

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

Robin Lovelace

February 20, 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 its 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 using various metrics of for determining its performance. 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 Notes for authors

Additonal papers to cite include:

- Harland et al. (2012)
- Smith et al. (2009)
- Cullinan (2010)
- 

## 2 Introduction

The aggregation of survey data has a long history in the social sciences. Generally this involves translating the raw data from a ‘long’ format (where each row represents an individual and each column a separate variable) to a ‘wide’ format (where each row represents a geographic area or other sub-group of the total population and each column represents a variable value). This process of aggregation is performed in three steps: a) 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), b) assigning a column to each category and c) assigning integer counts to cells in each column. Frequently large national social survey datasets such as censuses are *only* available in the wide, geographically aggregated form, which is problematic for the analysis of spatial phenomena (Openshaw, 1984; Lee, 2009).

Iterative proportional fitting (IPF) can help overcome these issues when used as a basis for *spatial microsimulation*, which can be understood in simplistic terms as the reversal of geographic aggregation to approximate the original ‘long’ dataset (ref). As illustrated in Section 2, the need for spatial microsimulation is driven by the tendency of statistical agencies to make the raw georeferenced microdata generated through large social surveys unavailable to the research community: spatial microsimulation is also known as

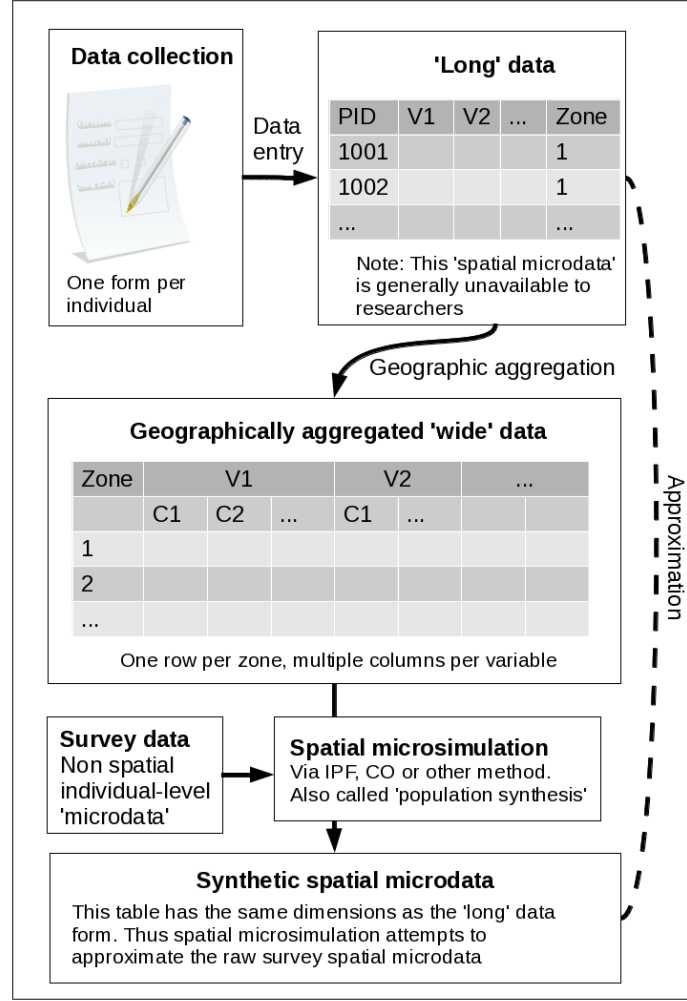


Figure 1: Schema of iterative proportional fitting (IPF) and combinatorial optimisation in the wider context of the availability of different data formats and spatial microsimulation.

*population synthesis*, especially in transport modelling, which conveys its core function: to generate — and then, in most cases, analyse — long spatial microdata that approximates the raw spatial microdata collected during surveys.

IPF is a statistical procedure that enables the calculation of *maximum likelihood* of values that are constrained by multiple variables. IPF is a general purpose tool with multiple uses. In this paper we focus on its ability to *disaggregate* the geographically aggregated data described above, resulting in the generation of synthetic *spatial microdata* — a long table of individual-level data whereby each individual is also allocated to a geographic zone.

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 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 contained in the variables to be cross tabulated.<sup>1</sup> 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 degree, are female and over 25 years old? This question is impossible to answer with a geographically aggregated (flat) dataset.

## 2.1 IPF in spatial microsimulation

Spatial microsimulation tackles the aforementioned problems associated with the prevalent geographically aggregated flat data form 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 associated attributes and weights for each zone (see Table 7 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 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

<sup>1</sup>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.

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...

## 2.2 Applications

**Dimitris and Eveline:** please fill in this section Applications of IPF include using it ...

The application of greatest interest to regional scientists is as a method of spatial microsimulation: generating individual level data allocated to specific zones... Papers that have used IPF in this context (or a substitutable method such as simulated annealing) have explored a range of regional economic problems that purely aggregate level analyses struggle to address. These include:

- using spatial microdata to analyses the wider impacts of major job losses at the local level (Ballas et al., 2006).
- an investigation into to local impacts of the marine sector in Ireland (Morrissey et al., 2013)
- ...
- ...

## 3 IPF: theory and a worked example

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...

### 3.1 IPF theory: a worked example

In most modelling texts there is a strong precedence of theory over application: the latter usually flows from the former. The location of this section after a description of the programming language R is therefore a little unconventional but there is a logic to this order. Having demonstrated the power and flexibility of the programming language in which the model is written, the next stage is to analyse the task to which it is to be set. More importantly for reproducible research, this theory section is illustrated with a simple worked example that culminates in a question to the reader, to test his or her understanding.

IPF is a simple statistical procedure, “in which cell counts in a contingency table containing the sample observations are scaled to be consistent with various externally given population marginals” (McFadden et al., 2006). In other words, and in the context of *spatial* microsimulation, IPF produces maximum likelihood estimates for the frequency with which people appear in different areas. The method is also known as ‘matrix raking’ or the RAS algorithm, and has been described as one particular instance of a more general procedure of ‘entropy maximisation’ (Johnston and Pattie, 1993; Blien and Graef, 1998). The mathematical properties of IPF have been described in several papers (Bishop et al., 1975; Fienberg, 1970; Birkin and Clarke, 1988). Illustrative examples of the procedure can be found in Saito (1992), Wong (1992) and Norman (1999). Wong (1992) investigated the reliability of IPF and evaluated the importance of different factors influencing the its performance. Similar methodologies have since been employed by Mitchell et al. (2000), Williamson et al. (2002) and Ballas et al. (2005) to investigate a wide range of phenomena.

To illustrate how IPF works in practice, a simplified example is described below. This is a modified version of a simpler demonstration from Ballas et al. (2005).<sup>2</sup> Table 1 describes a hypothetical microdataset

---

<sup>2</sup>In Ballas et al. (2005) the interaction between the age and sex constraints are assumed to be known. (Their equivalent of Table 3 contains data for every cell, not question marks.) This results in IPF converging instantly. However, in Census data, such cross-tabulation is often absent, and IPF must converge over multiple constraints and iterations. This latter scenario is assumed in the worked example below. Other worked examples of the principles are provided in Johnston (1985, Appendix 3) (for entropy maximisation), Norman (1999) and Simpson and Tranmer (2005) (using the proprietary statistical software SPSS).

comprising 5 individuals, who are defined by two constraint variables, age and sex. Each has two categories. Table 2 contains aggregated data for a hypothetical area, as it would be downloaded from census dissemination portal Casweb. Table 3 illustrates this table in a different form, which shows our ignorance of interaction between age and sex.

Table 1: A hypothetical input microdata set (the original weights set to one). The bold value is used subsequently for illustrative purposes.

Individual	Sex	Age-group	Weight
1	Male	Over-50	1
2	Male	Over-50	1
3	Male	Under-50	<b>1</b>
4	Female	Over-50	1
5	Female	Under-50	1

Table 2: 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 3: 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 4 presents the hypothetical microdata in aggregated form, that can be compared directly to Table 3.

Using these data it is possible to readjust the weights of the hypothetical individuals, so that their sum would add up to the totals given in Table 3 (12). In particular, the weights can be readjusted by multiplying them by the marginal totals, originally taken from Table 2 and then divided by the respective marginal total in 4. 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 2) and  $mT(n)_i$  is the marginal total of category  $j$  of the aggregated results of the weighted microdata,  $m$  (Table 4).  $n$  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 5 illustrates the weights that result. Notice that the sum of the weights is equal to the total population, from the constraint variables.

After the individual level data have been re-aggregated (Table 6), 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  $n$ :

Table 4: The aggregated results of the weighted microdata set ( $m(1)$ ). Note, these values depend on the weights allocated in Table 1 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	

Table 5: Reweighting the hypothetical microdataset in order to fit Table 2.

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	$1 \times 8/2 = 4$

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

To test your understanding of IPF, apply Eq. (2) to the information above and that presented in Table 6. This should result 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 4 and Table 6 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.

Table 6: The aggregated results of the weighted microdata set after constraining for age ( $m(2)$ ).

	Marginal totals	$i$			
	Age/sex	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	

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. As described in the introduction, this is not ideal for certain applications. Integer weights allow the results of spatial microsimulation to be further processed using dynamic microsimulation and agent based modelling techniques (Pritchard and Miller, 2012).

Spatial microsimulation can also provide insight into the likely distribution of individual level variables about which only geographically aggregated statistics have been made available. 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.

### 3.2 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 is implemented in R, using a slightly more complex example (Lovelace and Ballas, 2013).<sup>3</sup>

## 4 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. 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?” As we shall see below, ideally we would hope for a perfect fit between the input and output datasets during internal evaluation. *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.

This section outlines the evaluation techniques that are commonly used in the field, with a strong focus on internal validation. The options for external validation are heavily context dependent so must generally be decided on a case-by-case basis, and are mentioned in passing with a focus on general principles rather than specific advice. The methods are presented in ascending order of complexity and roughly descending order of frequency of use, ranging from Pearson’s coefficient of correlation to entropy based measures. Before these methods are described, it is worth stating the the purpose of this section 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. A final point is that 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 7), suggests that researchers need only to present one or two metrics (but should perhaps try more, for corroboration) to establish relative levels of goodness of fit between different model configurations.

### 4.1 Scatter plots and Pearson’s coefficient of correlation

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 (flat) 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.

Pearson’s coefficient of correlation (henceforth Pearson’s  $r$ ) 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  $r$  value of 1 represents a perfect fit; an  $r$  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 very high coefficient of determination for internal validation (the constraint) and strong positive correlation between target variables that have been estimated and external datasets (e.g. income).

### 4.2 Standard error around identity (SEI)

here SEI is the standard error around identity,  $y$  are the estimated values for each area,  $y_{rel}$  are the reliable estimates for each area from a census or other data source and  $\bar{y}_{rel}$  is the mean estimate for all areas where reliable data are available. This estimate has been used in validation by both Ballas and Tanton (see Ballas et al. 2007; Tanton et al. 2011).

---

<sup>3</sup>This tutorial is available from Rpubs, a site dedicated to publishing R analyses that are reproducible. It uses the RMarkdown mark-up language, which enables R code to be run and presented within documents. See <http://rpubs.com/RobinLovelace/5089>.

### 4.3 Total Absolute and Standardized Error (TAE and SAE)

The SAE addresses this issue by using the population size as the denominator, but generally, authors seem to use total population, rather than the population for that categorization of the variable, which may be deemed to understate the size of the errors. Also, this measure does not provide any information on whether any differences are statistically significant (Edwards and Tanton, 2013).

### 4.4 Z scores

Here, the evaluation mainly takes place at the cell level (for a single constraint in a single area). However, there should also be attention for the overall picture like spatial concentration.

## 5 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 on the internet. 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 a sample survey dataset of only five individuals, constrained by two variables (age and sex). 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 that fit on a single page (see Table 7 and Table 8). As with all data used in this paper, these tables can be viewed and downloaded online.

Table 7: Individual-level data for the simple scenario. The file, ‘vailable from the GitHub in the folder ‘input-data’

Id	Age	Sex
1	59	m
2	54	m
3	35	m
4	73	f
5	49	f

Table 8: Geographically aggregated constraint variables for the simple scenario, also available online

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 geographical unit in the UK), with an average population of 184 people aged over 16. 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 the hyperlinks embedded in digital versions of the previous sentence. A sample of these constraint variables is presented in Table 9 below.

The individual-level table that compliments these geographically aggregated count constraints is taken from Understanding Society wave 1. These have been randomized to ensure anonymity but they are still



Table 9: Sample of constraint variables from the first 5 rows of the small-area-eg input dataset.

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 to be viewed properly). As shown in Table 10 below, this consists of the same 3 constraint variables as provided in Table 9 but at aggregate level, in addition to one ‘target variable’: monthly pre-tax income in pounds.

Table 10: Randomly selected sample of 5 individuals taken from the individual-level dataset.

... More here on loading-up this data.

## 6 Method: model experiments

### 6.1 Baseline scenarios

In order to perform a wide range of model experiments, in which only one factor is altered at a time, it was necessary to clearly define baseline scenarios. These are model runs with sensible defaults, which can be referred back to as a reference point and re-used at any point in the future. The baseline ‘simple’ scenario consists of 3 iterations, leading to a total of 6 constraint-iteration permutations. The results of simplest baseline scenario are shown in Table 11. The results show that the model fit improves after every iteration and every constraint in every measure of model fit (except ‘p5’, which reaches zero in the second iteration) — *pcor* continues to improve, even though the printed value, to 6 decimal places, is indistinguishable from 1 after two complete iterations.

Table 11: Results for the baseline ‘simple’ scenario. Generated through by the ‘analysis’ script, and available as a .csv file online.

The baseline ‘small-area’ scenario consists of three iterations of IPF, in which all three constraints were used. The results for all 9 constraint-iteration permutations of this baseline scenario are illustrated in Fig. 2 and in the online ‘measures’ table. These results show that IPF successfully brings the simulated cell counts in-line with the census after each constraint is applied and that the fit improves with each iteration (only two iterations are shown here, as the fit improves very little beyond two, and to avoid clutter).

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 does not always improve from one constraint to the next (see the online ‘measures’ table).

### 6.2 Constraints and iterations

These were the simplest model experiments, requiring only the reporting of iterations that had already taken place, and re-arrangement of the IPF code such that constraints were applied in a different order than before. The re-ordered code for the ‘small-area’ scenario, for example can be seen in the script ‘etsim-reordered’ online.

### 6.3 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 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.

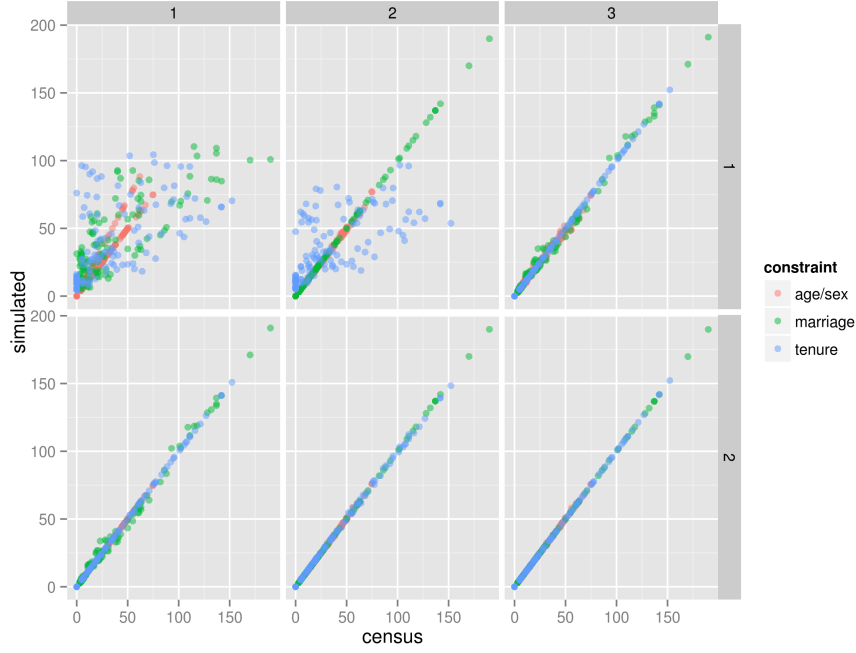


Figure 2: IPF in action: simulated vs census zone counts for the small-area baseline scenario after each constraint (columns) and iteration (rows).

## 6.4 Empty cells

Empty cells refer to cross-tabulations of 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 7. The method used to find-out if empty cells exist is to first select only individuals with unique characteristics (in the narrow terms of the constraint variables — of course every survey individual is unique). The column totals of the categorical (binary, 0 or 1) form of this dataset are compared: any categories with fewer than expected totals have missing cells.

Once the presence of empty cells is determined in the baseline scenarios, the next stage is to add or subtract empty cells by adding or subtracting individuals to the input data. After these changes were made, the baseline scenarios were run as before.

In the ‘simple’ scenario, this simply involves removing a single individual from the input data, which equates to deleting a single line of code (see ‘etsim-empty’). The question of which unique individual to remove is solved by random sampling.

## 6.5 Integerisation

This was the next variable that was altered, following (). 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 folder ‘small-area-weights’ (see ‘etsim.R’). The number of different initial weights 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 term defined as the sequence 1: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  $k <- u * 10^6$  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 7).

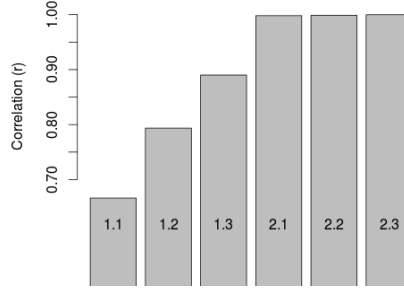


Figure 3: Improvement in goodness of fit with additional constraints and iterations, as illustrated by  $r$  values.

## 7 Results

### 7.1 Baseline scenarios

As expected from Fig. 2 above, the correlation of the baseline scenario improves over time, with each constraint and from one iteration to the next. Through the course of the first iteration, the correlation improves from 0.67 (the correlation between the frequencies of variables in the survey dataset and their frequencies in the census dataset, for each zone) to 0.9981. As illustrated in Fig. 3, the marginal improvement reduces from one iteration to the next: at the end of iteration 2, the correlation has improved to 0.999978.

### 7.2 Constraints and iterations

As illustrated above, the number of iterations required to reach a very close fit between the census and simulated data is very low, with rapid reductions in the marginal accuracy return on additional computing power. After four iterations, the result in R is indistinguishable from 1, with its default accuracy of 7 decimal places. The decay in the rate of improvement with additional iterations and constraints was found to super-exponential.

### 7.3 Initial weights

It was found that initial weights had little impact on the overall simulation. Taking the simplest case, a random individual ( $\text{id} = 228$ ) 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 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.

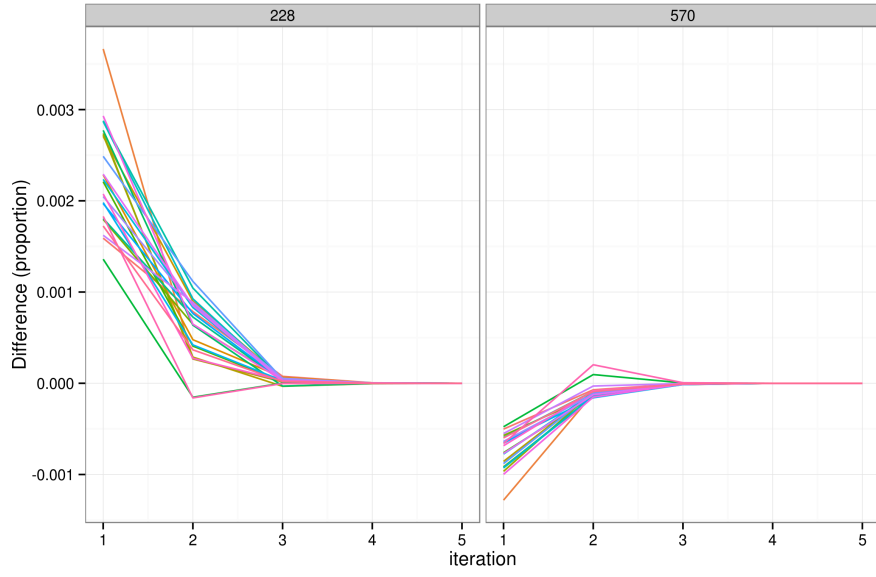


Figure 4: The impact of changing the initial weight of individual 228 on the simulated weights of two individuals after 5 iterations. Each of the 24 coloured lines represents a single zone.

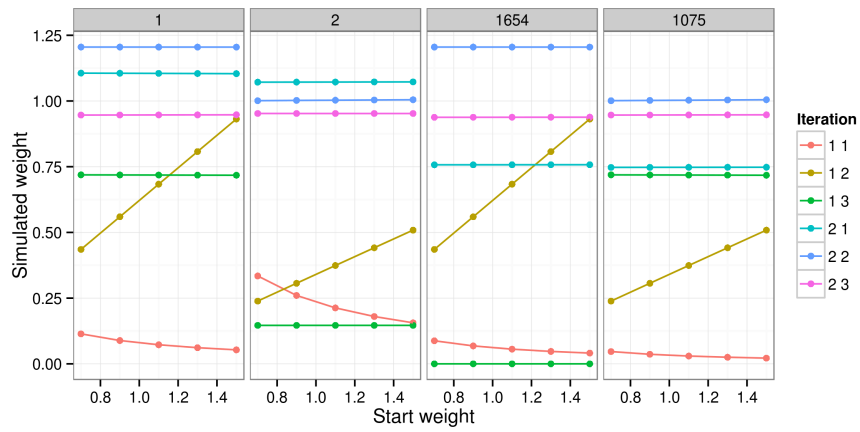


Figure 5: Initial vs simulated weight of four individuals for a single area (zone 1).

## 7.4 Empty cells

## 7.5 Integerisation

# 8 Conclusions

## References

- 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.
- 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.
- Bishop, Y., Fienberg, S., Holland, P., 1975. *Discrete Multivariate Analysis: theory and practice*. MIT Press, Cambridge, Massachusettes.
- Blien, U., Graef, F., 1998. Entropy optimizing methods for the estimation of tables, in: *Classification, Data Analysis, and Data Highways*. Springer, pp. 3–15.
- 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.
- Cullinan, J., 2010. Developing a Continuous Space Representation of a Simulated Population. *Spatial Economic Analysis* 5, 317–338.
- 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.
- Fienberg, S., 1970. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics* 41, 907–917.
- 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., 1985. *The geography of English politics: the 1983 general election*. Croom Helm London.
- Johnston, R.J., Pattie, C.J., 1993. Entropy-Maximizing and the Iterative Proportional Fitting Procedure. *The Professional Geographer* 45, 317–322.
- Lee, A., 2009. *Generating Synthetic Microdata from Published Marginal Tables and Confidentialised Files*. Technical Report. Statistics New Zealand. Wellington.
- 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.

- McFadden, D., Heiss, F., Jun, B.h., Winter, J., 2006. On testing for independence in weighted contingency tables. *Medium for Econometric Applications* 14, 11–18.
- Mitchell, R., Shaw, M., Dorling, D., 2000. *Inequalities in Life and Death: What If Britain Were More Equal?* Policy Press, London.
- 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.
- Mosteller, F., 1968. Association and estimation in contingency tables. *Journal of the American Statistical Association* 63, 1–28.
- 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.
- 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.
- Saito, S., 1992. A multistep iterative proportional fitting procedure to estimate cohortwise interregional migration tables where only inconsistent marginals are known. *Environment and Planning A* 24, 1531–1547.
- 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.M., Clarke, G.P., Harland, K., 2009. Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A* 41, 1251–1268.
- 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.
- Williamson, P., Mitchell, G., McDonald, A.T., 2002. Domestic Water Demand Forecasting: A Static Microsimulation Approach. *Water and Environment Journal* 16, 243–248.
- Wong, D.W.S., 1992. The Reliability of Using the Iterative Proportional Fitting Procedure. *The Professional Geographer* 44, 340–348.