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

Robin Lovelace

February 9, 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.[1] The 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 policies targeted at skilled unemployed young adults: how many unemployed people reside in each geographical area who at the same time have a university

[1]To provide a concrete example, a cross-tabulation of mode of travel to work by sex and social class would contain 540 (12 * 3 * 15) columns, due to the number of categories present in each variable. Given that there are dozen of additional variables in any large census, this clearly become unwieldy.

degree, are female and over 25 years old? This question is impossible to answer with a geographically aggregated (flat) dataset.

Spatial microsimulation tackles this problem by simulating the individuals within each administrative zone. The most detailed output of this process of spatial microsimulation is 'spatial microdata', a long table of individuals with associate attributes and weights for each zone (table 1). Because these individuals correspond to observations in the survey dataset, the spatial microdata can be stored with only 3 columns: person identification number (ID), zone and weight. Weights are critical to spatial microdata generated in this way, as they determine how representative each individual is of each zone. Thus, an even more compact way of storing spatial microdata taken from a single dataset is as a 'weight matrix' (table 2).

Representing the data in this form offers a number of advantages (Clarke and Holm, 1987). These include:

- The ability to investigate intra-zone variation (by analysing a the variable after subsetting a single area from the data frame),and inter-zone variation in the same dataset.
- Cross-tabulations can be created for any combination of variables (e.g. age and employment status).
- The presentation of continuous variables such as income as real numbers, rather than as categorical count data, in which there is always data loss.
- The data is in a form that is ready to be passed into an individual-level model. For example individual behaviour could simulate traffic, migration or the geographical distribution of future demand for healthcare.

These benefits have not gone unnoticed by the spatial modelling community. Spatial microsimulation is now more than a methodology, and can be considered as a research field in its own right, with a growing body of literature, methods and results (Tanton and Edwards, 2013). In tandem with these methodological developments, a growing number of applied studies has emerged to take advantage of the new possibilities opened up by such small area individual-level data, with applications as diverse as the location of health services (Tomintz et al., 2008) to the analysis of commuting (Lovelace et al., 2014). Various strategies for generating small area microdata are available, as described in recent overviews on the subject (Tanton and Edwards, 2013; Ballas et al., 2013; Hermes and Poulsen, 2012). However, it is not always clear which one is most appropriate, as highlighted in the conclusions of a recent Reference Guide to spatial microsimulation (Clarke and Harding, 2013, p 270): "It would be desirable if the spatial microsimulation community were able to continue to analyse which of the various reweighting/synthetic reconstruction techniques is most accurate — or to identify whether one approach is superior for some applications while another approach is to be preferred for other applications".

The longest-established and most straight-forward of these is Iterative Proportional Fitting (IPF), the performance of which is the subject of this paper...

## 2 IPF: theory and applications

Iterative Proportional Fitting has a long history in Statistics and the social sciences (especially in Economics and Geography). In statistical language, IPF is used to estimate the values of internal cells in a cross-tabulated table to fit known margins. The method is also known as matrix raking or scaling, biproportional fitting, RAS, and entropy maximising. There are a number of texts discussing the method in a mathematical formal way in some detail (for example Mosteller, 1968; Fienberg, 1970; Pritchard and Miller, 2012). (Birkin and Clarke, 1988) provide a succinct presentation of the mathematics of IPF and their work is the foundation of the work presented here...

## 3 Evaluation techniques

## 4 Input data

## 5 Method: model experiments

## 6 Results

## 7 Conclusions

## References


Ballas, D., Clarke, G., Hynes, S., Lennon, J., Morrissey, K., Donoghue, C.O., 2013. A Review of Microsimulation for Policy Analysis, in: O'Donoghue, C., Ballas, D., Clarke, G., Hynes, S., Morrissey, K. (Eds.), Spatial Microsimulation for Rural Policy Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg. Advances in Spatial Science. chapter 3, pp. 35–54.

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.

Birkin, M., Clarke, M., 1988. SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples. Environment and Planning A 20, 1645–1671.

Clarke, G., Harding, A., 2013. Conclusions and Future Research Directions, in: Tanton, R., Edwards, K. (Eds.), Spatial Microsimulation: A Reference Guide for Users. Springer, Berlin, Heidelberg. chapter 16, pp. 259–273.

Fienberg, S., 1970. An iterative procedure for estimation in contingency tables. The Annals of Mathematical Statistics 41, 907–917.

Hermes, K., Poulsen, M., 2012. A review of current methods to generate synthetic spatial microdata using reweighting and future directions. Computers, Environment and Urban Systems 36, 281–290.

Lovelace, R., Ballas, D., Watson, M., 2014. A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. Journal of Transport Geography 34, 282–296.

Mosteller, F., 1968. Association and estimation in contingency tables. Journal of the American Statistical Association 63, 1–28.

Pritchard, D.R., Miller, E.J., 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation 39, 685–704.

Tanton, R., Edwards, K., 2013. Spatial Microsimulation: A Reference Guide for Users. Springer.

Tomintz, M.N.M., Clarke, G.P., Rigby, J.E.J., 2008. The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services. Area 40, 341–353.