 2014

Abstract

Iterative Proportional Fitting (IPF) is an established technique in Regional Science and has been used for a variety of purposes in the field, primary amongst these being the generation of small area microdata. The technique is mature, widely used and relatively straight-forward yet there have been few studies focussed on evaluation of it performance. An additional research problem is the tendency of researchers to 'start from scratch', resulting in a variety ad-hock implementations, with little evidence of the merits of differing approaches. Although IPF is well described mathematically, reproducible examples of the algorithm, written in modern programming language, are rare in the academic literature. These knowledge gaps mean that answers to methodological questions must be guessed: Under what conditions can a 'perfect' match be expected, and how can one test whether these conditions exist in the input datasets? How many iterations should be used for any given application? What is the impact of initial weights on the final results? And what impact will integerisation have on accuracy? This paper tackles such questions, using a systematic methodology based on publicly available code and input data. The results confirm IPF's status as robust and widely applicable method for combining areal and individual-level data and demonstrate the importance of various parameters determining its performace. We conclude with an agenda for future tests and offer more general guidance on how the the method can be optimised to maximise its utility for future research.

Keywords: Iterative proportional fitting, spatial microsimulation, modelling

JEL Code: C

1 Introduction

The aggregation of survey data has a long history in the social sciences. The most common form of aggregation conducted by national statistical agencies collecting data about citizens is 'flattening': converting data from 'long' to 'flat' form. This involves involves a) converting all variables into discrete categories (e.g. an individual with an income of £13,000 per year would be allocated to band of £10,000 to £15,000 per year), b) assigning a column to each category and c) assigning integer counts to cells in each column (Fig. 1). Each row

in this aggregated form can represent the entire dataset or subsets thereof. (Geographical aggregation – one row per zone – is the most common form of aggregation, but any subsets, from social class, time period or other *cross-tablulations* of the dataset can be used).

The main advantage of 'flat' data is that it can summarise information about a very large number of individuals in a small number of rows. Often, such flattened datasets are the only format in which statistical agencies make large datasets (e.g. the National Census) available, to save storage space, ease data analysis and ensure the anonymity and unidentifiability of survey respondents. The main disadvantage is that the process of flattening is almost always associated with data loss: we do not know from a flat dataset if the hypothetical individual mentioned above earns £10,000, £15,000 or anywhere in between. Also, when flattening is associated with geographical aggregation, there is usually a loss of geographical resolution: instead of knowing the person's full postcode from the full dataset, we may only be able to pin them down to a relatively large administrative zone such as a Local Authority in the UK context in the flattened dataset. Finally, the cross-tabulations between the variables are generally during the process of flattening.

Both types of data representation are common in the physical sciences, although the raw data is usually available in long-form. In the social sciences, by contrast, the long format is often unavailable. As indicated above, this can be attributed to the long history of and confidentiality issues surrounding officially collected survey data. National Census datasets — in which characteristics of every individual are recorded — and some of the best datasets in social science are only available in flat form, however, posing major challenges to researchers in the sciences.

A related long-standing research problem especially prevalent in geography and regional science is the scale of analysis. There is often a trade-off between the quality and resolution of the data, which can result in a compromise between detail and accuracy of results (Ballas and Clarke, 2003). A common scenario facing researchers investigating a spatial problem is being presented with two datasets: one long aspatial survey table where each row represents the characteristics of an individual — such as age, sex, employment status and income — and another of 'flat', geographically aggregated dataset containing count data for each category.

As presented in Fig. 1, flat datasets are are wider and shorter, containing multiple columns for each variable from the raw dataset. The loss of information about the linkages between the different variables is common, but not essential, in flat datasets. The solution to this problem, however, makes the flat datasets even more unwieldy, as each new cross tabulation would add $nv1 \times nv2$ additional columns, where $nv1$ and $nv2$ are the number of discrete categories in contained in the variables to be cross tabluted. This lack of cross tabulation is problematic because in many cases information about the linkages between different variables is needed to address policy-relevant questions. To pick one example related to programs targetting skilled unemployed young adults: how many unemployed people reside in each geographical area who at the same time have a university degree, are female and over 25 years old? This question is generally impossible to answer with geographically aggregated flat datasets.

One approach to deal with this problem is through the estimation of small area microdata by combining the small area data with national survey data.

The result of this process of population synthesis (in its 'long' form) is a table in which rows represent each individual in the study region, with an additional column indicating the area of residence (Table 1).

2 IPF: theory and applications

3 Evaluation techniques

4 Input data

5 Method: model experiments

6 Results

7 Conclusions

References

Ballas, D., Clarke, G.P., 2003. Microsimulation and regional science: 30 years of spatial microsimulation of populations, in: 50th Annual North American Meeting of the Regional Science Association, Philadelphia, USA, pp. 19–22.