

Evaluating the performance of Iterative Proportional Fitting for spatial microsimulation: new tests for an established technique

Robin Lovelace

July 23, 2014

Abstract

Iterative Proportional Fitting (IPF), also known as biproportional fitting or the RAS algorithm, is an established technique in Regional Science and has been used for a variety of purposes in the field. Primary amongst these has been the use of IPF in static spatial microsimulation models for the generation of small area microdata. The technique is mature, widely used and relatively straight-forward yet there have been few studies focused on evaluation of its performance. An additional research problem is the tendency of researchers to ‘start from scratch’, resulting in a variety ad-hoc implementations, with little evidence about the relative merits of differing approaches. Although IPF is well described mathematically, reproducible examples of the algorithm, written in modern programming languages, are rare in the academic literature. These knowledge gaps mean that answers to methodological questions must be guessed: Under what conditions can convergence be expected? How many iterations should be used for any given application? And what are the impacts of initial weights and integerisation on the final result? 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 using various metrics of for determining its performance. The paper also sets an agenda for future tests and offers more general guidance on how the method can be optimised to maximise its utility for future research.

Keywords: Iterative proportional fitting, spatial microsimulation, modelling

JEL Code: C

1 Introduction

Surveys of the composition, attitudes and behavior of individual respondents constitute a ubiquitous empirical basis of social science research. The aggregation of survey data of this type also has a long history. Generally this involves translating raw data from a ‘long’ or ‘microdata’ format — as survey information is entered into a database, one individual at a time — to a ‘wide’ data format (Van Buuren, 2012). In the wide form, each row represents a geographic area or other sub-group of the total population and each column represents a variable value or bin (see Fig. 1). This transformation to a ‘wide’ (or geographically aggregated) form is typically carried out in order to preserve confidentiality, ensuring that no individual can be identified at the small area level (Dale and Marsh, 1993). Wide data is advantageous computationally, as it can represent information about millions of individuals in relatively little computer disk space. An additional benefit from the perspective of spatial analysis is that the wide data form is generally easy to load into geographical information systems (GIS). The process of aggregation is performed in three steps:

1. conversion of 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)
2. assignment of a column to each unique category
3. assignment of integer counts to cells in each column to represent the number of people with each characteristic.

The process of aggregation, represented in above bullet points, has some important disadvantages, however. It is these disadvantages, summarised in the next paragraph, that spatial microsimulation and small area estimation seek to address.¹ Both combine national social survey microdata (such as the Understanding

¹There is a great deal of overlap between methods for spatial microsimulation and small area estimation (the terms are sometimes used interchangeably). It is useful to have two terms, however. Small area estimation seeks to estimate summary statistics for a variable whose geographical distribution is unknown, sometimes via a spatial microdataset. With spatial microsimulation, on the other hand, the emphasis is on generating and analysing spatial microdata.

Society panel dataset UK) with small area census data. A number of techniques are available, including deterministic reweighting (Birkin and Clarke, 1989), genetic algorithms and simulated annealing (Williamson et al., 1998) as well as approaches based on linear regression (Harding et al., 2011). It is an active field of development with on-going refinements (e.g. (Lovelace and Ballas, 2013; Pritchard and Miller, 2012)) and debates about which approaches are most suited in different situations (Harland et al., 2012; Hermes and Poulsen, 2012; Smith et al., 2009; Whitworth, 2013; Williamson, 2013)

These methods are important because large national social survey datasets such as censuses are often *only* available in the wide, geographically aggregated form. Information loss occurs when georeferenced individual level survey data — *spatial microdata* — is aggregated in this way. This problem is exacerbated when continuous variables are converted into discrete categories and when not cross-tabulations, such as age/income, are not provided, as is often the case. Geographically aggregated datasets are problematic for the analysis of spatial phenomena (Openshaw, 1984) and can prevent analysis of the interdependencies individual and household variables at the local level (Lee, 2009). In an effort to address the drawbacks associated with this lack of microdata, Samples of Anonymised Records (SARs) were released from the 1991 Census (Dale and Marsh, 1993; Middleton, 1995) and this was followed by similar SARs releases in 2001 and 2011. Nevertheless, the SARs are not suitable for intra-city and neighbourhood level analysis, as the smallest geographical level at which such microdata are publicly available is the Local Authority, to preserve confidentiality. There has been discussion ways to further disaggregate the geography of the SARs, with due emphasis on risks to confidentiality (Tranmer et al., 2005). Furthermore, there are many policy relevant variables (e.g. income and wealth, social attitudes) that are not collected by censuses of population, making them unavailable at the small area level. Geographers and regional scientists have addressed this data paucity through the development and application of small area estimation methods (Ballas et al., 2005; Hermes and Poulsen, 2012).

Iterative proportional fitting (IPF) can help overcome the limitations of wide, geographically aggregated data sources when used in *spatial microsimulation*. In simplistic terms spatial microsimulation can be understood as the reversal of geographic aggregation described in points 1–3 above, to approximate the original ‘long’ dataset (Fig. 1). Thus the need for spatial microsimulation is driven by the tendency of statistical agencies to make the raw georeferenced microdata generated by large social surveys unavailable to the research community (Lee, 2009). In fact spatial microsimulation, the term favoured by geographers and regional scientists, is generally known as *population synthesis* in transport modelling, which is an apt name: its core function is to simulate populations based on imperfect data. Spatial microsimulation can also refer to a wider approach which harnesses spatial microdata in the analysis stage (Clarke and Holm, 1987; Lovelace et al., 2014). In this paper we focus on the former meaning, which is also known as *static spatial microsimulation* (Ballas et al., 2005).

1.1 What is IPF and why use it?

In abstract terms, IPF is a procedure for assigning values to *internal cells* based on known marginal totals in a multidimensional matrix (Cleave et al., 1995). More practically, this enables the calculation of the *maximum likelihood* estimate of the presence of given individuals in specific zones. Some confusion surrounds the procedure because it has multiple names: the ‘RAS algorithm’, ‘biproportional fitting’ and to a lesser extent ‘raking’ have all been used to describe the procedure at different times and in different branches of academic research (Lahr and de Mesnard, 2004). ‘Entropy maximisation’ is a related technique, of which IPF is considered to be a subset (Johnston and Pattie, 1993; Rich and Mulalic, 2012).² Usage characteristics of these different terms are summarised in Table 1.

Table 1: Usage characteristics of different terms for the IPF procedure, from the Scopus database

| Search term | N. citations | Commonest subject area | % Since 2010 | % Mentioning “regional” |
|----------------------------------|--------------|------------------------|--------------|-------------------------|
| Entropy maximisation | 564 | Engineering | 30 | 9 |
| “Iterative proportional fitting” | 135 | Mathematics | 22 | 16 |
| “RAS Algorithm” | 25 | Comp. science | 24 | 24 |
| “Iterative proportional scaling” | 15 | Mathematics | 27 | 0 |
| Matrix raking | 12 | Engineering | 50 | 0 |
| Biproportional estimation | 12 | Social sciences | 17 | 33 |

²As an aside, IPF is considered by Rich and Mulalic (2012) to be a particular efficient for of entropy maximisation: “The popularity of the IPF is therefore mainly due to the fact that it provides a solution which is equivalent to that of the ME approaches, but attained in a much more computationally efficient way” (p. 469).

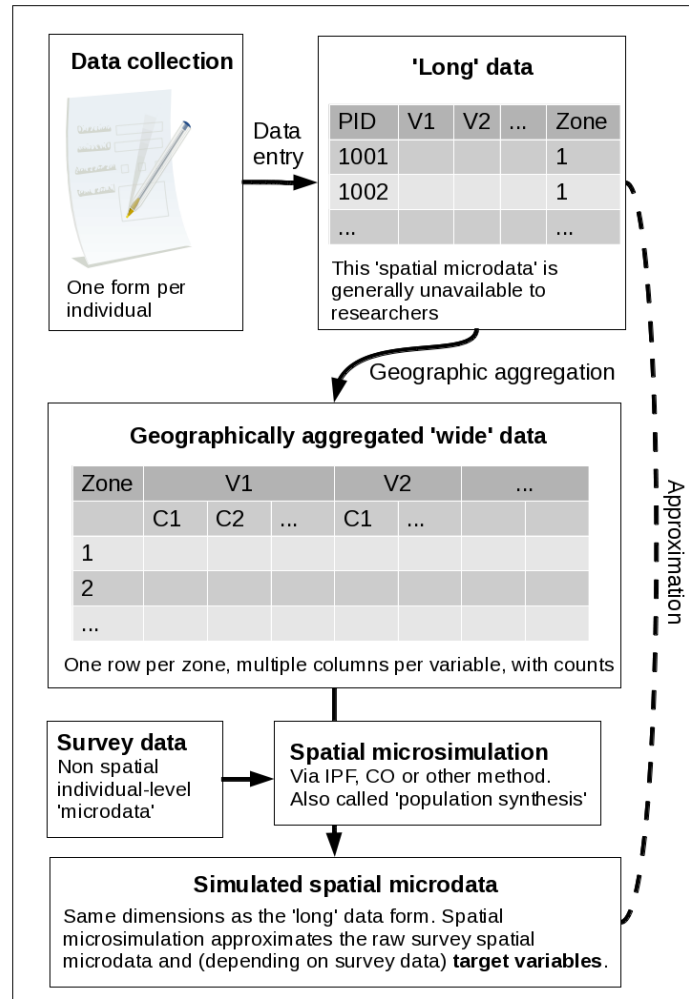


Figure 1: Schema of iterative proportional fitting (IPF) and combinatorial optimisation in the wider context of the availability of different data formats and spatial microsimulation.

Ambiguity in terminology surrounding the procedure has existed for a long time: when Michael Bacharach began using the term ‘biproportional’ late 1960s, he did so with the following caveat: “The ‘biproportional’ terminology is not introduced in what would certainly be a futile attempt to dislodge ‘RAS’, but to help to abstract the mathematical characteristics from the economic associations of the model” (Bacharach, 1970). The original formulation of the procedure is generally ascribed to Deming and Stephan (1940), who presented an iterative procedure for solving the problem of how to modify internal cell counts of census data in ‘wide’ form, based on new marginal totals. Without repeating the rather involved mathematics of this discovery, it is worth quoting again from Bacharach (1970), who provided as clear a description as any of the procedure in plain English: “The method that Deming and Stephan suggested for computing a solution was to fit rows to the known row sums and columns to the known column sums by alternate [iterative] scalar multiplications of rows and columns of the base matrix” (p. 4). The main purpose for revisiting and further developing the technique from Bacharach’s perspective was to provide forecasts of the sectoral disaggregation of economic growth models. It was not until later (e.g. Clarke and Holm, 1987) that the procedure was rediscovered for geographic research.

Thus IPF is a general purpose tool with multiple uses. In this paper we focus on its ability to *disaggregate* the geographically aggregated data typically provided by national statistical agencies to synthesise *spatial microdata*, as illustrated in Fig. 1. But why would one want to do this in the first place?

To answer this question, let us briefly consider the advantages and drawbacks of the ubiquitous wide data format. The main advantages of wide datasets are that they can summarise information about very many individuals in a relatively few rows, saving storage space, easing the analysis and, critically for policy makers, ensuring the anonymity of survey respondents. But the process of aggregation and associated information loss leads to the following disadvantages:

- *Binning of continuous variables:* we do not know from a wide dataset if the aforementioned person earning £13,000 is in fact on a salary of £10,000, £15,000 or anywhere in between
- *Loss of geographical resolution:* when aggregation is done by geographical zones spatial information is lost. Instead of knowing person’s precise home location, from the full dataset, we may only be able to pin them down to a relatively large administrative area.
- *Loss of cross-tabulation:* during the process of aggregation the links between different variables, often expressed as ‘contingency tables’, is usually lost. For example a female cyclist may be represented as additional units in the “cyclist” and “female” categories, without knowledge of their interrelation.

In the *physical sciences* both data formats are common and the original long dataset is generally available. In the social sciences, by contrast, the long format is often unavailable due to the long history of and confidentiality issues surrounding officially collected survey data. It is worth emphasising that National Census datasets — in which characteristics of every individual living in a state are recorded and which constitute some of the best datasets in social science — are usually only available in wide form. This poses major challenges to researchers in the sciences who interested in intra-zone variability at the local level (Whitworth, 2012; Lee, 2009).

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 wide, geographically aggregated dataset containing count data for each category. IPF solves this problem when used in spatial microsimulation by proportionally assigning individuals to zones of which they are most representative.

1.2 IPF in spatial microsimulation

Spatial microsimulation tackles the aforementioned problems associated with the prevalent geographically aggregated wide data form by simulating the individuals within each administrative zone. The most detailed output of this process of spatial microsimulation is *synthetic spatial microdata*, a long table of individuals with associated attributes and weights for each zone (see Table 2 for a simple example). Because these individuals correspond to observations in the survey dataset, the spatial microdata generated in this way, of any size and complexity, 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 synthetic spatial microdata

taken from a single dataset is as a ‘weight matrix’, provided the original individual-level dataset is available, to recover individuals’ attributes. the .

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

- The ability to investigate intra-zone variation (variation within each zone can be analysed by subsetting a single area from the data frame).
- 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 considered by some to be more than a methodology: it can be considered as a research field in its own right, with a growing body of literature, methods and software tools (Tanton and Edwards, 2013). In tandem with these methodological developments, a growing number of applied studies take advantage of the possibilities opened up by spatial microdata. Applications are diverse, including the following:

- Projecting the wider individual and local impacts of job losses and gains from the closure (Ballas et al., 2006) or opening (Van Leeuwen, 2010) of new businesses.
- A model of the housing market, where outputs for a spatial microsimulation model were used as an input into an agent based model to assess the likely outcome of urban regeneration interventions (Jordan et al., 2011).
- Inequalities between individuals and zones in educational opportunities and attainment (Kavroudakis et al., 2012).
- Analysis of commuting patterns and the distributional impacts of new transport policies (Lovelace et al., 2014)
- Investigation of the local impacts of economic development in the marine energy sector in Ireland (Morrissey et al., 2013).
- EUROMOD, a model for analysing the impact of changes in taxation and financial laws on European citizens (Sutherland and Figari, 2013).
- Evidence to support the optimal location of health services based on estimates of individual level behaviour and obesity rates (Tomintz et al., 2008; Edwards et al., 2010).
- The use of spatial microsimulation in combination with spatial interaction and agent based models to simulated the spatial distribution and movement of student populations (Wu et al., 2010).
- Estimation of the energy impacts of freight transport at the national level, by using spatial microsimulation alongside a spatial interaction model (Zuo et al., 2013).

Various strategies for generating the synthetic spatial microdata used in such studies 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 is most appropriate in different situations. There has been work testing and improving model results (Teh and Welling, 2003), yet research focused on systematically and transparently testing the underlying algorithms for generating synthetic spatial microdata is limited (Harland et al., 2012). This research gap, addressed in this paper, is 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”.

2 The IPF algorithm: a worked example

In most modelling texts there is a precedence of theory over application: the latter usually flows from the former (e.g. Batty, 1976). As outlined in Section 1.1, however, the theory underlying IPF have been thoroughly demonstrated in research stemming back to the 1940s. *Practical* demonstrations are comparatively rare.

That is not to say that reproducibility is completely absent in the field. Simpson and Tranmer (2005) and Norman (1999) exemplify their work with code snippets from an implementation of IPF in the proprietary software SPSS and reference to an Excel macro that is available on request, respectively. Williamson (2007) also represents good practice, with a manual for running combinatorial optimisation algorithms described in previous work. Ballas et al. (2005) demonstrate IPF with a worked example to be calculated using mental arithmetic, but do not progress to illustrate how the examples are then implemented in computer code. This section builds on and extends this work, by first outlining an example that can be completed using mental arithmetic alone before illustrating its implementation in code. A comprehensive introduction to this topic is provided in a tutorial entitled “Introducing spatial microsimulation with R: a practical” (Lovelace, 2014). This section builds on the tutorial by providing context and a higher level description of the process.

We begin with some hypothetical data. Table 2 describes a hypothetical microdataset comprising 5 individuals, who are defined by two constraint variables, age and sex. Table 3 contains aggregated data for a hypothetical area, as it might be downloaded from census dissemination portals such as Casweb. Table 4 illustrates this table in a different form, demonstrating the unknown links between age and sex. The job of the IPF algorithm is to establish estimates for the unknown cells, which are optimized to be consistent with the known row and column marginals.

Table 2: A hypothetical input microdata set (the original weights set to one). The bold value is used subsequently for illustrative purposes.

| Individual | Sex | Age | Weight |
|------------|--------|-----|----------|
| 1 | Male | 59 | 1 |
| 2 | Male | 54 | 1 |
| 3 | Male | 35 | 1 |
| 4 | Female | 73 | 1 |
| 5 | Female | 49 | 1 |

Table 3: Hypothetical small area constraints data (s).

| Constraint \Rightarrow | i | | j | |
|--------------------------|----------|---------|-------|--------|
| Category \Rightarrow | i_1 | i_2 | j_1 | j_2 |
| Area \Downarrow | Under-50 | Over-50 | Male | Female |
| 1 | 8 | 4 | 6 | 6 |

Table 4: Small area constraints expressed as marginal totals, and the cell values to be estimated.

| | Marginal totals | j | | |
|-----|-----------------|------|--------|----|
| | Age/sex | Male | Female | T |
| i | Under-50 | ? | ? | 8 |
| | Over-50 | ? | ? | 4 |
| | T | 6 | 6 | 12 |

Table 5 presents the hypothetical microdata in aggregated form, that can be compared directly to Table 4. Note that the integer variable age has been converted into a categorical variable in this step, a key stage in geographical aggregation.

Table 5: The aggregated results of the weighted microdata set ($m(1)$). Note, these values depend on the weights allocated in Table 2 and therefore change after each iteration

| | Marginal totals | j | | | |
|-----|-----------------|----------|--------|---|--|
| | Age/sex | Male | Female | T | |
| i | Under-50 | 1 | 1 | 2 | |
| | Over-50 | 2 | 1 | 3 | |
| | T | 3 | 2 | 5 | |

Using these data it is possible to readjust the weights of the hypothetical individuals, so that their sum adds up to the totals given in Table 4 (12). In particular, the weights can be readjusted by multiplying them by the marginal totals, originally taken from Table 3 and then divided by the respective marginal total in 5. Because the total for each small-area constraint is 12, this must be done one constraint at a time. This can be expressed, for a given area and a given constraint (i or age in this case), as follows:

$$w(n+1)_{ij} = \frac{w(n)_{ij} \times sT_i}{mT(n)_i} \quad (1)$$

where $w(n+1)_{ij}$ is the new weight for individuals with characteristics i (age, in this case), and j (sex), $w(n)_{ij}$ is the original weight for individuals with these characteristics, sT_i is element marginal total of the small area constraint, s (Table 3) and $mT(n)_i$ is the marginal total of category j of the aggregated results of the weighted microdata, m (Table 5). n represents the iteration number. Although the marginal totals of s are known, its cell values are unknown. Thus, IPF estimates the interaction (or cross-tabulation) between constraint variables. (Follow the emboldened values in the tables to see how the new weight of individual 3 is calculated for the sex constraint.) Table 6 illustrates the weights that result. Notice that the sum of the weights is equal to the total population, from the constraint variables.

Table 6: Reweighting the hypothetical microdataset in order to fit Table 3.

| Individual | Sex | age-group | Weight | New weight, $w(2)$ |
|------------|--------|-----------|--------|--------------------------------------|
| 1 | Male | Over-50 | 1 | $1 \times 4/3 = \frac{4}{3}$ |
| 2 | Male | Over-50 | 1 | $1 \times 4/3 = \frac{4}{3}$ |
| 3 | Male | Under-50 | 1 | $1 \times 8/2 = 4$ |
| 4 | Female | Over-50 | 1 | $1 \times 4/3 = \frac{4}{3}$ |
| 5 | Female | Under-50 | 1 | $1 \times 8/2 = 4$ |

After the individual level data have been re-aggregated (Table 7), the next stage is to repeat Eq. (1) for the age constraint to generate a third set of weights, by replacing the i in sT_i and $mT(n)_i$ with j and incrementing the value of n :

$$w(3)_{ij} = \frac{w(2)_{ij} \times sT_j}{mT(2)_j} \quad (2)$$

Apply Eq. (2) to the information above and that presented in Table 7 results in the following vector of new weights, for individuals 1 to 5:

$$w(3) = \left(\frac{6}{5}, \frac{6}{5}, \frac{18}{5}, \frac{3}{2}, \frac{9}{2} \right) \quad (3)$$

As before, the sum of the weights is equal to the population of the area (12). Notice also that after each iteration the fit between the marginal totals of m and s improves. The total absolute error (TAE) from $m(1)$ to $m(2)$ improves from 14 to 6 in Table 5 and Table 7 above. TAE for $m(3)$ (not shown, but calculated by aggregating $w(3)$) improves even more, to 1.3. This number would eventually converge to 0 through subsequent iterations, as there are no empty cells in the input microdataset, a defining feature of IPF (see Section 5.4 for more on the impact of empty cells).

The above process, when applied to more categories (e.g. socio-economic class) and repeated iteratively until a satisfactory convergence occurs, results in a series of weighted microdatasets, one for each of the small areas being simulated. This allows for the estimation of variables whose values are not known at the local level (e.g. income) (Ballas et al., 2005). An issue with the results of IPF (absent from combinatorial optimisation methods), however, is that it results in non-integer weights: fractions of individuals appear in simulated areas. This is not ideal for certain applications, leading to the development of strategies to

Table 7: The aggregated results of the weighted microdata set after constraining for age ($m(2)$).

| Marginal totals | Age/sex | i | | |
|-----------------|----------|----------------|----------------|----|
| | | Male | Female | T |
| j | Under-50 | 4 | 4 | 8 |
| | Over-50 | $\frac{8}{3}$ | $\frac{4}{3}$ | 4 |
| | T | $6\frac{2}{3}$ | $5\frac{1}{3}$ | 12 |

‘integerise’ the fractional weights illustrated in Table 7. Methods for performing integerisation are discussed in (Ballas et al., 2005), (Pritchard and Miller, 2012) and, most recently in (Lovelace and Ballas, 2013), who systematically benchmarked and compare different procedures. It was found that the ‘truncate, replicate, sample’ (TRS) and ‘proportional probabilities’ methods were most accurate. These methods are therefore used to test the impact of integerisation in this paper.

2.1 Implementing IPF in R

The above example is best undertaken by hand, probably with a pen and paper to gain an understanding of IPF, before the process is automated for larger datasets. This section explains how the IPF algorithm described above can be implemented in R.

3 Evaluation techniques

To verify the integrity of any model, it is necessary to compare its outputs with empirical observations or prior expectations gleaned from theory. The same principles apply to spatial microsimulation, which can be evaluated using both *internal* and *external* validation methods (Edwards and Clarke, 2009). *Internal validation* is the process whereby the aggregate-level outputs of spatial microsimulation are compared with the input constraint variables. It is important to think carefully about evaluation metrics because results can vary depending on the measure used (Voas and Williamson, 2001).

Internal validation typically tackle such issues as “do the simulated populations in each area microdata match the total populations implied by the constrain variables?” and “what is the level of correspondence between the cell values of different attributes in the aggregate input data and the aggregated results of spatial microsimulation?” A related concept from the agent based modelling literature is *verification*, which asks the more general question, “does the model do what we think it is supposed to do?” (Ormerod and Rosewell, 2009, p. 131) *External validation* is the process whereby the variables that are being estimated are compared to data from another source, external to the estimation process, so the output dataset is compared with another known dataset for those variables. Both validation techniques are important in modelling exercises for reproducibility and to ensure critical advice is not made based on coding errors (Ormerod and Rosewell, 2009).

For dynamic microsimulation scenarios, calculating the sampling variance (the variance in a certain statistic, usually the mean, across multiple samples) across multiple model runs is recommended to test if differences between different scenarios are statistically significant (Goedemé et al., 2013). In static spatial microsimulation, sub-sampling from the individual level dataset or running stochastic processes many times are ways to generate such statistics for tests of significance. Yet these statistics will only apply to a small part of the modelling process and are unlikely to relate to the validity of the process overall (Lovelace and Ballas, 2013).

This section briefly outlines the evaluation techniques used in this study for *internal validation* of static spatial microsimulation models. The options for external validation are heavily context dependent so should be decided on a case-by-case basis. External validation is thus mentioned in the discussion, with a focus on general principles rather than specific advice. The three quantitative methods used in this study (and the visual method of scatter plots) are presented in ascending order of complexity and roughly descending order of frequency of use, ranging from the coefficient of determination (R^2) through total and standardised absolute error (TAE and SAE , respectively) to root mean squared error ($RMSE$). This latter option is our preferred metric, as described below.

Before these methods are described, it is worth stating that the purpose is not to find the ‘best’ evaluation metric: each has advantages and disadvantages, the importance of which will vary depending on the nature of the research. The use of a variety of techniques in this paper is of interest in itself. The high degree of correspondence between them (presented in Section 6), suggests that researchers need only to present one

or two metrics (but should try more, for corroboration) to establish relative levels of goodness of fit between different model configurations.

The problem of internal validation (described as ‘evaluating model fit’) was tackled in detail by Voas and Williamson (2001). Chi-squared, Z-scores, absolute error and entropy-based methods were tested, with no clear ‘winner’ other than the advocacy of methods that are robust, fast and easy to understand (standardized absolute error measures match these criteria well). The selection of evaluation methods is determined by the nature of available data: “Precisely because a full set of raw census data is not available to us, we are obliged to evaluate our simulated populations by comparing a number of aggregated tables derived from the synthetic data set with published census small-area statistics” (Voas and Williamson, 2001, p. 178). A remaining issue is that the results of many evaluation metrics vary depending on the areal units used, hence advocacy of ‘scale free’ metrics (e.g. (Malleson and Birkin, 2012)).

3.1 Scatter plots

A scatter plot of cell counts for each category for the original and simulated variables is a basic but very useful preliminary diagnostic tool in spatial microsimulation (Ballas et al., 2005; Edwards and Clarke, 2009). In addition, marking the points, depending on the variable and category which they represent, can help identify which variables are particularly problematic, aiding the process of improving faulty code (sometimes referred to as ‘debugging’).

In these scatter plots each data point represents a single zone-category combination (e.g. 16-25 year old female in zone 001), with the x axis value corresponding to the number of individuals of that description in the input constraint table. The y value corresponds to the aggregated count of individuals in the same category in the aggregated (wide) representation of the spatial microdata output from the model. This stage can be conducted either before or after *integerisation* (more on this below). As taught in basic statistics courses, the closer the points to the 1:1 (45 degree, assuming the axes have the same scale) line, the better the fit.

3.2 The coefficient of determination

The coefficient of determination (henceforth R^2) is the square of the Pearson correlation coefficient for describing the extent to which one continuous variable explains another. R^2 is a quantitative indicator of how much deviation there is from this ideal 1:1 line. It varies from 0 to 1, and in this context reveals how closely the simulated values fit the actual (census) data. An R^2 value of 1 represents a perfect fit; an R^2 value close to zero suggests no correspondence between then constraints and simulated outputs (i.e. that the model has failed). We would expect to see R^2 values approaching 1 for internal validation (the constraint) and a strong positive correlation between target variables that have been estimated and external datasets (e.g. income). A major limitation of R^2 is that it takes no account of *additive* or *proportional* differences between estimated and observed values, meaning that it ignores systematic bias (Legates and McCabe, 1999). This limitation is less relevant for spatial microsimulation results because they are generally population-constrained.

3.3 Total Absolute and Standardized Error (TAE and SAE)

Total absolute error (TAE) is the sum of the absolute (positive) discrepancies between the simulated and known population in an area. The standardized absolute error (SAE) is simply TAE divided by the total population. TAE provides no level of significance, so thresholds have been used. Clarke and Madden (2001) used an error threshold of 80 per cent of the areas with less than 20 per cent error ($SAE < 0.20$). Smith et al. (2007) work with a model that simulates with diabetes and use an error threshold of 10% ($SAE < 0.1$) in 90 per cent of the output areas. Due to its widespread use, simplicity and scale-free nature — allowing the fit for scenarios 1, 2 and 3 to be compared — SAE is our preferred metric in this study, although we also calculated the root mean square error for each test.

3.4 Root mean square error

Root mean square error ($RMSE$) and to a lesser extent mean absolute error (MAE) have been frequently used in both social and physical sciences as a metric of model fit (Legates and McCabe, 1999; Ngo and See, 2012). The former has the benefit of being widely used and understood, but is influenced by the mean value in each zone. Research in climate model evaluation suggests that the lesser known but simpler MAE metric is preferable to $RMSE$ in many ways (Willmott and Matsuura, 2005). Therefore both metrics are presented in this paper.

4 Input data

Three input datasets form the basis of the scenarios tested in this paper, ‘simple’, ‘small-area’ and ‘sheffield’. Each consists of a set of aggregate constraints and corresponding individual-level survey tables. The latter are anonymous and have been ‘scrambled’ to enable their publication online. This enables the results to be replicated by others, without breaching the user licenses of the data. To this end, all the input data (and code) used to generate the results reported in the paper are available online on the code-sharing site GitHub. The simplest scenario consists of 6 zones which must be populated by the data presented in the previous section (see Table 2). This scenario is purely hypothetical and is used to represent the IPF procedure and the tests in the simplest possible terms. Its small size allows the entirety of the input data to be viewed in a couple of tables (see Table 2 and Table 8, the first row of which is used to exemplify the method the previous section). As with all data used in this paper, these tables can be viewed and downloaded online.

Table 8: Geographically aggregated constraint variables for the simple scenario.

| Zone | Age constraint | | Sex constraint | |
|------|----------------|-----|----------------|---|
| | 16–49 | 50+ | m | f |
| 1 | 8 | 4 | 6 | 6 |
| 2 | 2 | 8 | 4 | 6 |
| 3 | 7 | 4 | 3 | 8 |
| 4 | 5 | 4 | 7 | 2 |
| 5 | 7 | 3 | 6 | 4 |
| 6 | 5 | 3 | 2 | 6 |

The next smallest input dataset, and the most commonly used scenario in the paper, is ‘small-area’. The area constraints consist of 24 Output Areas (the smallest administrative unit in the UK), with an average adult population of 184. Three files describe the number of hours worked, the marital status and the tenancy of these individuals (although tenancy in fact counts houses, not people): the files can be accessed from ‘small-area-eg’, within the ‘input-data’ folder of the online repository.³

The individual-level table that compliments these geographically aggregated count constraints is taken from Understanding Society wave 1. These have been randomised to ensure anonymity but they are still roughly representative of real data. The table consists of 1768 rows of data and is stored in R’s own space-efficient filetype (see “ind.RData”, which must be opened in R).

5 Method: model experiments

5.1 Baseline conditions

In order to perform a wide range of model experiments, in which only one factor is altered at a time, baseline scenarios were created. These are model runs with sensible defaults and act as reference points against which alternative runs are compared. The baseline scenarios consist of 3 iterations. In the ‘simple’ scenario (with 2 constraints), there are 6 constraint-iteration permutations. Table 10 shows fit improving with each step.⁴

The main characteristics of the baseline model for each scenario area presented in Table 9. The meaning of each of the attributes in the row names of will be explained later in this section.

5.2 Constraints and iterations

These were the simplest model experiments, requiring only the reporting of how the results changed with more or fewer iterations and re-arrangement of to apply constraints in a different order. 1, 3 and 10 are reported in the final results and ‘convergence graphs’ were created to show how model fit improved from one iteration-constraint combination to the next.

To alter the number of iterations used in each model, one simply changes the number in the following line of code (the default is 3):

³These datasets can be found here: <https://github.com/Robinlovelace/IPF-performance-testing/tree/master/input-data/small-area-eg>

⁴The exception is *p5*, which reaches zero in the second iteration. *pcor* continues to improve however, even though the printed value, to 6 decimal places, is indistinguishable from 1 after two complete iterations.

Table 9: Characteristics of the baseline model run for each scenario.

| Scenario | Simple | Small area | Sheffield |
|----------------------|--------|------------|-----------|
| N. constraints | 2 | 3 | 4 |
| N. zones | 6 | 24 | 71 |
| Av. individuals/zone | 10 | 184 | 3225 |
| Microdataset size | 5 | 1768 | 4933 |
| Average weight | 1.3 | 0.1 | 0.65 |
| % empty cells | 0% | 21% | 85% |

```
num.its <- 3
```

The code for running the constraints in a different order is saved in ‘etsim-reordered.R’.

5.3 Initial weights

The impact of initial weights was tested in two ways: first, the impact of changing a sample of one or more initial weight values by a set amount was explored. Second, the impact of gradually increasing the size of the initial weights was tested, by running the same model experiment several times, but with incrementally larger initial weights applied to a pre-determined set of individuals.

In the simplest scenario, the initial weight of a single individual was altered repeatedly, in a controlled manner. The code enabling this model experiment is contained in the file ‘etsim-weights.R’. The number of different initial weight runs can be controlled by the parameter ‘num.ws’ which is set to 5 as an arbitrary default. Initial weights are set to k , a vector of length ‘num.ws’ which are by default set as a function of ‘u’ which is in turn defined as the sequence $1:\text{num.ws}$:

```
k <- u/5 + 0.5 # initial weight of sample individuals for testing
```

In this case the resultant weights are set to 0.7, 0.9, 1.1, 1.3 and 1.5 for each model experiment. For factors of one million, for example, the above would be replaced by $k <- u * 10^6 \text{ yrm}$. To observe the impacts of altering the weights in this way the usual tests of model fit were conducted. In addition, the capacity of initial weights to influence an individual’s final weight/probability of selection was tested by tracing its weight into the future after each successive constraint and iteration. This enabled plotting original vs simulated weights under a range of conditions (Section 6).

5.4 Empty cells

Empty cells refer to combinations individual-level attributes that are not represented by any single individual in the survey input data. An example of this would be if no female in the 16 to 49 year age bracket were present in Table 2. Formalising the definition, survey dataset that contains empty cells may be said to be *incomplete*; a survey dataset with no empty cells is *complete*.

No method for identifying whether or not empty cells exist in the individual-level dataset for spatial microsimulation has been previously published to the authors’ knowledge. We therefore start from first principles. The number of different constraint variable permutations (N_{perm}) is defined by Eq. (4), where n_{cons} is the total number of constraints and n_{cat_i} is the number of categories within constraint i :

$$N_{perm} = \prod_{i=1}^{n_{cons}} n_{cat_i} \quad (4)$$

To exemplify this equation, the number of permutations of constraints in the ‘simple’ example is 4: 2 categories in the sex variables multiplied by 2 categories in the age variable. Clearly, N_{perm} depends on how continuous variables are binned, the number of constraints and diversity within each constraint. Once we know the number of unique individuals (in terms of the constraint variables) in the survey (N_{uniq}), the test to check a dataset for empty cells is straightforward, based on Eq. (4):

$$is.complete = \begin{cases} TRUE & \text{if } N_{uniq} = N_{perm} \\ FALSE & \text{if } N_{uniq} < N_{perm} \end{cases} \quad (5)$$

Once the presence of empty cells is determined in the baseline scenarios, the next stage is to identify which individuals are missing from the individual-level input dataset (Ind) — the combination of attributes

that no individual in *Ind* has. To do this, we created an algorithm to generate the the complete input dataset (*Ind.complete*). The ‘missing’ individuals, needed to be added to make *Ind* complete can be defined in terms of set theory (Goldrei, 1996) by Eq. (6):

$$Ind.missing = Ind.complete \setminus Ind = \{x \in Ind.complet : x \notin Ind\} \quad (6)$$

This means simply that the missing cells are defined as individuals with constraint categories that are present in the complete dataset but absent from the input data. The theory to test for and identify empty cells described above is straightforward to implement in practice with R, using the commands `unique` and `%in%` respectively (see the ‘empty-cells.R’ script in the ‘model’ folder of our code repository).

To test the impact of empty cells on model results, an additional empty cell was added to the input dataset in each example.⁵ Experiments in which empty cells were removed through the addition of individuals with attributes defined by Eq. (6) were also conducted, and the model fit was compared with the baseline scenarios. The baseline scenarios were run using these modified inputs.

5.5 Integerisation

Integerisation is the process of converting the fractional weights produced by IPF into whole numbers. It is necessary for some applications, including the analysis of intra-zone variability and statistical analyses of intra-zone variation such as the generation of Gini indices of income for each zone.

A number of methods for integerisation have been developed, including deterministic and probabilistic approaches (Ballas et al., 2005). We used modified R implementations of two probabilistic methods — ‘truncate replicate sample’ and ‘proportional probabilities’ — reported in (Lovelace and Ballas, 2013) and the supplementary information associated with this paper (Lovelace, 2013), as these were found to perform better in terms of final population counts and accuracy than the deterministic approaches.

6 Results

6.1 Baseline results

After only 3 complete iterations under the baseline conditions set out in Section 5.1, the final weights in each scenario converged towards a final result.

The baseline ‘Sheffield’ scenario consists of the same basic set-up, but with 4 constraints. As before, the results show convergence towards better model fit, although the fit did not always improve from one constraint to the next (see the online ‘measures’ table).

Table 10: Results for the baseline ‘simple’ scenario. Generated through by the ‘analysis’ script, and available as a .csv file online.

6.2 Constraints and iterations

As expected from Fig. 3 above, fit improved with each iteration. This improvement generally happened for each constraint and always from one iteration to the next in every scenario. Through the course of the first iteration in the ‘simple’ senario, for example, R^2 improved from 0.67 to 0.9981. As illustrated in Fig. 2, the rate of improvement in model fit plateaus after the first iterate: at the end of iteration 2 in the simple scenarios, R^2 had improved to 0.999978.

The results for the first 6 constraint-iteration permutations of this baseline scenario are illustrated in Fig. 3 and in the online ‘measures’ table. These results show that IPF successfully took the simulated cell counts in-line with the census after each constraint and that the fit improves with each iteration (only two iterations are shown here, as the fit improves very little beyond two).

Rapid reductions in the accuracy improvement from additional computation were observed in each scenario, as illustrated by the falling percentage change in TAE in Fig. 3. After four iterations, in the ‘simple’ scenarios, the fit is near perfect, with an R^2 value indistinguishable from one to 7 decimal places. The decay in the rate of improvement with additional iterations is super-exponential.

⁵In the ‘simple’ scenario, this simply involves removing a single individual from the input data, which equates to deleting a single line of code (see ‘empty-cells.R’ in the ‘simple’ folder to see this).

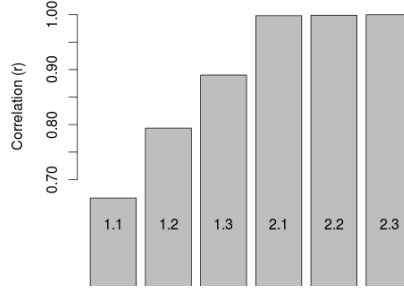


Figure 2: Improvement in goodness of fit with additional constraints and iterations, as illustrated by R^2 values.

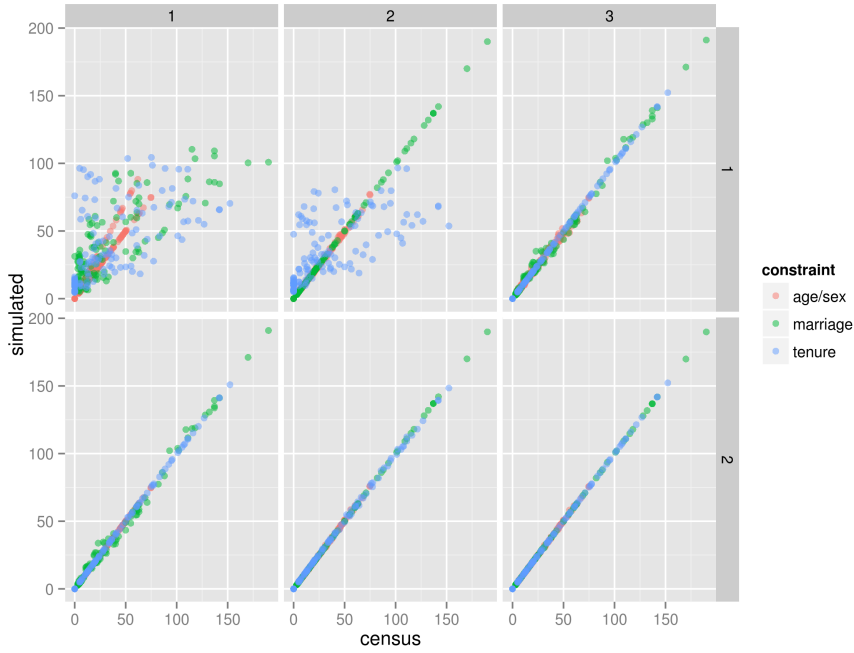


Figure 3: IPF in action: simulated vs census zone counts for the small-area baseline scenario after each constraint (columns) and iteration (rows).

6.3 Initial weights

It was found that initial weights had little impact on the overall simulation. Taking the simplest case, a random individual (id = 228) was selected from the ‘small area’ data and his/her initial weight was set to 2 (instead of the default of 1). The impact of this change on the individual’s simulated weight decayed rapidly, tending to zero after only 5 iterations in the small-area scenario. It was found that altered initial weights affect the simulated weights of all other individuals; this is illustrated in Fig. 4, which shows the impact on individual 228 and a randomly selected individual whose initial weight was not altered, for all areas.

Fig. 4 shows that changing initial weights for individuals has a limited impact on the results that tends to zero from one iteration to the next. After iteration one the impact of a 100% increase in the initial weight of individual 228 had reduced to 0.3%; after 3 iterations the impact was negligible. The same pattern can be observed in the control individual 570, although the impact is smaller and in the reverse direction. The same pattern can be observed for other individuals, by altering the ‘analysis.min’.

To analyse the impacts of changing the initial weights of multiple individuals, the same experiment was conducted, but this time the initial weights were applied to the first five individuals. The results are illustrated in Fig. 5, which shows the relationship between initial and final weights for the first two individuals and another two individuals (randomly selected to show the impact of altering weights on other

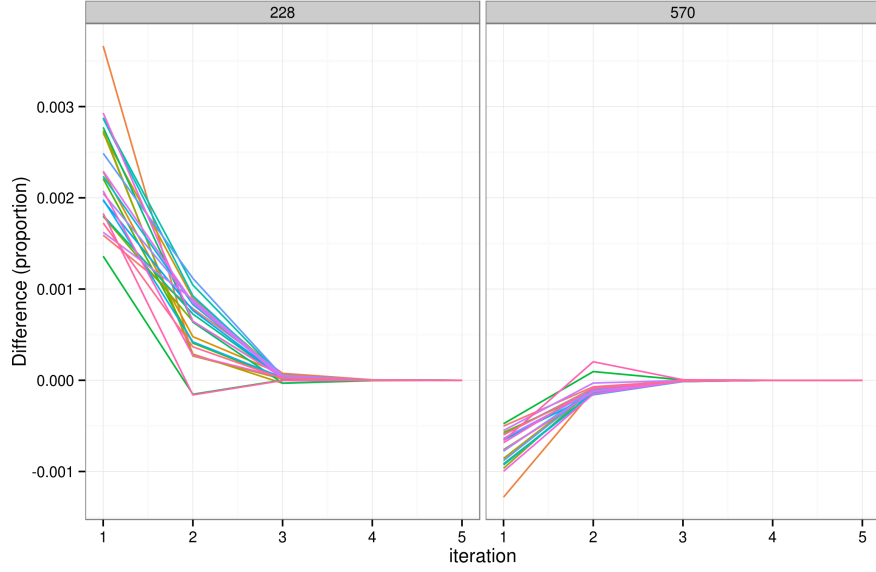


Figure 4: The impact of changing the initial weight of individual 228 on the simulated weights of two individuals after 5 iterations. Each of the 24 coloured lines represents a single zone.

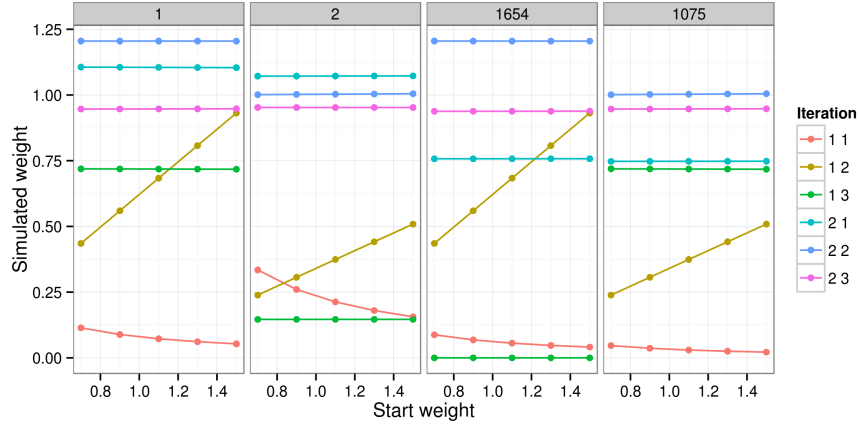


Figure 5: Initial vs simulated weight of four individuals for a single area (zone 1).

individual weights).

Fig. 5 shows that the vast majority of the impact of altered starting weights occurs in iteration-constraint combinations 1.1 and 1.2, before a complete iteration has even occurred. After that, the initial weights continue to have an influence, of a magnitude that cannot be seen. For individual 1, the impact of a 53% difference in initial weights (from 0.7 to 1.5) shrinks to 2.0% by iteration 1.3 and to 1.1% by iteration 2.3.

These results suggest that initial weights have a limited impact on the results of IPF, that declines rapidly with each iteration.

6.4 Empty cells

There were found to be 0, 62 and 8,105 empty cells in the simple, small-area and sheffield examples respectively. Only the smallest input dataset was *complete*. Due to the multiplying effect of more variables (Eq. (4)), N_{perm} was found to be much larger in the complex models, with values rising from 4 to 300 and 9504 for the increasingly complex test examples.

As expected, it was found that adding additional empty cells resulted in worse model fit, but only slightly worse fit if few empty cells (1 additional empty cell to 10% non-empty cells) were added. The addition of new rows of data with attributes corresponding to the empty cells to the input dataset allowed the small-area

and sheffield models to be re-run without empty cells. This resulted in substantial improvements in model fit that because relatively more important with each iteration: model runs with large numbers of empty cells reached minimum error values quickly, beyond which there was no improvement.

6.5 Integerisation

After passing the the final weights through the integerisation algorithms mentioned in Section 5.5, the weight matrices were of identical dimension and total population as before. It was found that, compared with the baseline scenarios, that integerisation had a negative impact on model fit.

For the ‘simple’ and ‘small area’ scenarios, the impact of integerisation was proportionally large, greater than the impact of any other changes. For the more complex Sheffield example, by contrast, the impact was moderate, increasing RMSE by around 20%.

6.6 Summary of results

The impact of each model experiment on each of the three scenarios is summarised in Table 11. It is notable that the variable with the greatest impact on model fit varied from one scenario to the next, implying that the optimisation of IPF for spatial microsimulation is context-specific. Changes that had the greatest positive impact on model fit were additional iterations for the simpler scenarios and the removal of empty cells from the input microdata for the more complex ‘Sheffield’ scenario.

Likewise, the greatest negative impact on model fit was due to the same variable for the simpler scenarios: integerisation. For the ‘Sheffield’ example, the use of only one complete iteration instead of 3 or more had the most damaging impact, highlighting the importance of multiple iterations of IPF for more complex spatial microsimulation models.

Table 11: Summary results of root mean square error (RMSE) between simulated and constraint data.

| Model test | Simple | Small area | Sheffield |
|-----------------------------------|-----------|------------|-----------|
| Baseline | 0.0001 | 0.018 | 25.5 |
| Constraints and iterations | | | |
| 1 iteration | 0.221 | 1.91 | 78.2 |
| 5 iterations | <0.000001 | 0.00001 | 20.9 |
| 10 iterations | <0.000001 | 0.000002 | 13.8 |
| Reordered constraints | 0.0001 | 0.018 | 14.5 |
| Initial weights | | | |
| 10% sample set to 10 | 0.0001 | 0.034 | 25.5 |
| 10% sample set to 1000 | 0.0001 | 0.998 | 25.5 |
| Empty cells | | | |
| 1 additional empty cell | 0.852 | 0.018 | 25.5 |
| 10% additional empty cells | — | 0.050 | 25.6 |
| No empty cells | — | 0.005 | 1.24 |
| Integerisation | | | |
| TRS integerisation | 0.577 | 3.68 | 29.9 |
| Proportional probabilities | 1.29 | 3.91 | 30.0 |

7 Conclusions

In this paper we have systematically reviewed and tested modifications to the design of the IPF algorithm to generate spatial microdata. This is the first time to our knowledge that such tests have been performed in a consistent and reproducible manner, that will allow others in to community to cross-check the results and enable further model testing. The model experiments conducted involved changing one of the following variables at a time:

- Order of constraints and number of iterations.

- Modification of initial weights.
- Addition and removal of ‘empty cells’.
- Interisation of the fractional weights using a couple of probabilistic methods.

Some of the variables altered

In every case, the direction of change in model fit was as expected, confirming prior expectations about the influence of model set-up on performance. The finding that the initial weights have very little influence on the final result, particularly after multiple iterations of IPF, is new and relevant to the spatial microsimulation community. This finding all but precludes the modification of initial weights as a method for altering IPF’s performance and suggests that research time would be best focused on other variables in the model run, notably that of empty cells.

The problem of empty cells has received little attention in the literature and this paper presents both the theory and methodology demonstrating why they are important (Section 5.4). We have demonstrated a method for identifying whether empty cells are present in the input data. More important for the spatial microsimulation community, we have shown for the first time (to our knowledge) that empty cells have a large adverse effect on model fit.

We conclude that the removal of empty cells can greatly improve model fit, especially for complex scenarios, where it can have the single greatest positive impact on model fit out of the variables modified in this paper. The identification and removal of empty cells, as demonstrated in this paper and accompanying code, can substantially enhance the performance of spatial microsimulation models yet appears to be rarely used. On potential downside of removing empty cells is that it involves adding fictitious new individuals to the input microdata. This could potentially cause problems if the purpose of spatial microsimulation is small area estimation, where the value of target variables (e.g. income) must be set, based on imperfect information. Another potential problem caused by removal of empty cells is that it could lead to the creation of impossible combinations of variables such as car drivers under the age of 16, if applied indiscriminately.

Despite these concerns, our finding that empty cells has a large impact on model results is robust across all scenarios and should be taken seriously in future research. A direct recommendation from these findings is therefore that researchers should test for missing cells, as defined in Section 5.4, before reweighting.

There are many opportunities for further research raised by this paper. There is clearly a need for further model testing in the area of spatial microsimulation, and IPF is only one of several options available. Research questions raised include: What are the impacts of additional variables affecting the performance of IPF? What are the interactions between different variables — e.g. what is the combined impact of altering the number of empty cells *and* number of iterations simultaneously? Can change in one variable reduce the need for change in another?

Moreover, there are broader questions raised by this research about validation of models that are used for policy making. Scarborough et al. (2009) highlight the paucity of evaluation studies of model based local estimates of heart disease, even when these estimates were being widely used by the medical profession for decision making. In this wider context our study has only scratched the surface, providing insight into the impact of a number of variables on the internal fit of one particular method for producing such small area estimates. Clearly there is much else to do. Based on the preceding literature review, it would seem that continued model testing and external validation are critical to the use of spatial microsimulation in academia and decision making based on imperfect information. This paper has provided new insights into the important factors affecting model fit using IPF. Its potentially more important wider contribution is to demonstrate the importance of systematic model testing and evaluation to ensure that spatial microsimulation models are serving the aims of researchers and society at large, and not vice versa.

References

- Bacharach, M., 1970. Biproportional matrices & input-output change. volume 16. Cambridge University Press, Cambridge.
- Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., Rossiter, D., 2005. SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place* 11, 13–34.
- 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.
- Ballas, D., Clarke, G.P., Dewhurst, J., 2006. Modelling the Socio-economic Impacts of Major Job Loss or Gain at the Local Level: a Spatial Microsimulation Framework. *Spatial Economic Analysis* 1, 127–146.
- Batty, M., 1976. Urban modelling; algorithms, calibrations, predictions. Cambridge University Press.
- Birkin, M., Clarke, M., 1989. The generation of individual and household incomes at the small area level using synthesis. *Regional Studies* 23, 37–41.
- 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.
- Clarke, G., Madden, M., 2001. *Regional science in business*. Springer.
- Clarke, M., Holm, E., 1987. Microsimulation methods in spatial analysis and planning. *Geografiska Annaler. Series B. Human Geography* 69, 145–164.
- Cleave, N., Brown, P.J., Payne, C.D., 1995. Evaluation of Methods for Ecological Inference. *Journal of the Royal Statistical Society. Series A* 158, pp. 55–72.
- Dale, A., Marsh, C., 1993. *The 1991 census users guide*.
- Deming, W., Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*.
- Edwards, K.L., Clarke, G.P., 2009. The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity. *Social science & medicine* 69, 1127–34.
- Edwards, K.L., Clarke, G.P., Thomas, J., Forman, D., 2010. Internal and External Validation of Spatial Microsimulation Models: Small Area Estimates of Adult Obesity. *Applied Spatial Analysis and Policy* 4, 281–300.
- Goedemé, T., den Bosch, K., Salanauskaite, L., Verbist, G., 2013. Testing the Statistical Significance of Microsimulation Results: Often Easier than You Think. A Technical Note. Technical Report. Herman Deleeck Centre for Social Policy, University of Antwerp.
- Goldrei, D.C., 1996. *Classic Set Theory: For Guided Independent Study*. CRC Press, London.
- Harding, A., Vidyattama, Y., Tanton, R., 2011. Demographic change and the needs-based planning of government services: projecting small area populations using spatial microsimulation. *Journal of Population Research* 28, 203–224.
- Harland, K., Heppenstall, A., Smith, D., Birkin, M., 2012. Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *Journal of Artificial Societies and Social Simulation* 15, 1.
- 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.
- Johnston, R.J., Pattie, C.J., 1993. Entropy-Maximizing and the Iterative Proportional Fitting Procedure. *The Professional Geographer* 45, 317–322.
- Jordan, R., Birkin, M., Evans, A., 2011. Agent-based simulation modelling of housing choice and urban regeneration policy, in: *Multi-agent-based simulation XI*. Springer, pp. 152–166.
- Kavroudakis, D., Ballas, D., Birkin, M., 2012. Using Spatial Microsimulation to Model Social and Spatial Inequalities in Educational Attainment. *Applied Spatial Analysis and Policy* 6, 1–23.
- Lahr, M., de Mesnard, L., 2004. Biproportional Techniques in Input-Output Analysis: Table Updating and Structural Analysis. *Economic Systems Research* 16, 115–134.
- Lee, A., 2009. *Generating Synthetic Microdata from Published Marginal Tables and Confidentialised Files*. Technical Report. Statistics New Zealand. Wellington.

- Legates, D.R., McCabe, G.J., 1999. Evaluating the use of “goodness-of-fit” Measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35, 233–241.
- Lovelace, R., 2013. Supplementary information: a user manual for the integerisation of IPF weights using R , 1–18.
- Lovelace, R., 2014. Introducing spatial microsimulation with R: a practical. *National Centre for Research Methods Working Papers* 08.
- Lovelace, R., Ballas, D., 2013. ‘Truncate, replicate, sample’: A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems* 41, 1–11.
- 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.
- Malleson, N., Birkin, M., 2012. Analysis of crime patterns through the integration of an agent-based model and a population microsimulation. *Computers, Environment and Urban Systems* 36, 551–561.
- Middleton, E., 1995. Samples of anonymized Records. *Census Users’ Handbook* , 337–362.
- Morrissey, K., O’donoghue, C., Farrell, N., 2013. The Local Impact of the Marine Sector in Ireland: A Spatial Microsimulation Analysis. *Spatial Economic Analysis* 00, 1–20.
- Ngo, T., See, L., 2012. Calibration and Validation of Agent-Based Models of Land Cover Change, in: Heppenstall, A.J., Crooks, A.T., See, L.M., Batty, M. (Eds.), *Agent-Based Models of Geographical Systems SE - 10*. Springer Netherlands, pp. 181–197.
- Norman, P., 1999. Putting Iterative Proportional Fitting (IPF) on the Researcher’s Desk. Technical Report October. School of Geography, University of Leeds.
- Openshaw, S., 1984. Ecological fallacies and the analysis of areal census data. *Environment & planning A* 16, 17–31.
- Ormerod, P., Rosewell, B., 2009. Validation and verification of agent-based models in the social sciences, in: *Epistemological Aspects of Computer Simulation in the Social Sciences*. Springer, pp. 130–140.
- 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.
- Rich, J., Mulalic, I., 2012. Generating synthetic baseline populations from register data. *Transportation Research Part A: Policy and Practice* 46, 467–479.
- Scarborough, P., Allender, S., Rayner, M., Goldacre, M., 2009. Validation of model-based estimates (synthetic estimates) of the prevalence of risk factors for coronary heart disease for wards in England. *Health & place* 15, 596–605.
- Simpson, L., Tranmer, M., 2005. Combining sample and census data in small area estimates: iterative proportional fitting with standard software. *The Professional Geographer* 57, 222–234.
- Smith, D., Clarke, G., Harland, K., 2007. SimHealth: Estimating Small Area Populations Using Deterministic Spatial Microsimulation in Leeds and Bradford. *University of Leeds Working Papers* .
- Smith, D.M., Clarke, G.P., Harland, K., 2009. Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A* 41, 1251–1268.
- Sutherland, H., Figari, F., 2013. EUROMOD: the European Union tax-benefit microsimulation model. *International Journal of Microsimulation* 6, 4–26.
- Tanton, R., Edwards, K., 2013. *Spatial Microsimulation: A Reference Guide for Users*. Springer.
- Teh, Y.W., Welling, M., 2003. On improving the efficiency of the iterative proportional fitting procedure, in: *Proceedings of the International Workshop on Artificial Intelligence and Statistics*.
- 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.

- Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin, D., Steel, D., Gardiner, C., 2005. The Case for Small Area Microdata. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 168, pp. 29–49.
- Van Buuren, S., 2012. Flexible imputation of missing data. CRC press.
- Van Leeuwen, E.S., 2010. The effects of future retail developments on the local economy: Combining micro and macro approaches. *Papers in Regional Science* 89, 691–710.
- Voas, D., Williamson, P., 2001. Evaluating Goodness-of-Fit Measures for Synthetic Microdata. *Geographical and Environmental Modelling* 5, 177–200.
- Whitworth, A., 2012. Local Inequality and Crime: Exploring how Variation in the Scale of Inequality Measures Affects Relationships between Inequality and Crime. *Urban Studies* .
- Whitworth, A., 2013. Evaluations and improvements in small area estimation methodologies .
- Williamson, P., 2007. CO Instruction Manual: Working Paper 2007/1 (v. 07.06.25). Technical Report June. University of Liverpool.
- Williamson, P., 2013. An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimisation, in: Tanton, R., Edwards, K. (Eds.), *Spatial Microsimulation: A Reference Guide for Users SE - 3*. Springer Netherlands. volume 6 of *Understanding Population Trends and Processes*, pp. 19–47.
- Williamson, P., Birkin, M., Rees, P.H., 1998. The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A* 30, 785–816.
- Willmott, C., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.
- Wu, B.M., Birkin, M.H., Rees, P.H., 2010. A Dynamic MSM With Agent Elements for Spatial Demographic Forecasting. *Social Science Computer Review* 29, 145–160.
- Zuo, C., Birkin, M., Clarke, G., McEvoy, F., Bloodworth, A., 2013. Modelling the transportation of primary aggregates in England and Wales: Exploring initiatives to reduce CO2 emissions. *Land Use Policy* 34, 112–124.