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

NA

December 9, 2014

Abstract

Iterative Proportional Fitting (IPF), also known as biproportional fitting, 'raking' or the RAS algorithm, is an established procedure used in a variety of applications across the social sciences. Primary amongst these for urban modelling has been its use in static spatial microsimulation to generate small area microdata — individual level data allocated to administrative zones. The technique is mature, widely used and relatively straight-forward. Although IPF is well described mathematically, ~~reproducible~~ accessible examples of the algorithm ~~,~~ written in modern programming languages ~~, are rarein the academic literature. Therefore, there~~ are rare. There is a tendency for researchers to 'start from scratch', resulting in a variety of ad hoc implementations and little evidence about the relative merits of differing approaches. These knowledge gaps mean that answers to methodological questions must be guessed: What impact can initial weights have on the final result? How can 'empty cells' be identified and how do they influence model fit? Can IPF be made more computationally efficient? This paper tackles such questions ~~,~~ using a systematic methodology based on publicly available code and data. The results demonstrate the sensitivity of the results to initial conditions, notably ~~to~~ the presence of 'empty cells', and the ~~importance of software choices for IPF's~~ dramatic impact of software decisions on computational efficiency. The paper concludes by proposing an agenda for robust and ~~reproducible future tests~~ transparent future tests in the field. ~~for IPF and other methodologies used in spatial microsimulation.~~

Keywords: Iterative proportional fitting, spatial microsimulation, modelling

1 Introduction

Surveys of the ~~composition~~characteristics, attitudes and behaviour of individual ~~respondents constitute a ubiquitous empirical basis of~~ people or households constitute ubiquitous empirical bases for 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~~ format to a 'wide' data format (?) . The impacts of this process of aggregation should not be underestimated: a large proportion of administrative data that is available to the public and under academic license is provided in this form, with implications for the types of analyses that can be conducted. Notably for modellers, it is difficult to build an agent-based model from geographically aggregated count data.

Long data is often described as 'microdata' in the spatial microsimulation literature. Long data is analogous to the ideal of 'tidy data' advocated by ? , who has created software for converting between the two data forms. Long data can be regarded as being higher in fidelity than long data and is closer to the form in which survey information is entered into a database, one individual at a time ~~— to a 'wide' data format (?) .~~ (??).

In wide or 'areal' datasets, by contrast, each row represents a contiguous and mutually exclusive geographic area containing a subset of the total population~~,~~. In wide data each column represents a variable value or bin (?; see ??). The transformation from 'long' to 'wide' is typically carried out for the following reasons: to preserve anonymity of participants (??); to ease storage and retrieval of information — aggregate-level datasets can represent information about millions of individuals in relatively little computer disk space; and to ease the loading, visualisation and analysis of official data in geographical information systems (GIS).

The process of aggregation is generally performed in three steps:

1. Conversion of variables into discrete categories (e.g. a person aged 21 could be allocated to the age band 20–25).

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.

However, the process of aggregation has some important disadvantages. Geographically aggregated datasets are problematic for the analysis of ~~spatial~~ multi-level phenomena (?) and can prevent the analysis of ~~the~~ interdependencies between individual and ~~household variables at the local level~~ higher-level variables across space (?). In an effort to address the drawbacks associated with this lack of microdata, statistical authorities in many countries have made available samples of individual-level data, usually with carefully designed measures to prevent identification of individuals. ? provide an insightful discussion of this subject from the perspective of the publication of ‘public use microdata samples’ ~~PUMS~~ (PUMS) in the USA, where the “benefits of data dissemination are potentially enormous” (p. 1)~~but where~~. Yet there is little consensus about the right balance ~~right ‘tolerance to risk’ of individuals being identified~~between the costs and benefits of making more detailed data available. In the UK, to take another example, a 2% Sample of Anonymised Records (~~SARs~~SAR) taken from the 1991 Census was released to academic researchers (??). This has been followed by similar ~~SARs~~ SAR releases in 2001 and 2011. Due to increased ability to to store, process and transfer such ‘big microdata’, alongside increasing awareness of its potential benefits for public policy making, ~~? suggests that researchers “will soon have~~ the availability of such data is rapidly increasing. “[~~access to~~I]~~individual-level~~ ndividual-level information about entire populations or large samples covering most of the world~~’~~’s population with multiple observations at high geographic resolution” ~~(p. 297). Indeed~~will soon become available, enthuses ?, p. 297 . Furthermore, the mass collection and analysis of individual-level information from mobile phones and other sources is already a reality in some sectors.

Nevertheless, most official microdatasets ~~,~~ such as the ~~SARs and the PUMS ,~~ SAR and PUMS are not suitable for ~~local and neighbourhood level~~ spatial analysis analysis due to the low level of geographic resolution~~allocated to each individual to preserve confidentiality~~. There has been discussion of ways to further disaggregate the geography of such microdata, with due emphasis on risks to confidentiality (?). ~~Furthermore, there are many~~ This is linked with calls to collect more sensitive policy relevant variables (e.g. income and wealth, social attitudes) ~~that are not collected by the Census at all in many countries, making these variables unavailable at the small area level and increasing the incentives to produce local level estimates.~~

often ignored in national surveys. To summarise the motivations for spatial microsimulation in relation to data accessibility, there is an absence of rich geocoded individual-level data in most countries. ~~One way geographers and regional scientists have addressed these data issues and lack of small area information is through the development and application of the methods of small area estimation and spatial microsimulation (?) . There is much overlap between these two methods and the terms are sometimes used interchangeably. It is useful to make the distinction, however, as~~ Spatial microsimulation and ‘small area estimation~~refers to methods to estimate summary statistics for geographical units — sometimes via a spatial microdataset. With spatial microsimulation, on the other hand, the emphasis is on generating and analysing spatial microdata.~~’ were originally developed in the field of applied geography to address these issues (???) .[1]

Spatial microsimulation combines ~~national~~ social survey microdata (such as the Understanding Society panel dataset UK) with small area census data to synthesis geographically specific populations. A number of techniques are available — see ? for a recent review — including deterministic reweighting (?), genetic algorithms and simulated annealing (?) as well as approaches based on linear regression (?). ~~It~~ Methodological advance in spatial microsimulation is an active field of development with ongoing refinements (e.g. ??) and debates about which approaches are most suited in different situations (????)

~~These methods are important because large national social survey datasets are often *only* available in the wide, geographically aggregated form. Information~~ Insufficient data collection for research needs is not the only motivation for spatial microsimulation. Even when the ideal variables for a particular study *are* collected locally, information loss occurs when ~~georeferenced individual level survey data — spatial microdata — is aggregatedin this way~~they are are geographically aggregated. This problem is exacerbated when continuous variables are converted into discrete categories and ~~when~~ cross-tabulations, such as age/income, are not provided~~, as~~. Unfortunately this is often the case.

~~One~~ Within this wider context, Iterative Proportional Fitting (IPF) is important as one of the key techniques used in spatial microsimulation~~and small area estimation is Iterative proportional fitting (IPF)~~. 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 (??). Thus

[1] There is much overlap between spatial microsimulation and small area estimationand the terms are sometimes used interchangeably. It is useful to make the distinction, however, as small area estimation refers to methods to estimate summary statistics for geographical units — sometimes via a spatial microdataset. With spatial microsimulation, on the other hand, the emphasis is on generating and analysing spatial microdata.

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.

the need for spatial microsimulation is ~~driven by the tendency of statistical agenciesto make the~~ largely driven by statistical agencies' reluctance to 'open-up' raw georeferenced microdata ~~generated by large social surveys unavailable to the research community (?) . In fact spatial microsimulation , the term favoured by geographers and regional scientists, is generally~~ (?) .

Reflecting its role in overcoming these data limitations, spatial microsimulation is 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~~ . Population synthesis captures the central role of spatial microsimulation, to simulate geographically specific populations based. Some ambiguity surrounds the term spatial microsimulation as it can also refer ~~to~~ a wider approach ~~which harnesses spatial microdata in the analysis stage~~ based on spatial microdata (??). In this paper we focus on the former meaning, which is also known as *static spatial microsimulation* (?).

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 (?). 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 (?). 'Entropy maximisation' is a related technique, of which IPF is considered to be a subset (??). Usage characteristics of these different terms are summarised in ??.

Table 1: Usage characteristics of different terms for the IPF procedure, from the Scopus database. The percentages refer to the proportion of each search term used to refer to IPF in different subsets of the database: articles published since 2010 and mentioning "regional", respectively.

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

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" (?). The original formulation of the procedure is generally ascribed to ?, 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 ?, 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. ?) 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 ??.

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 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 are interested in intra-zone variability at the local level (**??**).

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. Often, researchers investigating spatial phenomena at many scales must use two datasets: one long non-spatial survey table where each row represents the characteristics of an individual — such as age, sex, employment status and income — and another wide, geographically aggregated dataset. 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 **??** 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 three 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. Synthetic spatial microdata taken from a single dataset can therefore be stored in a more compact way: a 'weight matrix' is a matrix with rows representing individuals and columns geographic zones. The internal cells reflect how reflective each individual is of each zone.

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 (**?**). 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 studies on:

- The individual and local impacts of job losses and gains from the closure (**?**) or opening (**?**) of new businesses.
- Inequalities between individuals and zones in educational opportunities and attainment (**?**).
- Commuting patterns and the distributional impacts of new transport policies (**?**).
- The spatial behavior of farmers and rural residents (**??**).
- Investigation of the local impacts of economic development in the marine energy sector (**?**).
- The socio-spatial impacts of changes in taxation policy.
- The optimal location of health services based on estimates of individual level behaviour and obesity rates (**??**).

- The use of spatial microsimulation in combination with agent based models to simulated student populations (?), estimate the energy impacts of freight transport (?) and to estimate the impacts of urban regeneration (?).

~~Various~~ IPF is used in the microsimulation literature to reweight individual-level data to fit a set of aggregate-level constraints. A clear and concise definition is provided by ? in a *Reference Guide* (?) : ”the method allocates all households from the sample survey to each small area and then, for each small area, reweights each household” (p. 53). ? also provides a simple explanation of the mathematics underlying the method. Issues to consider when performing IPF include selection of appropriate constraint variables, number of iterations and ‘internal’ and ‘external’ validation (???) . It has been noted that IPF is an alternative to ’combinational optimisation (CO) and ’generalised regression weighting’ (GREGWT) (?) . One issue with the IPF algorithm is that unless the fractional weights it generates *intergerised*, the results cannot be used as an input for agent-based models. ? provide an overview of methods for integerisation to overcome this problem. To further improve the utility of IPF there has been methodological work to convert the output into household units (?) .

In addition to IPF, other strategies for generating ~~the~~ synthetic spatial microdata~~used in such studies are available, as~~. These are described in recent overviews ~~on~~ of the subject (??). However, it is not always clear which is most appropriate in different situations. There has been work testing and improving model results (?) ~~, yet research focused on~~ but very little work systematically and transparently testing the ~~underlying algorithms for generating synthetic spatial microdata is limited (?) . This research gap ,~~ algorithms underlying population synthesis (?) .

The research gap addressed in this paper ~~,~~ is highlighted in the conclusions of a recent Reference Guide to spatial microsimulation: “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. ?). As outlined in ??, however, the theory underlying IPF has 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. ? and ? 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. ? also represents good practice, with a manual for running combinatorial optimisation algorithms described in previous work. ? 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” (?). This section builds on the tutorial by providing context and a higher level description of the process.

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” (?) . 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’ (??) .

The mathematical properties of IPF have been described in several papers (???). Illustrative examples of the procedure can be found in ? , ? and ? . ? investigated the reliability of IPF and evaluated the importance of different factors influencing the its performance. Similar methodologies have since been employed by ? , ? and Ballas et al. (?) to investigate a wide range of phenomena.

We begin with some hypothetical data. Table ?? describes a hypothetical microdataset comprising 5 individuals, who are defined by two constraint variables, age and sex. Table ?? contains aggregated data for a hypothetical area, as it might be downloaded from census dissemination portals such as Casweb.[2] ?? illustrates this table in a different form, demonstrating the unknown links between age and sex. The job

[2] Casweb (casweb.mimas.ac.uk/) is an online portal for academics and other researchers to gain access to the UK’s official census data in geographically aggregated form.

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 ⇒	i		j	
Category ⇒	i_1	i_2	j_1	j_2
Area ⇓	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 **??** presents the hypothetical microdata in aggregated form, that can be compared directly to Table **??**. Note that the integer variable age has been converted into a categorical variable in this step, a key stage in geographical aggregation. 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 **??** (12). In particular, the weights can be readjusted by multiplying them by the marginal totals, originally taken from Table **??** and then divided by the respective marginal total in **??**. Because the total for each small-area constraint is 12, this must be done one constraint at a time. This is expressed, for a given area and a given constraint (i, age in this case) in **??**.

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

In **??**, $w(n+1)_{ij}$ 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 **??**) and $mT(n)_i$ is the marginal total of category j of the aggregated results of the weighted microdata, m (Table **??**). 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 **??** to see how the new weight of individual 3 is calculated for the sex constraint. Table **??** illustrates the weights that result. Notice that the sum of the weights is equal to the total population, from the constraint variables.

Table 5: Aggregated results of the weighted microdata set ($m(1)$). Note, these values depend on the weights allocated in Table **??** 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 6: Reweighting the hypothetical microdataset in order to fit Table **??**.

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	$\mathbf{1} \times \mathbf{8/2} = 4$
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 (**??**), the next stage is to repeat **??** 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:

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

Apply **??** to the information above and that presented in **??** results in the following vector of new weights, for individuals 1 to 5:

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

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, described in the next section) improves in $m(1)$ to $m(2)$ from 14 to 6 in **??** and **??** 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~~ ‘empty cells’ (?) in the input microdataset, a defining feature of IPF (see **??** for more on the impact of empty cells).

Table 7: 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) (**?**). An issue with the results of IPF 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 'integerise' the fractional weights illustrated in **??**. Methods for performing integerisation are discussed in (**?**), (**?**) and, most recently in (**?**), who systematically benchmarked and compared different procedures. It was found that the probabilistic ‘truncate, replicate, sample’ (TRS) and ‘proportional probabilities’ methods — which use repeat sampling on the raw weights and non-repeat sampling (whereby a selected individual cannot be selected again) of the truncated weights, respectively — were most accurate. These methods are therefore used to test the impact of integerisation in this paper.

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 (**?**). *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 (**?**).

Internal validation typically tackles 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?" (**?**, 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 (**?**).

Typically only aggregate-level datasets are available, so researchers are usually compelled to focus internal validation as the measure model performance. Following the literature, we also measure performance in terms of aggregate-level fit, in addition to computational speed. The danger with such a focus on aggregate-level fit, prevalent in the literature, is that it can occur to the detriment of alternative measures of performance. Provided with real spatial microdata, researchers could use individual-level distributions of continuous variables and comparison of joint-distributions of cross-tabulated categorical variables between input and output datasets as additional measures of performance. However one could argue that the presence of such detailed datasets removes the need for spatial microsimulation (**?**) . Real spatial microdata, where available, offer an important opportunity for the evaluation of spatial microsimulation techniques. However, individual-level external validation is outside the scope of this study.

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 (**?**). In static spatial microsimulation, subsampling 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 (**?**).

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 **??**), 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 **?**. The Chi-squared statistic, 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" (**?**, 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 (**?**) .

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 (**??**). 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 undertaken either before or after *integerisation* (more on this below). As is typical in such circumstances, 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 (see **?** for detail). 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 (always positive) discrepancies between the simulated and known population in each cell of the geographically aggregated (wide format) results. The standardised absolute error (SAE) is simply TAE divided by the total population (**?**). TAE provides no level of significance, so thresholds are used. **?** used an error threshold of 80 per cent of the areas with less than 20 per cent error (SAE <0.20); an error threshold of 10% (SAE <0.1) can also been used **?**. Due to its widespread use, simplicity and scale-free nature, SAE is an often preferred metric.

3.4 The modified Z-score

The Z-statistic is a commonly used measure, reporting the number of standard deviations a value is from the mean. In this work we report *zm*, a modified version to the Z-score metric that is robust to very small expected cell values (**?**) . Z-scores apply to individual cells (for a single constraint in a single area), and also provide a global statistic on model fit (*Zm*), defined as the sum of all zm^2 values (**?**) . The measure of fit has the advantage of taking into account absolute, rather than just relative, differences between simulated and observed cell count:

$$Zm_{ij} = (r_{ij} - p_{ij}) \Bigg/ \left(\frac{p_{ij}(1 - p_{ij}))}{\sum\limits_{ij} U_{ij}} \right)^{1/2} \tag{4}$$

where

$$p_{ij} = \frac{U_{ij}}{\sum\limits_{ij} U_{ij}} \qquad \textit{and} \qquad r_{ij} = \frac{T_{ij}}{\sum\limits_{ij} U_{ij}}$$

To use the modified Z-statistic as a measure of overall model fit, one simply sums the squares of *zm* to calculate Zm^2. This measure can handle observed cell counts below 5, which chi-squared tests cannot (**?**) .

3.5 Proportion of values deviating from expectations $E > 5\%$

The proportion of values which fall beyond 5% of the actual values is a simple metric of the quality of the fit. It focusses on the proportion of values which are close to expected, rather than the overall fit, so is robust to outliers. The $E > 5\%$ metric implies that getting a perfect fit is not the aim, and penalises relationships that have a large number of moderate outliers, rather than a few very large outliers. The precise definition of 'outlier' is somewhat arbitrary (one could just as well use 1%).

3.6 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 (?). Like SAE, the RMSE takes the discrepancies between the simulated and known values, squares these and divides the sum by the total number of observations (areas in our case), after which the root is taken. The MAE basically does the same without the square and root procedure. The former has the benefit of being widely used and understood, but is influenced by the mean value in each zone. Recent research in climate model evaluation suggests that the lesser known but simpler MAE metric is preferable to RMSE in many ways (?). We use the RMSE error in the final results as this is a widely understood and applied method with a proven track record in the social sciences.

The relationship between these different goodness-of-fit measures was tested, the results of which can be seen towards the end of Section ??.

4 Input data

The size and complexity of input data ~~have a strong influence on~~ influence the results of spatial microsimulation models. Especially important are the number of constraint variables and categories. We wanted to test the method under a range of conditions. Therefore, three unrelated scenarios, each with their own input datasets were used. These are referred to as 'simple', 'small-area' and '~~sheffield~~Sheffield'. Each consists of a set of aggregate constraints and corresponding individual-level survey tables. To ensure reproducibility, all the input data (and code) used to generate the results reported in the paper are available online on the code-sharing site GitHub.[3] An introductory tutorial explains the process of loading and 'cleaning' input data (?) . This tutorial also introduces the basic code used to create and modify the weight matrices with reference to the 'simple' example described below.

The simplest scenario consists of 6 zones which must be populated by the data presented in the previous section (see ??). 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 ?? and ??, the first row of which is used to exemplify the method). 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.

	Age constraint		Sex constraint	
Zone	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 constraint files describe the number of ~~hours worked, the marital status and the tenancy of these individuals (although tenancyin fact counts houses, not people) : the files can be accessed from '~~people (number of households for tenancy) in each category for the following variables:

- hours worked (see 'input-data/small-area-eg~~', within the 'input-data' folder of the online repository (see the 'input-data'sub-directory of the) .~~/hrs-worked.csv', which consists of 7 categories from '1-5' to '49+' hourse per week)
- marital status (see 'marital-status.csv', containing 5 categories)
- tenancy type ('tenancy.csv', containing 7 categories).

[3] The individual level survey tables are anonymous and have been 'scrambled' to enable their publication online and can be downloaded by anyone from http://tinyurl.com/lgw6zev tinyurl.com/lgw6zev. This enables the results to be replicated by others, without breaching the user licenses of the data.

The individual-level table that complements these geographically aggregated count constraints for the 'small area' exaple 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). This and all other data files are contained within the 'input-data' folder of the online repository (see the 'input-data' sub-directory of the Supplementary Information).

The largest example dataset is 'Sheffield', which represents geographically aggregated from the UK city of the same name, provided by the 2001 Census. The constraint variables used in the *baseline* version of this model can be seen in the folder 'input-data/sheffield' of the Supplementary Information. They are:

- A cross-tabulation of 12 age/sex categories
- Mode of travel to work (11 variables)
- Distance travelled to work(8 variables)
- Socio-economic group (9 variables)

As with 'small-area', the individual-level dataset associated with this example is a *scrambled* and *anonymised* sample from the Understanding Society. In addition to those listed above at the aggregate level, 'input-data/sheffield/ind.RData' contains columns on whether their home location is urban/rural (variable 'urb'), employment status ('stat') and home region ('GOR').

5 Method: model experiments

5.1 Baseline conditions

Using the three different datasets, the effects of five model settings were assessed: number of constraints and iterations; initial weights; empty cell procedures; and integerisation. In order to perform a wide range of 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. The fit improves with each step. The 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.

The main characteristics of the baseline model for each scenario area presented in **??**. The meaning of the row names of **??** should become clear throughout this section.

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%

5.2 Constraints and iterations

~~These~~

The number of constraints refers to the number of separate categorical variables (e.g. age bands, income bands). The number of iterations is the number of times IPF adjusts the weights according to all available constraints. Both are known to be important determinants of the final output, but there has been little work to identify the optimal number of either variable. ? suggested including the "maximum amount of information in the constraints", implying that more is better. ? , by contrast, notes that the weight matrix can become overly sparse if the model is over-constrained. Regarding the number of iterations, there is no clear concensus on a suitable number, creating a clear need for testing in this area.

Number of constraints and iterations were the simplest model experiments~~, requiring only additional iterations and re-arrangement~~. These required altering the number of iterations or rearranging of the order in which constraints were applied. 1, 3 and 10 ~~iteration (up to 20 are reported in the literature, but we found no difference in the results beyond 10) are reported in the final results.~~iterations were tested and all combinations of constraint orders were tested.[4] The order with the greatest impact on the results ~~is the one reported in the final results table (~~for each example is reported in **??**~~)~~.

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. 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 (**??**).

In the final results, extreme cases of setting initial weights are reported: setting a 10% sample to 100 and 1000 times the default initial weight of 1. In the baseline scenario there are only 3 iterations, allowing the impacts of the initial weights to linger. This influence was found to continue to drop, tending towards zero influence after 10 iterations even in the extreme examples.

5.4 Empty cells

Empty cells refer to combinations 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 **??**. Formalising the definition, survey dataset that contains empty cells may be said to be *incomplete*; a survey dataset with no empty cells is *complete*. **?** was the first to describe the problem of empty cells in the context of spatial microsimulation, and observed that the procedure did not converge if there were too many in the input microdata. However, no methods for identifying their presence were discussed, motivating the exploration of empty cells in this paper.

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 ($Nperm$) is defined by **??**, where $n.cons$ is the total number of constraints and $n.cat_i$ is the number of categories within constraint i:

$$Nperm = \prod_{i=1}^{n.cons} n.cat_i \tag{5}$$

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, $Nperm$ 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 ($Nuniq$), the test to check a dataset for empty cells is straightforward, based on **??**:

$$is.complete = \left\{ \begin{array}{ll} TRUE & \text{if } Nuniq = Nperm \\ FALSE & \text{if } Nuniq < Nperm \end{array} \right\} \tag{6}$$

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 complete input dataset ($Ind.complete$). The 'missing' individuals, needed to be added to make Ind complete can be defined in terms of set theory (**?**) by **??**:

$$Ind.missing = Ind.complete \setminus Ind = \{x \in Ind.complet : x \notin Ind\} \tag{7}$$

[4] Up to 100 iterations are used in the literature (**?**) . We found no difference in the results beyond 10 iterations so a maximum of 10 iterations is reported. For an explanation of how to modify model parameters including the number of iterations, see the 'README.md' text file in the root directory of the Supplementary Information.

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 the 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. 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 'empty-cells.R' in the 'simple' folder to see this). Experiments in which empty cells were removed through the addition of individuals with attributes defined by **??** 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 (**?**). We used modified R implementations of two probabilistic methods — 'truncate replicate sample' (TRS) and 'proportional probabilities' — reported in (**?**) and the Supplementary Information associated with this paper (**?**), as these were found to perform better in terms of final population counts and accuracy than the deterministic approaches.

5.6 Computational efficiency

The code used for the model runs used throughout this paper is based on that provided in the Supplementary Information of **?**. This implementation was written purely in R, a vector based language, and contains many 'for' loops, which are generally best avoided for speed-critical applications. R was not designed for speed, but has interfaces to lower level languages ~~(?) . Programmers can take advantages of these interfaces when efficiency — the amount of information processed per unit time — is a priority (?)~~ (?) . Another advantage of R from the perspective of synthetic populations is that it has an interface to NetLogo, a framework for agent-based modelling (?) . One potential use of the IPF methods outlined in this paper is to generate realistic populations to be used for model experiments in other languages (see ?).

To increase the computational efficiency of R, experiments were conducted using the ~~recently published (July 2014)~~ **ipfp** package, ~~which implements IPF in the low-level C programming language. The package is hosted on the Comprehensive R Archive Network (CRAN) website and is available to download under the terms of the open source Apache License. See for the package's documentation~~a C implementation of IPF (?) .

For each scenario in the Supplementary Information, there is a file titled 'etsim-ipfp.R', which performs the same steps as the baseline 'etsim.R' model run, but using the `ipfp()` function of the **ipfp** package. The results of these experiments for different scenarios and different numbers of iterations are presented in section **??**.

6 Results

6.1 Baseline results

After only 3 complete iterations under the baseline conditions set out in **??**, 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.

6.2 Constraints and iterations

As shown in **??**, fit improved with each iteration. This improvement generally happened for each constraint and always from one iteration to the next. Through the course of the first iteration in the 'simple' scenario, for example, R^2 improved from 0.67 to 0.9981. There is a clear tendency for the rate of improvement in model fit to drop with subsequent computation after the first iteration: 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 **??** and in the online 'measures' table. These results show that IPF successfully took the simulated cell

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).

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 rapid alignment of zone/constraints to the 1:1 line in **??**. 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.

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 **??**, which shows the impact on individual 228 and a randomly selected individual whose initial weight was not altered, for all areas.

?? 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 **??**, 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).

?? 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 62 and 8,105 empty cells in the 'small-area' and 'Sheffield' baseline scenarios, respectively. Only the smallest input dataset was *complete*. Due to the multiplying effect of more variables (**??**), *Nperm* 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 became 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.

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).

6.5 Integerisation

After passing the final weights through the integerisation algorithms mentioned in ??, 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 Computational efficiency

The times taken to perform IPF in the original R code and the new low-level C implementation are compared in ??. This shows that running the repetitive part of the modelling in a low-level language can have substantial speed benefits — more than 10 fold improvements were found for all but the simplest model runs with the fewest iterations. ?? shows that the computational efficiency benefits of running IPF in C are not linear: the relative speed improvements increase with the both the number of iterations and the size of the input datasets.

This suggests two things:

- Other parts of the R code used in both the 'pure R' and C implementations via **ipfp** could be optimised for speed improvements.
- Greater speed improvements could be seen with larger and more complex model runs.

Table 10: Time taken for different spatial microsimulation models to run in microseconds, comparing R and C implementations code. 3 (left hand columns) and 10 (right hand columns) iterations were tested for each. 'Improvement' represents the relative speed improvement of the C implementation, via the **ipfp** package.

		3 iterations			10 iterations	
Scenario	R code	C code (ipfp)	Improvement	R code	C code (ipfp)	Improvement
Simple	73	8	9.1	231	8	28.9
Small-area	3548	204	17.4	10760	288	37.4
Sheffield	44051	1965	22.4	135633	2650	51.2

6.7 Summary of results

The impact of each model experiment on each of the three scenarios is summarised in ??. 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 'Sheffield', the use of only one complete iteration (not shown in ??) instead of 3 or more had the a more damaging impact than any of those presented, highlighting the importance of performing multiple iterations of IPF.

Substantial improvements in execution speed have been demonstrated using a lower-level implementation of the IPF algorithm than has been used in previous research. If computational efficiency continues to increase, the code could run 100 times faster than current implementations on larger datasets (see ??).

The results presented in Table ?? were robust to the type of measure. We calculated values for the six goodness of fit metrics described in Section ?? and illustrated in Fig. ?? for each of the three example datasets. The results after 10 iterations of 'baseline' scenario are saved in table files titled 'measures.csv' in each of the model folders. The results for each metric in the Sheffield model, for example, can be found in 'models/sheffield/measures.csv'. The calculation, storage and formating of this arrary of results is performed by the script 'analysis.R' in the models folder.

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

Test scenario	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

Overall it was found that there was very good agreement between the different measures of goodness-of-fit. Despite often being seen as crude, the r measure of correlation was found to be very closely related to the other measures. Unsurprisingly, the percentage of error metric (‘E5’ in Fig. ??) fitted worse with the others, because this is a highly non-linear measure that rapidly tends to 0 even when there is still a large amount of error.

The conclusion we draw from this is that although the *absolute* goodness-of-fit varies from one measure to the next, the more important *relative* fit of different scenarios remains unchanged. The choice of measure becomes more important when we want to compare the performance of different models which vary in terms of their total population. To test this we looked at the variability between different measures between each scenario, the results of which are shown in table ??. It is clear that r, SAE and $E5$ metrics have the advantage of providing a degree of cross-comparability between models of different sizes. TAE, $RMSE$ and the sum of $Z - squared$ scores, by contrast, are highly dependent on the total population.

Table 12: Summary statistics for the 6 goodness of fit measures for each iteration of each ’baseline’ scenario.

Model (mean)	r	TAE	SAE	RMSE	Zs	E5
Simple	0.942	4.304	0.036	0.287	2.3640e+00	0.132
Small-area	0.964	968.934	0.073	4.195	1.1936e+07	0.160
Sheffield	0.972	89531.053	0.098	68.703	6.8430e+04	0.355
Model (SD)						
Simple	0.138	8.762	0.073	0.584	5.6490e+00	0.215
Small-area	0.074	1734.436	0.131	7.224	2.3784e+07	0.204
Sheffield	0.046	89626.742	0.098	61.599	1.1313e+05	0.206

corplot.pdf

Figure 6: Correlation matrix comparing 6 goodnes-of-fit measure for internal validation. The numbers to the bottom-left of the diagonal line represent Pearson's r, the correlation between the different measures. The circles represent the same information in graphical form. The abbrevitiation are as follows. r: Pearson's coefficient of correlation; TAE: Total Absolute Error; SAE: Standardised Absolute Error; Zs: Z-score; E5: Errors greater than 5% of the observed value.

7 Conclusions

In this paper we have systematically reviewed and tested modifications to the implementation of the IPF algorithm for spatial microsimulation — i.e. the creation of 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 to cross-check the results and facilitate 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'.
- Integerisation of the fractional weights using a couple of probabilistic methods — 'TRS' and 'proportional probabilities'.
- The language in which the main IPF algorithm is undertaken, from R to C via the **ipfp** package.

In every case, the direction of change 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 results, 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, notably empty cells.

A new contribution of this paper is its analysis of empty cells — individuals absent in the individual-level dataset but implied by the geographically aggregated constraints. The methods presented in **??** should enable others to identify empty cells in their input datasets and use this knowledge to explain model fit or to identify synthetic individuals to add. This is the first time, to our knowledge, that empty cells have been shown to have large adverse effect on model fit compared with other variables in model design.

The removal of empty cells greatly improved fit, especially for complex scenarios. This straightforward change had the single greatest positive impact on the results, leading to the conclusion that removing empty cells can substantially enhance the performance of spatial microsimulation models, especially those with many constraint variables and relatively small input input datasets. On the other hand, a 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. Another potential problem caused by removal of empty cells that must be avoided is the creation of impossible combinations of variables such as car drivers under the age of 16.

~~The most~~ A useful practical finding ~~for future researchers dealing with large spatial microsimulation models and limited computational resources is likely to relate to computational efficiency. The implementation of the IPF procedure in a low-level language via the recently published~~ is that IPF implemented using the **ipfp** package ~~yielded execution times dozens of times~~ yields execution much faster than the ~~original R implementation~~R code described in (?) . The speed benefit of this low-level implementation rose with additional iterations and model complexity, suggesting that the method of implementation may be more influential than the number of iterations. Regarding the optimal number of iterations, there is no magic number identified in this research. However, it is clear that the marginal benefits of additional iterations decline rapidly: we conclude that 10 iterations should be sufficient for complete convergence in most cases.

~~There are many~~ Many opportunities for further research ~~raised by this paper . There is clearly a need for further testing in the area~~ emerge from this paper on the specific topic of IPF and in relation to the field of spatial microsimulation ~~overall, and we have demonstrated this for the established IPF method. Research questions~~ more generally.

Questions raised relating to IPF ~~specifically include:~~ include:

- What are the impacts of constraining by 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? - What are the opportunities for further improvements in computational efficiency of the code? - How does IPF respond to unusual constraints, such as cross-tabulated contingency tables with shared variables (e.g. age/sex and age/class)? - How are the results of IPF (and other reweighting methods) affected by correlations between the individual-level variables?

There is clearly a need for further testing in the area of spatial microsimulation overall, beyond the specific example of IPF in R used here. The generation of synthetic populations allocated to specific geographic zones can be accomplished using a variety of other techniques, including:

- The Flexible Modelling Framework (FMF), a open sourcesoftware written in Java for spatial modelling. The FMF can perform spatial microsimulation using simulated annealing.
- Generalised regression weighting (GREGWT), an alternative to IPF written in SAS (?) .
- SimSALUD, a spatial microsimulation application written in Java, which has an online interface (?) .
- CO, a spatial microsimulation package written in Fortran and SPSS (?) .

A research priority emerging from this research is to set-up a framework for systematically testing and evaluating the relative performance of each of these approaches. This is a challenging task beyond the scope of this paper, but an important one, providing new users of spatial microsimulation a better idea of which options are most suitable for their particular application.

Moreover, there are broader questions raised by this research concerning the validation of models that are used for policy making. **?** 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. We contend that such issues, with real-world consequences, are exacerbated by a lack of reproducibility, model evaluation and testing across spatial microsimulation research. The ~~paper's methods — supplied in the~~ underlying data and code are supplied in Supplementary Information ~~— and reproducible results address these issues, providing~~ allow others to experiment with the procedure to shed new insight into the ~~likely impacts of different variables on the performance of the IPFprocedure~~factors affecting IPF. There is ~~however much further work to be done, including research into~~ much scope for further research in this area including further work on external validation, ~~handling of very large datasets, and grouping output~~ synthesis of large spatial microdatasets, the grouping of spatial microdata into household units and the integration of methods of spatial microsimulation into ABM. Seen in this wider context, the model experiments presented in this paper can be seen as part of a wider program of systematic testing, validation and benchmarking in the field of spatial microsimulation.