Journal of Urban Mobility 2 (2022) 100027

Contents lists available at ScienceDirect

Journal of Urban Mobility

journal homepage: www.elsevier.com/locate/urbmob

Simulating micro-level attributes of railway passengers using big data

Eusebio Odiari[a,b,*], Mark Birkin[a]

[a] Consumer Data Research Centre, Leeds Institute for Data Analytics, LS2 9JT, United Kingdom
[b] School of Geography, University of Leeds, Leeds LS2 9JT, United Kingdom

ARTICLE INFO

Keywords:
Spatial microsimulation
Population synthesis
Micro-level attributes
Railways
Big data
Consumer data
Railway ticketing
Passenger mobility

ABSTRACT

In the absence of a comprehensive, representative, and attribute-rich population, a spatial microsimulation is necessary to simulate or reconstruct a population for use in the analysis of complex mobility on the railways. Novel consumer datasets called 'big-data' are exhaustive but they only reveal a subset of the wider population who consume a specific digital service. Further, big-data are measured for a particular purpose and so do not have the broad spectrum of attributes required for their wider application. Harnessing big-data by spatial microsimulation has the potential to resolve the above shortcomings. This paper explores the relative merits of different spatial microsimulation methodologies, and a case study illustrates how best to simulate a micro-population linking rail ticketing big-data with the 2011 Census commute to work data and a National Rail Travel Survey (NRTS). The result is a representative attribute-rich micro-level population, which is likely to have a significant impact on the quality of inputs to strategic, tactical and operational rail-sector analysis planning models.

1. Introduction

Progressively complex societal urban mobility has meant that transport planners require insights into robust rules governing movement patterns of people and the interdependencies with demography, municipal parameters, space and time, in order to provide sustainable efficient services. There is a growing need for novel geographical modelling tools, as well as attribute-rich, comprehensive and representative data to assist in such research and decision-making. Today's wide use of electronic devices has meant that novel sources of large consumer datasets are now increasingly readily available, however it is unrealistic to expect all the information required to be provided by a single dataset (De Montjoye, Hidalgo, Verleysen & Blondel, 2013; Lynch, 2008; Manyika et al., 2011). As such, methods have to be developed to combine various datasets, within mobility analysis frameworks.

Once the relevant datasets are identified, it is often found that the process of integration requires adjustments in resolution and in geographies of scale to maintain consistency between the datasets (Deming & Stephan, 1940). Often one dataset has to be systematically adjusted to fit the resolution, and system process associated with other datasets (Howe et al., 2008; Weber, Mandl & Kohane, 2014), and the hypothesis is that the adjustment methodology adopted impinges on the quality of subsequent postulations based on such integrated data. In this paper, we investigate the different methodologies for combining disparate datasets to integrate their resolution and geographies and to simulate a population represented at micro-scale (i.e. individual/household) levels. A case study illustrates how best to apply such methodologies to urban mobility in West Yorkshire. The datasets combined are the 2011 Census interaction data, the National Rail Travel Survey and big railway ticketing data for West Yorkshire study area procured from the Association of Train Operating Companies' (ATOCs) Latest Earnings Network Nationally over Night (LENNON) database.

A LENNON dataset including every railway ticket sold in the UK is large and exhaustive of railway passengers (ORR, 2016). However, for use in research to relate population mobility to demographic and other likely drivers of behaviour, LENNON data is lacking because it does not contain information on say passenger residence, final destination, nor does it identify any socio-demographic and urban morphology characteristics associated with the passenger. In order to use such ticketing data for mobility analysis, it would be necessary to combine with other relevant datasets that contain the sought attributes. The 2011 Census measures the interaction between each geography (OA, LSOA or MSOA) zone, measuring the volume of passengers commuting from locations of usual residence to places of work (Rees, Martin & Williamson, 2002; Stillwell & Duke-Williams, 2003). These measures are aggregated for a range of socio-demographic attributes (age, gender, mode of commute, occupation, ethnicity, and the cars, children, and type of household). The National Rail Travel Survey (NRTS) is another relevant railway passenger survey conducted over 2001 to 2005; it aimed to produce a comprehensive picture of weekday rail travel across Great Britain. The NRTS identifies places of usual residence and final destination (DfT, 2013), as well as a range of social-demographic and, sample passenger train station access and egress attributes. Considerations outlined below form the basis for reconciling NRTS and Census, despite that they were acquired at different times.

[*] Corresponding author.
E-mail address: e.a.odiari@gmail.com (E. Odiari).

https://doi.org/10.1016/j.urbmob.2022.100027
Received 31 October 2020; Received in revised form 2 June 2022; Accepted 17 June 2022

As discussed, big-data[1] are advantageous in that each set offers a unique and distinctive view of real events. Combining such datasets reveal a more complete picture of reality, and this is a particular advantage of big-data which make them attractive for the analysis of a wide range of phenomena. The NRTS and Census are random samples of the UK population taken at different times; hence their variables have different crystallization times (i.e. times at which they were digitally measured). Within the time when the 2001/5 NRTS and 2011 Census were measured, there were increases in passenger volumes, changes in travel behaviour, etc., so the datasets have different conditional distributions. Whilst the NRTS is conditional on time (i.e. 2001/5) and a transport-mode (i.e. train), the Census is conditional on a different time (i.e. 2011) and different transport mode (i.e. all modes of transport). This brought about reports in the literature (Gower, 2021) that structural changes occurred in the UK rail demand, which resulted in annual sales differential growth rates in non-season tickets (~38%) and in season tickets (~26%) between 2005 and 2011. Further, in this period, rail commuters were observed to travel longer distances but less frequently (Le Vine, Polak & Humphrey, 2017; ONS-UK, 2013). This paper addresses the concern about combining such structurally different data like the NRTS and Census acquired at different times. The hypothesis developed in this paper is that disparate datasets like the NRTS and Census can be objectively combined if they do not represent concept or data drift within the model development process. Whilst concept drift occurs in a model where properties of the dependant variable change, data drift occurs where those of independent variables change.

Big-data like LENNON[2] ticketing information are exhaustive, comprehensive, and representative of individual-level rail travel within the UK. However, they are not intended to capture information on the UK population and should only be linked to a subset that uses the trains. Expectedly the distribution of LENNON will be different from that of the UK Census. Each (LENNON or Census) contains data drawn from the same population (the UK full joint distribution). Each dataset is conceived as a conditional distribution, giving the probabilities contingent upon other variables. In the instance of the LENNON and Census, the conditional on the former is 'mode of travel is by train', whilst in the later the conditional is 'mode of travel is by all modes'. The Census being representative of the full UK population joint distribution is taken as the reference. The Census frequency distribution is made up of 6% rail commuters whilst LENNON is 62% rail commuters. In this context, the 'travel-purpose' frequency distributions of the Census and LENNON are markedly structurally different. As the Census is the reference, we refer to LENNON as a conditional distribution of the Census and in so doing we concede that this does not imply underlying LENNON bias. This terminology is consistently adopted in this paper to highlight the sort of issues typically associated with disparate big-datasets. Any dataset that is not a representative sample of the wider population is a conditional distribution of the Census. Being a conditional is typical of the nature of big-data as they are typically acquired at different times for a specific purpose. This character of big-data poses challenges in integrating with other datasets from measured stated surveys; designed to be representative of the wider population.

Two broad population simulation strategies exist: the first is population synthesis (or reweighting) whereby a seed sample is optimal replicated to fit the dimensions of an aggregate target population. In this case, the simulated population created is made up of multiples (weights) of the seed sample. The second strategy is population reconstruction whereby a constructed model is used to generate a new population such that it fits the dimensions of a target population. In this second strategy the population created has no direct reference to an existing population. The population synthesis and reconstruction methodologies are both further divided into deterministic and stochastic methodologies.

The key shortcomings of the existing population synthesis and the population reconstruction strategies are summarized below. First for the population synthesis:

- Only the marginal distribution of the target is exploited, and not the full joint distribution of the target (Odiari, 2018).
- Only permits sequential fitting of disparate contingency seed data (Odiari, Birkin, Grant-Muller & Malleson, 2021).
- Yields clusters of passengers with same attribute if seed is of small sample ratio (Zhu & Ferreira Jr, 2014).
- Assumes that the seed is representative of the target and has similar distribution (Lomax & Norman, 2016).
- Only a limited number of attributes can be combined due to computational limitations (Müller & Axhausen, 2010; Odiari, 2018).
- Does not incorporate a robust statistical strategy for assessing the errors and uncertainty in results (Lovelace, Dumont, Ellison & Založnik, 2017; Whitworth, Carter, Ballas & Moon, 2017).

For the population reconstruction:

- Accuracy of the simulated population is dependant on model constructed (Farooq, Bierlaire, Hurtubia & Flötteröd, 2013).
- The simulated population is not directly related to the actual population (Müller & Axhausen, 2010).
- Incorporates assumptions inherent in regression models (Tanton & Vidyattama, 2010).

In this paper we review and validate the range of population simulation strategies for use to adjust 'big data'[3] for consistency with survey data[4] and established theory. The science behind the range of population simulation methodologies is presented in a practice orientated way, as a pre-cursor to a case study of, for the first time simulating a representative population portrayed at micro-scale interacting between zones and through the railway network.

2. Literature review

Census exigencies around confidentiality meant that despite complete counts of the populace, only a small sample of anonymized records (SAR) of cross-tabulations of all individual attributes were released. This SAR data was released alongside comprehensive sets of aggregate tables typically limited to only three variables per table (Williamson, Birkin & Rees, 1993). Despite the anonymized data and aggregation in respect of individual privacy, population geographers and transport planners realize that comprehensive and disaggregated records would enable the construction of more detailed pictures of geographic and transport phenomena. The need to create such micro-data led to develop the field of

[1] Big data is a term coined to describe the range of consumer datasets that are increasingly readily available from sensors of consumers of digital services. The data analysts refer to datasets as being 'tall' or 'fat' depending respectively on whether the data has a commensurately larger number of observations or covariates. Under such conditions, data that is both fat and tall is described as 'big'. Business analysts with an aptness for buzzwords refer to the five 'V's, and describe big data as having volume, velocity, variety, veracity and value.

[2] In Geographical Information Systems (GIS), big data refers to data that are large and unstructured that they do not fit on to conventional hardware and software information tools. About 1.3bn tickets are sold in the UK mostly recorded on LENNON, with over 200 ticket types. This is reflective of volume and variety. This dataset is further exhaustive as it is assumed to represents all railway tickets sold. The LENNON database is the digital service for recording tickets, albeit tailored to train operating companies (RDG-ATOC). As such, the LENNON ticking information is classed here as big-data.

[3] Big-data tend to come from consumers of a specific digital service, putting context on the data. This specific context often makes big-data a specific conditional distribution of the population of interest. This concept has been adopted within this paper.

[4] Survey data tend to be regular measured and stated data so they are random samples representative of a population of interest. Conventional established methods for integrating datasets are typically based on that each dataset being combined is a representative random sample of the population of interest.

study of population synthesis and population reconstruction (Birkin & Clarke, 1988).

2.1. Population synthesis

The first application of population synthesis, often ascribed spatial micro-simulation, was reported in the literature as far back as 1940 (Deming & Stephan, 1940). The method was applied to the USA Census. A synthetic population was formed from an estimate of a combination of multiples of individuals in a SAR dataset that will best fit the aggregate values defined in Census tables, using the Lagrange multiplier constrained optimization method (Bertsekas, 2014). With the development of computers and numerical analysis, instead of solving the linear equations from the Lagrange multiplier method, iterative methods gradually converging towards an optimum were proposed (Fletcher, 2013; Kelley, 1999), and it became increasingly more efficient to resort to these iterative proportional fitting (IPF) strategies, as they are faster to compute, less sensitive to numerical and round-off errors and simpler in algebra, than comparative formulations by Lagrange and Fermat (Fermat, 1891; Lagrange, 1867). The IPF algorithm has found applications in diverse fields as economics (Bacharach, 1970), transport engineering (Fratar, 1954; Furness, 1965), statistics and computer science (Lavrakas, 2008) under various names.

In population geography and demography, IPF came to prominence relatively recently with the range of policy relevant applications and solutions proffered. First, (Birkin & Clarke, 1989) used IPF to simulate individual and household incomes at small area levels, (Rees, 1994) used IPF to project age and gender structure of urban areas, (Ballas, Kingston, Stillwell & Jin, 2007) addressed in detail the use of spatial micro-simulation as a framework for decision support for policy analysis, and (Ballas & Clarke, 2001) assessed the impact of aggregate national policies within segregate local communes. The IPF methods have improved with burgeoning use, and while earlier effort concentrated on developing the microsimulation steps on different platforms, the advent of suites of statistical software packages like R-studio (Team, 2016) have enabled an automation of the processes and a look beyond the steps and stages of iteration onto the characteristics of the solution. Effort latterly has shifted to concerns of numerical stability and propagation of errors in IPF methods (Birkin & Clarke, 1995; Wong, 1992). Further concerns relate to converting fractional values to integer counts of individuals, and in developing strategies for internally and externally validating IPF results (Lovelace & Ballas, 2013; Upton, 1985).

Spatial microsimulation is now considered a mature application (Lomax & Norman, 2016), re-visiting questions like whether we can be confident that unconstrained attributes are reproduced reliably by the IPF process, thereby accurately replicating the distribution of those attributes not included in the spatial microsimulation (Birkin & Clarke, 2011). Further unanswered questions concern whether, more benchmark constraining variables generally translate to better microsimulation results (Markham, Young & Doran, 2017; Smith, Clarke & Harland, 2009; Tanton & Edwards, 2012; Tanton & Vidyattama, 2010), the effect of the disparity in the distributions of the seed sample and target population (Tanton & Edwards, 2012), and the sensitivity of the different spatial micro-simulation strategies to sample ratios of the seed data (Tanton, 2014). These are questions investigated in this paper to explore the potential of spatial micro-simulation methods for use in harnessing mobility (spatial interaction) data acquired from consumers of digital services provided on the UK railways.

Apart from the above deterministic (IPF) spatial micro-simulation strategies which yield the same result on repeat, the alternative strategies are stochastic, typically Monte Carlo Markov chain (MCMC) based methods. The MCMC methods are efficient for converging to solutions of intractable complex constrained optimization problems (Asmussen & Glynn, 2007; Brooks, Gelman, Jones & Meng, 2011). Population geographers have developed a range of MCMC variants to complement traditional deterministic IPF. Some of these methods have been branded hill-climbing and have been compared to IPF methods (Kurban, Gallagher, Kurban & Persky, 2011). The strategies branded simulated annealing have been compared with IPF (Harland, Heppenstall, Smith & Birkin, 2012). Another MCMC based strategy the genetic algorithm, was used to simulate network traffic by reconciling observed and estimated flows, opening the realm for application to transport problems (Dimitriou, Tsekeris & Stathopoulos, 2006). Hill-climbing, simulated annealing and genetic algorithms are local search optimisation methods which start with an arbitrary solution to a problem and then iteratively makes incremental improvements to the solution by sampling alternatives. The objective functions in these methods would be similar, with algorithm differences lying in the (stochastic chain) rule for accepting or rejecting a sample that forms the alternative solution (Kavroudakis, Ballas & Birkin, 2008). Other stochastic methods include the Bayesian expectation maximization (EM) strategy (Dempster, Laird & Rubin, 1977) which performs optimization by iteratively estimating the maximum likelihood.

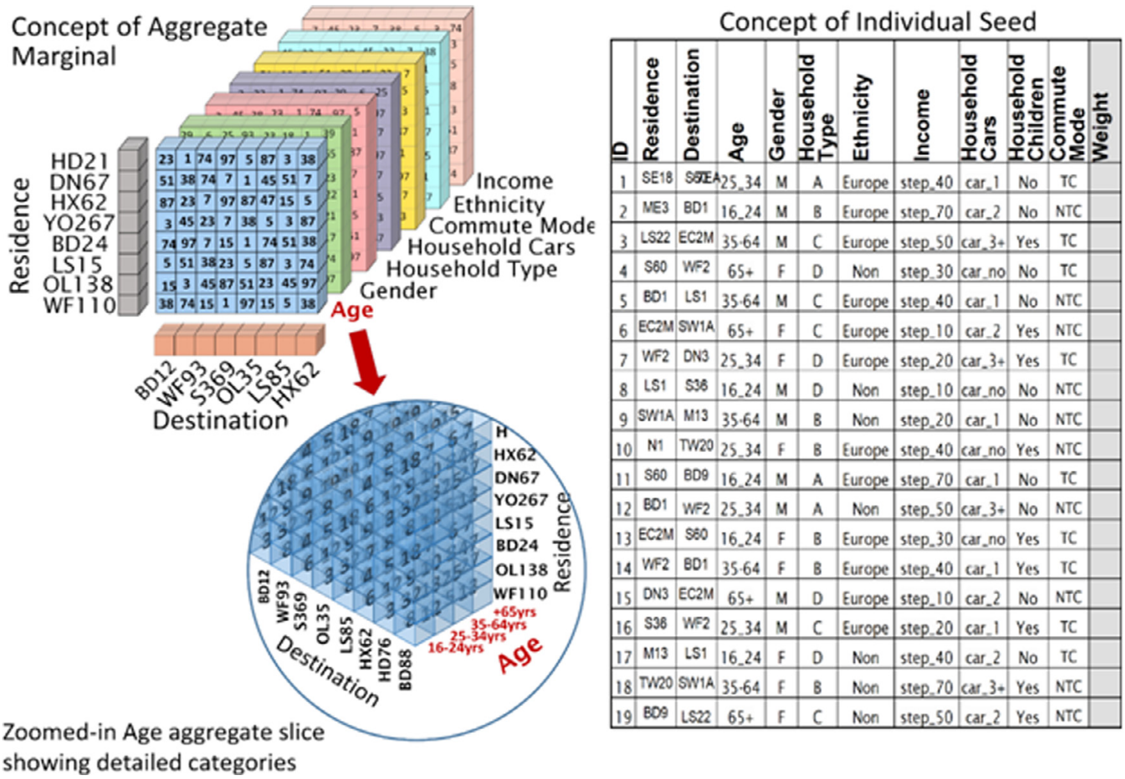
2.2. Population reconstruction

The synthetic reconstruction is another spatial micro-simulation strategy developed in the UK (Birkin & Clarke, 1988; Birkin, Turner & Wu, 2006), exploiting the probabilistic indicative potentials of the Sample of Anonymized Records (SARs) dataset from the Office for National Statistics. The synthetic reconstruction methodology is akin to the Gibbs sampling procedure (Gelfand, Hills, Racine-Poon & Smith, 1990), a variant of the MCMC. With the advent of novel consumer datasets and so-called big data which tend to be a conditional distribution of the wider population, the assumption of normally distributed errors between the target and proposal distributions no longer holds as big data tends to suffer sample bias (Kitchin, 2014). As a result, deterministic and stochastic strategies traditionally set up for representative population seed samples need to be validated for consumer data which often tend to be a conditional distribution of the target population, and structurally dissimilar to the conventional seed data. This validation forms the crux of the first part of this paper. Thereafter, the case study presented in this paper is the first application of spatial microsimulation to synthesize individual railway passengers interacting between geographic locations, using data from the Census, the NRTS and comprehensive LENNON ticketing big data.

3. Deterministic and stochastic formulation

Spatial micro-simulation fall in the wider description of constrained optimization problems (Sun & Yuan, 2006). The typical aim is to estimate a best possible (i.e. optimal) synthetic population limited such that the zonal aggregate characteristics (i.e. the constraints) are fulfilled. The concept of spatial micro-simulation is illustrated in Fig. 1. In practical applications in population and transport geography, the right table depicts a representative sample taken from the population, each row (tuple) an individual with attributes listed in the columns (fields). In literature, the sample table is severally referred to as seed sample, individual-level data, and survey or simply as the sample. The left in Fig. 1 represents the cross-tabulated form of aggregated zonal population attributes.

Each dimension of the aggregate structure typically represents the attributes of a zone and the categories therein. The vertical grey pillar (on its own) could represent the ‘Residence’ variable made up of categories of geographic zones. If values were included in each cube that made up the vertical grey pillar, these would be the populations associated with each residential zone. A similar description follows for the horizontal brown pillar which (on its own) represents the ‘Destination’ attributes and the zone categories that make up the destinations. The vertical pillar represents a one-dimensional (1*D*) aggregate constraint, just as the horizontal pillar also forms an (1*D*) aggregate constraint. If the vertical or horizontal pillars are further sliced along their lengths, they would



Concept of Individual Seed

ID	Residence	Destination	Age	Gender	Household Type	Ethnicity	Income	Household Cars	Household Children	Commute Mode	Weight
1	SE18	SW1A	25_34	M	A	Europe	step_40	car_1	No	TC	
2	ME3	BD1	16_24	M	B	Europe	step_70	car_2	No	NTC	
3	LS22	EC2M	35-64	M	C	Europe	step_50	car_3+	Yes	TC	
4	S60	WF2	65+	F	D	Non	step_30	car_no	No	TC	
5	BD1	LS1	35-64	M	C	Europe	step_40	car_1	No	NTC	
6	EC2M	SW1A	65+	F	C	Europe	step_10	car_2	Yes	NTC	
7	WF2	DN3	25_34	F	D	Europe	step_20	car_3+	Yes	TC	
8	LS1	S36	16_24	M	D	Non	step_10	car_no	No	NTC	
9	SW1A	M13	35-64	M	B	Non	step_20	car_1	No	NTC	
10	N1	TW20	25_34	F	B	Europe	step_40	car_no	Yes	NTC	
11	S60	BD9	16_24	M	A	Europe	step_70	car_1	No	TC	
12	BD1	WF2	25_34	M	A	Non	step_50	car_3+	No	NTC	
13	EC2M	S60	16_24	F	B	Europe	step_30	car_no	Yes	TC	
14	WF2	BD1	35-64	F	B	Europe	step_40	car_1	Yes	TC	
15	DN3	EC2M	65+	M	D	Europe	step_10	car_2	No	NTC	
16	S36	WF2	25_34	M	C	Europe	step_20	car_1	Yes	TC	
17	M13	LS1	16_24	F	D	Non	step_40	car_2	No	TC	
18	TW20	SW1A	35-64	F	B	Non	step_70	car_3+	Yes	NTC	
19	BD9	LS22	65+	F	C	Non	step_50	car_2	Yes	NTC	

Fig. 1. | Concept of spatial microsimulation depicted by the aggregate marginal totals on the left and the individual-level seed table on the right.

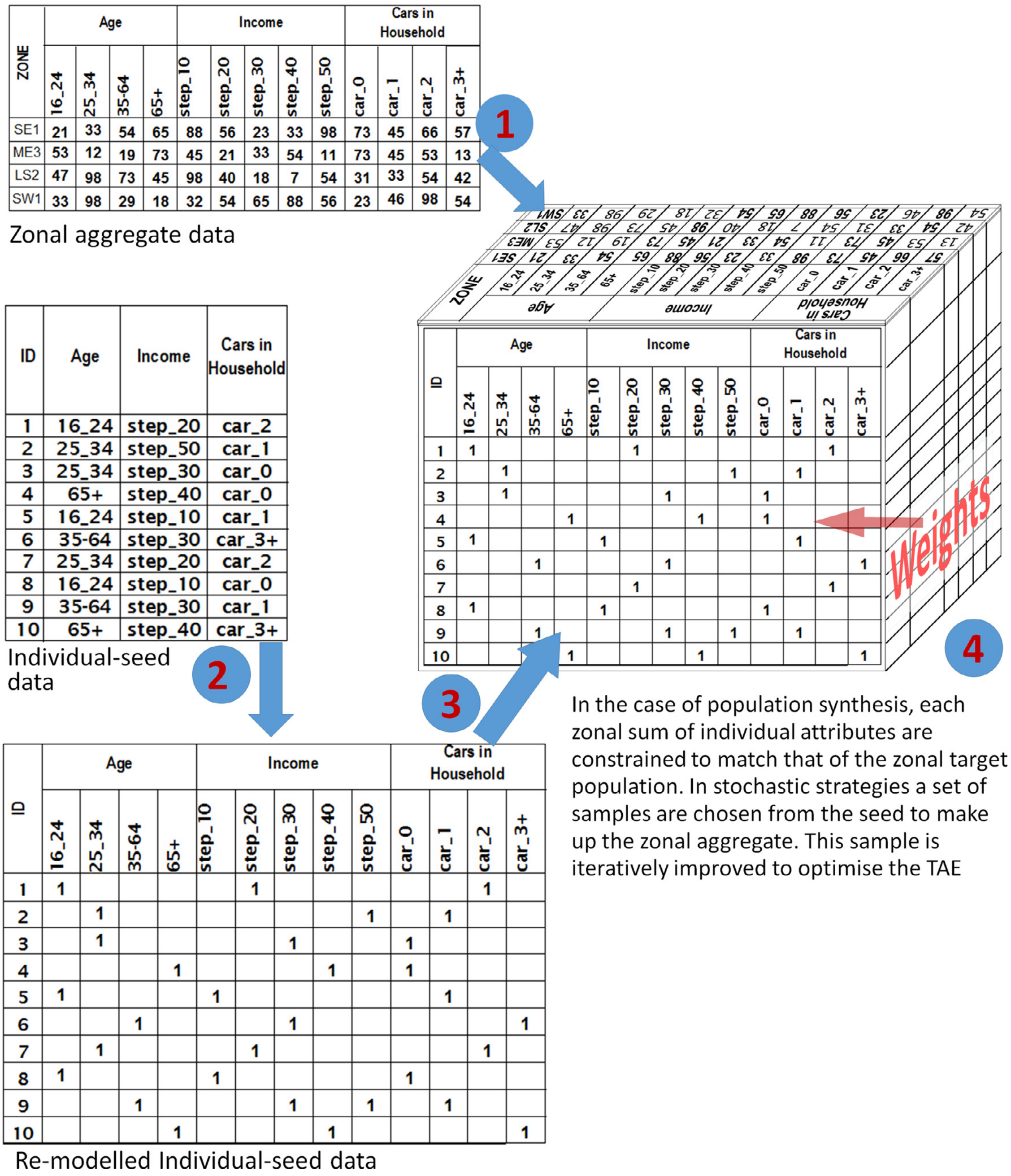
form a 2D constraint, with the second dimension representing say the age variable, sliced into the categories (16–24yrs, 25–34yrs, 35–64yrs, etc.) that make up the age range within the zone. In this illustration the blue slab in front represents an aggregate array cross-tabulation of 'Residence' versus 'Destination'.

A closer inspection of Fig. 1 however reveals that the blue slab is further sliced along the vertical axis, creating the third 'Age' variable, with categories therein. In such an instance, the blue slab aggregate constraint would be an (3*D*) array. The categories making up these 'Residence', 'Destination' and 'Age' variables are illustrated in the blown up section (highlighted in the circle). To further emphasize the nature of the aggregate data, in the blown up section, for passengers residing in Postcode WF110 and working at destinations in Postcode BD88, there are populations of 8, 12, 5 and 13 of ages 16–24yrs, 25–34yrs, 35–64yrs and over 65yrs respectively. The subsequent slabs coloured green, pink, purple, etc. also represent (3*D*) aggregate constraints ('Residence', 'Destination' and a third demographic attribute). As seen, the variables involved in each slab are 'Residence', 'Destination', and another demographic attribute. Each of the variables is further sub-divided into categories (typical of the blown up 'Age' variable in the circle). The 2011Census interaction data consists of several 3D arrays (like those in Fig. 1) made up of aggregates for location of usual residence ('Residence') and place of work ('Destination') for a range of socio-demographic attributes (age, gender, income, mode of commute etc.). These zonal population cross-table data are also referred to as aggregate, marginal, constraint, count, target or Census (as its structural form is typical of published Census counts).

In practice however, the constraints can be any combination of (1*D*),(2*D*),(3*D*), up to *nD* where ($0 \leq n \leq \infty$). Effectively implying that in practice the constraint could be made up of several differently sourced disparate multi-dimensional arrays. The 2011 Census interaction data implies a marginal constraint made up of eight (8) sets of (3*D*) slabs, a slab for the age, gender, ethnicity, commute mode, occupation (used in lieu of income), and attributes for cars, children, and type of household, all separately cross-tabulated against the residence and destination variables and categories therein (Upton, 1985). The intuition behind creating a detailed micro-level population from a sample individual-level seed and zonal aggregates of the population, is that if a zone consists of an aggregate of say 20 people, with particular zonal characteristic, for instance that are mostly aged and on high incomes: then if a seed sample representative of the entire region is available, the zone can be reconstituted from the sample by making an optimized selection of aged people on high incomes from the sample. To satisfy the constraint of sustaining the volume of people in the zone, the limited available aged and high income individuals in the sample might have to be replicated many times, hence the concept of weights which is indicative of how many times a type of individual is replicated to fulfil the synthetic population.

In our particular case of the Census interaction, the weight assigned to an individual would be indicative of how representative the individual is of passengers on a particular residence-destination flow. The concept can be extended to other scenarios involving aggregate constraints and seed data. It is noteworthy to also point out that in situations involving sample data that are a conditional distribution of the target, the weight assigned to an individual could also reflect that the measured seed data did not have a commensurate representative proportion of that particular individual when compared to the proportion in the target population. In essence the adequate population proportion of that individual is 'missing' from the sample seed. The difference in the alternative spatial micro-simulation (population synthesis) strategies lies in the numerical algorithm or probabilistic method adopted for estimating or calculating the weights assigned to individuals in the seed data. Typically, the seed sample measures each attribute for each individual forming a rich variable joint distribution which is limited only by being a sample, instead of the whole population.

On the other hand, the aggregate data (like the Census) is anonymized for confidentiality and the zonal attributes are aggregated to form a cross tabulation. This cross-tabulation creates the challenge of relating individuals within an age category say, to the same individual



ZONE	Age				Income					Cars in Household			
	16_24	25_34	35-64	65+	step_10	step_20	step_30	step_40	step_50	car_0	car_1	car_2	car_3+
SE1	21	33	54	65	88	56	23	33	98	73	45	66	57
ME3	53	12	19	73	45	21	33	54	11	73	45	53	13
LS2	47	98	73	45	98	40	18	7	54	31	33	54	42
SW1	33	98	29	18	32	54	65	88	56	23	46	98	54

Zonal aggregate data

ID	Age	Income	Cars in Household
1	16_24	step_20	car_2
2	25_34	step_50	car_1
3	25_34	step_30	car_0
4	65+	step_40	car_0
5	16_24	step_10	car_1
6	35-64	step_30	car_3+
7	25_34	step_20	car_2
8	16_24	step_10	car_0
9	35-64	step_30	car_1
10	65+	step_40	car_3+

Individual-seed data

ID	Age				Income					Cars in Household			
	16_24	25_34	35-64	65+	step_10	step_20	step_30	step_40	step_50	car_0	car_1	car_2	car_3+
1	1					1						1	
2		1							1		1		
3		1					1			1			
4				1				1		1			
5	1				1						1		
6			1				1						1
7						1						1	
8	1				1					1			
9			1				1		1		1		
10				1				1					1

In the case of population synthesis, each zonal sum of individual attributes are constrained to match that of the zonal target population. In stochastic strategies a set of samples are chosen from the seed to make up the zonal aggregate. This sample is iteratively improved to optimise the TAE

ID	Age				Income					Cars in Household			
	16_24	25_34	35-64	65+	step_10	step_20	step_30	step_40	step_50	car_0	car_1	car_2	car_3+
1	1					1						1	
2		1							1		1		
3		1					1			1			
4				1				1		1			
5	1				1						1		
6			1				1						1
7		1				1						1	
8	1				1					1			
9			1				1		1		1		
10				1				1					1

Re-modelled Individual-seed data

Fig. 2. | Process of remodelling the seed-data in spatial microsimulation.

within the gender, income and the other demographic attribute categories. The solution is to optimize the choice of individuals to minimize the difference between the marginal totals of the aggregate data and the synthesized population. In other words satisfying the condition that the micro-population created from the seed have margins adjusted and constrained to the aggregate margins, thus minimizing the residual between the aggregate and the synthesized population marginal.

Fig. 2 illustrates the procedure of spatial micro-simulation, showing the re-modelling process of the datasets used. The table of individual-level seed data (on the middle left side of Fig. 2) is converted into a binary data table (shown by the bottom left table). The binary table has attributes and categories now represented by data columns re-modelled such that they are commensurate with the columns within the aggregate zonal data (shown as the top left table). This concept illustrates the basis for comparison of the two (seed data and aggregate data) tables. Reconciliation of the seed and aggregate tables is depicted by the cubic array within Fig. 2 (on the right side). For each geographic zone, the seed is sampled to make up the population of the zone, and this initial sample is called the proposal distribution. The proposal distribution is optimised to yield the synthetic population for a zone.

In deterministic methods arithmetic fractions of the proposal distribution are iteratively improved to fit the aggregate constraints. In stochastic methods, random samples are taken with replacement from the seed data to improve the objective function (subject to a proposal distribution), until convergence to the target distribution. As a result deterministic strategies yield fractions of individuals, while stochastic procedures yield integer multiple counts of the individuals in the seed sample. Widely used deterministic strategies are reported in literature (Barthelemy, Suesse, Namazi-Rad & Barthelemy, 2016; Lovelace & Dumont, 2016), as well as stochastic strategies (Kavroudakis, 2015). For successful implementation of the deterministic and stochastic procedures, the same set of variables ought to exist within the seed and aggregate datasets (top left and bottom left tables in Fig. 2).

4. Merits of micro-simulation methods

The strategy in this paper is to explore the relative merits and behaviours of different deterministic and stochastic spatial micro-simulation methodologies, highlighting the problems in practical implementation and advantages associated with the different strategies. This is a pre-cursor to practical implementation on the case study of West Yorkshire UK railways, for the first time simulating representative mobility behaviour of a population of railway passengers at micro-scale (i.e. individual/household levels). The case study illustrates how best to simulate a micro-population linking big data on rail trip-making with information on socio-demographic characteristics.

In assessing the controlled behaviour of the deterministic and stochastic micro-simulation strategies, a subset of the NRTS data (for West Yorkshire) made up of about 23,000 samples is used. There are a maximum of eight (8) variables for each individual in the pre-processed NRTS dataset, as such the aggregate constraints are formed by cross-tabulating and aggregating these variables forming sets of (1*D*), (2*D*), (3*D*) up to (8*D*) target constraints. The seed sample on the other hand is constructed by taking different random samples from the NRTS dataset (to reflect different sample ratios by taking more or less random samples, to reflect a conditional distribution (which is dissimilar to the target distribution) by sampling the upper or lower half of the dataset, and without replacement, etc.).

The NRTS dataset used for the test cases includes variables for passenger residence and destination, mode of commute and a number of other demographic attributes (age, gender, cars in household and household type, ethnicity and children in household). The NRTS data is zoned to geographies of Postcode Sector boundaries. The postcode geography in the UK was created for the purpose of disseminating postal mail. Within this geography, there are 124 Areas, 2987 Districts, 11,192 Sectors and about 2 million Units. As such, on average a Postcode Area will consist of 24 Districts, 90 Sectors, and 16,125 Units (each of about 15 addresses). To manage the computational demands of the spatial micro-simulation algorithms, the NRTS which was originally zoned to Postcode Sector boundaries were re-zoned to larger Postcode Areas. As such, Postcode Sectors, originally 357 were converted to Postcode Areas numbering about 5. The re-zoning averts sparse data-frames which compromise the quality of the results and require special treatment (as discussed in this paper's case study).

As a consequence of re-zonation, remnants of some peripheral Postcode Sectors existed. For instance in a geographic location in the north west corner of the West Yorkshire county, a part sub-set of the Oldham Postcode Area (OL148, OL138, OL145 etc.) was captured and was included in the analysis. Similarly, other sub-sections of other Postcode Areas were captured, including Sheffield, York, Harrogate, Blackburn and Doncaster (S, YO, HG, BB and DN) respectively. Perhaps these might have been removed from the analysis, but they have however been included to replicate situations where a zone has a commensurately low number of counts. In effect, inclusion of these marginal Postcode Sectors enables an assessment of the sensitivity of the different spatial micro-simulation strategies to nominally low seed counts.

This section investigates the impact on the results of four separate aspects of the microsimulation process. Each of the four investigative aspects consisted of several individual distinct runs of the microsimulation procedure using a Monte Carlo experiment (totalling over 500 runs to capture the full spectrum of results), detailed in Section 4.1 to Section 4.2. The objectives of each aspect are respectively:

1) to use a Monte Carlo experiment to investigate the effect of the number of constraints by increasing these from one through to seven (which is one less the total of eight variables).

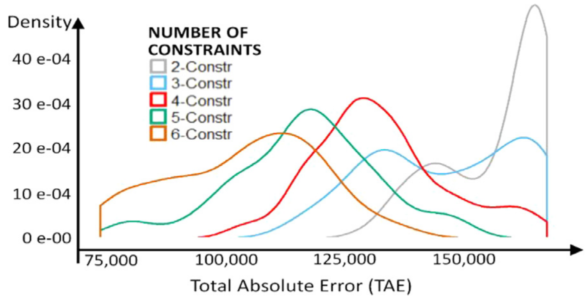


Fig. 3. | Variation in TAE with the number of constraints for the deterministic IPF spatial micro-simulation.

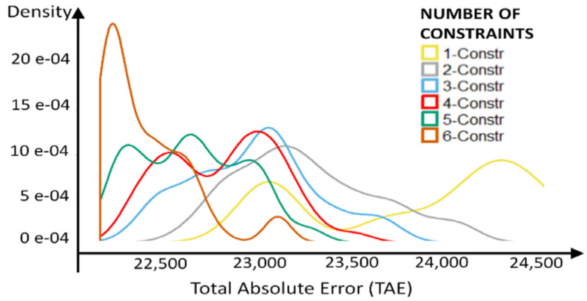


Fig. 4. | TAE from repeated random sets of a fixed number of constraint variables for simulated annealing (SA) spatial micro-simulation.

2) to ascertain ability to predict values of the non-constrained attributes by using a Monte Carlo experiment to compute the predictive accuracies of the methods for those variables not included as constraints.
3) to assess the effect of sample-ratio on simulated population by using a Monte Carlo experiment to apply different random sample ratios ranging from 0.05%, to 85%.
4) and to identify the influence of the nature of the conditional distribution of a seed sample by using a Monte Carlo experiment to recreate by sampling such scenarios where the seed data is increasingly structurally dissimilar to the target data.

4.1. Effect of number of constraints

The particular choice set of number of constraint variables would vary, as for instance there are $8P_3$ ways of choosing a set of three variables from eight options (assuming order is also important). To capture the full range of possible choices and thereby objectively establish the effect of number of constraint variables, a Monte Carlo sampling was implemented on the choice set of constraints, and on each sample occasion a distinct run of the microsimulation procedure took place. The densities of the *TAE*'s produced are displayed in Fig. 3 and Fig. 4. The density plots are normalized frequency distributions of the *TAE* values for each set of constraints. In the deterministic IPF, the scenario of 1 constraint yields a particularly high TAE which detracts the display, hence it has been excluded.

The results show that as the number of constraints increase, the *TAE* (across all constrained and unconstrained variable reduces as indicated by a drift of the density plots towards the origin as constraints increase. The deterministic methods have higher overall *TAE* values, as a full volume of the population is not generated whenever the seed sample is not fully reflective of the variety in the aggregate population. The stochastic strategies on the other hand produces a full population, with a broader range of results with lower *TAE* values, albeit yielding simulated results not reflective of the true distribution of the target population. These results may seem counter intuitive at first sight as reported in literature (Markham et al., 2017; Tanton & Edwards, 2012; Tanton & Vidyattama, 2010), as more constraints would imply a more

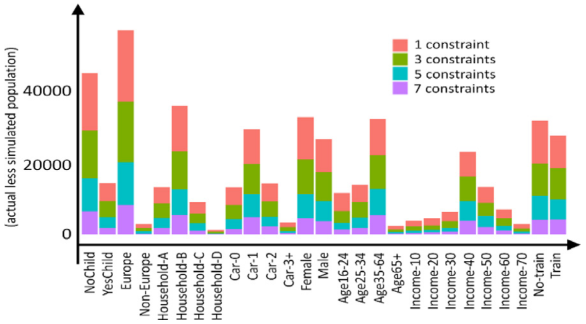


Fig. 5. | Total absolute error (TAE) for different number of constraints in the IPF deterministic spatial micro-simulation.

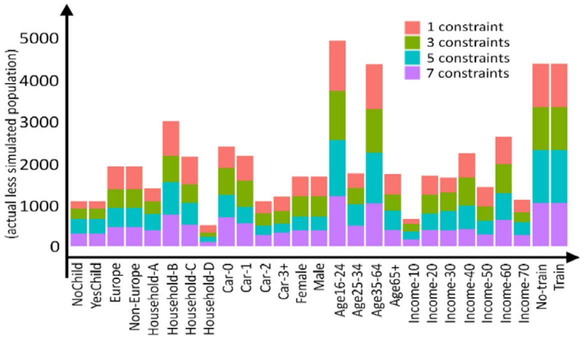


Fig. 6. | Discrepancy in the SA and actual simulated populations.

difficult set of constraints to fulfil, and the *TAE* typically adopted is the sum across constraint tables. However, provided the precautions highlighted in Fig. 2 are adhered to such that the sets of variable categories within the seed and aggregate match, the higher the number of categories would imply a provision for more consistency between the seed and aggregate data, thereby yielding lower TAE values.

4.2. Prediction of non-constrained attributes

The second experiment addresses the question: how well predicted are the values of those variables not included as constraints in the spatial micro-simulation? If a subset of the variables within NRTS is used for spatial micro-simulation, how well are the simulated zonal aggregates predicted for those variables not included in the set of aggregate constraints used in the micro-simulation procedure? In Fig. 5 and Fig. 6, 1 constraint refers to the categories within the first listed variable i.e. household children (with the two categories 'NoChild' and 'YesChild'). 3 constraints similarly refer to the categories within the first three variables of household children, ethnicity and household type. These categories are NoChild, YesChild, Europe, Non-Europe, Household-A, B, C and D). Similarly 5 constraints refers to the categories within the first five variables (on the horizontal axes of Fig. 5 and Fig. 6), i.e. household children, ethnicity, household type, cars and gender. A similarly reference is made for the case of 7 constraints. In Fig. 5 for the deterministic procedure, whilst the trend of the lines for the different number of constraints seems systematic, with higher constraint lines taking lower positions reflective of the lower *TAE* magnitudes for these lines, there are no systematic trends recorded for those variables not included as constraints.

For the stochastic procedure, it is observed that when the constrained and non-constrained variables are compared, the non-constrained ones are predicted with comparatively lower levels of accuracy. In Fig. 6, when only one variable is constrained (depicted by the first two orange bars), the errors are minimal for that constrained variable (in this case the household children variable). This is reflected in the first two red *TAE* bars having lowest values at locations between variables 'NoChild' to 'YesChild'. When three (3) constraints are adopted made up of household child, ethnicity and household cars, the *TAE*'s are depicted by the green bars associated with these variable categories. This is reflected by the height of the green bars over the range of categories 'NoChild' to 'Household-D'. This trend is continued when there are five (5) and seven (7) constraint variables, reflecting in the blue and the purple TAE bars having lowest sizes at locations on the x-axis between categorical variables 'Car-0' to 'Male', and 'Income-10' to 'Income-70' respectively.

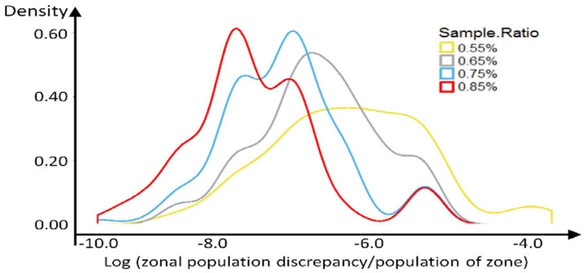


Fig. 7. | Effect of sample ratio on TAE estimates for IPF simulation.

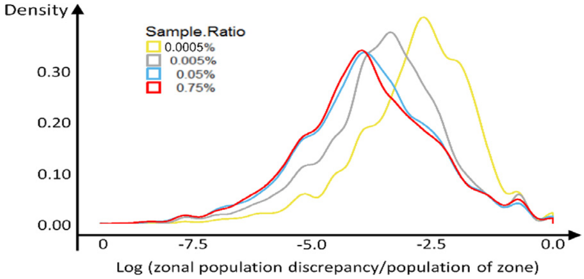


Fig. 8. | Effect of sample ratio on TAE estimates for SA simulation.

4.3. Effect of sample-ratio on simulation

A Monte Carlo sampling was implemented to choose various seed samples of fixed sample ratio to form the individual-level seed data for spatial micro-simulation. The procedure consisted of randomly selecting a sample with ratio v, ($0.05\% \leq v \leq 85\%$) from the full set of NRTS data, and repeating this for 250 times for each value of v, to capture and average out any variability due to the choice of particular individuals in the sample. On each occasion the difference between the simulated and actual populations (*TAE*) is computed, yielding the plots in Fig. 7 and Fig. 8. This is repeated for the deterministic and stochastic spatial micro-simulation strategies.

The *TAE* plot of Fig. 7 below shows that the accuracy of the simulated population from the deterministic method increases with sample ratio, indicated by a drift towards the left of the plots as sample ratio increases from 55% to 85%. The deterministic strategy is found to be highly susceptible to sampling ratios. The procedures fail whenever the sample ratios are lower than 25%, since at these values many of the lower population zones do not have adequate representation in the sample, reflecting where not enough samples were taken to represent zones. For the stochastic strategies, as the percentage of samples increases, Fig. 8 shows that the scatter in the *TAE* increases relative to the *TAE* scatter observed for the deterministic procedure. The *TAE* values in the stochastic procedure also show a reduction as the sample ratio is increased. The stochastic procedures are robust to low sample ratios with the procedures running successfully for the clipped NRTS dataset, for sample ratios as low as 0.05% (as seen in Fig. 8).

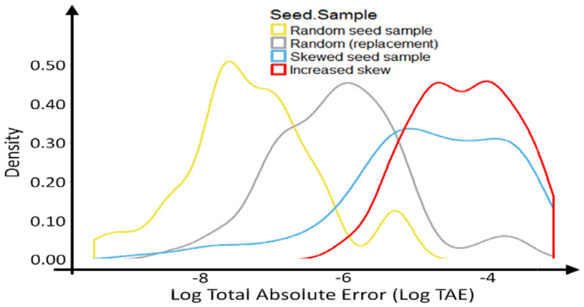


Fig. 9. | Effect of randomness of sample seed on simulated population TAE using a deterministic strategy.

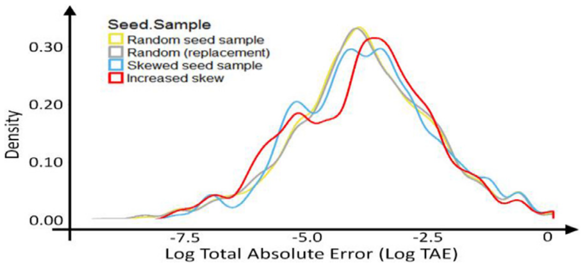


Fig. 10. | Effect of randomness of sample seed on simulated population TAE using the SA stochastic spatial micro-simulation.

4.4. *Influence of level of structural dissimilarity in seed sample*

The spatial micro-simulation process typically creates a simulated representative population by combining zonal aggregate data like the Census, with an individual-level survey like the NRTS or LENNON ticketing data. Traditionally it is assumed that the disparity in the datasets is normally distributed (Ireland & Kullback, 1968; Namazi-Rad, Mokhtarian & Perez, 2014). The robustness of the different deterministic and stochastic strategies are assessed, when for instance the seed sample is a structurally dissimilar representation of the population, typical of 'big-data' which are exhaustive within a specific coverage, and as such are not representative of the entire population. In addition, the effects of 'not having enough' sample data, (whereby the sample does not include enough information on all the zones) is assessed for both the deterministic and stochastic strategies.

A Monte Carlo experiment is conducted, sampling the seed from the NTRS data and then assessing the performance of the deterministic and stochastic strategies. The sensitivity of deterministic methods to choice of seed is depicted in Fig. 9 showing the seed progressing from a random sample representative of the wider population, to an increasingly non-representative sample achieved by sampling with replacement and then by selecting the first N individuals. As seen from Fig. 10, seeds that are better representative of the wider population produce consistently better *TAE* values, indicative of a more accurate simulated representative population. The stochastic procedure is less susceptible to the quality of the seed data. The *TAE* distribution depicted in Fig. 10 only nominally increases as the seed becomes less random, further buttressing the robustness of stochastic strategies to changes in the seed sample.

4.5. *Differences within deterministic and stochastic methods*

Reference has been made to the range of deterministic and stochastic strategies; however results are presented for just one of each of these strategies, i.e. the multi-iterative proportional filling (m-IPF) deterministic strategy (Barthelemy et al., 2016), and the simulated annealing (SA) stochastic strategy (Kavroudakis, 2015). This is because these ones are now the state-of-the-art implementations of the two strategies. Other variants exist, for instance the least squares, Chi-squares, maximum likelihood deterministic methods, and the hill climbing stochastic procedure. These alternative methods show similar trends, but are less robust to the range of behavioural attributes investigated and hence not discussed in detail. For instance, the m-IPF, Chi-squared, least-squared and maximum likelihood methods when broadly compared revealed stability of the iterative proportional fitting, whilst the Chi squares, maximum likelihood and least squares methods did not converge and required over 500GB RAM and 66 h to run, when compared to extreme cases of the IPF which converged, requiring 136MGB RAM and 35 hrs on a 758 GB Arc3-HPC. Fig. 11 shows a comparison of the hill climbing (HC) and simulated annealing (SA) stochastic methods. Whilst the HC algorithm consistently produces more accurate simulated populations than the SA procedure, the HC strategies are subject to a high number of failures by converging to sub-optimal solutions. This phenomenon has been reported in literature (Williamson et al., 1993) and is associated with the definition of the optimization algorithm part of the HC procedure.

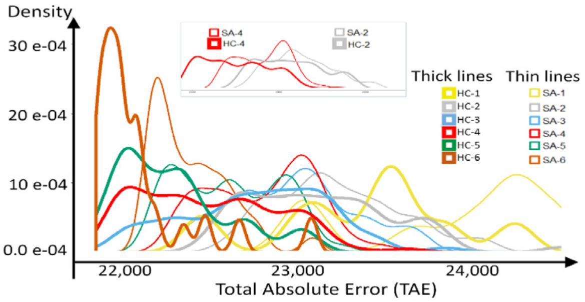


Fig. 11. | Comparison of hill-climbing (HC) and simulated-annealing (SA), both stochastic micro-simulation methods.

5. A case study of mobility on West Yorkshire railways

Having explored the relative merits of the different spatial microsimulation methodologies, the case study illustrates how best to simulate a micro-population linking big-data on rail trip making with information on socio-demographic characteristics. The case study combines the 2011 Census interaction data with the 2001/2004/2005 National Rail Travel Survey (NRTS), and the 2011 LENNON ticketing data to create the mobility behaviour of a population represented at micro-scale. The NRTS does not represent the same year as the Census (and LENNON), and as pointed out earlier in the paper, rail demand levels are likely to be markedly different in different years. Below we detail the basis for adopting the two datasets (2001/5 NRTS and 2011 Census) within the modelling framework of spatial microsimulation, despite being acquired in different years.

Our hypothesis was that disparate datasets like the 2001/5 NRTS and 2011 Census can be combined if they do not present 'concept' or 'data' drift within the model where they are used. We explain this by assessing the use of such datasets within a spatial **interaction** model (Clarke & Birkin, 2018), and then within a spatial **microsimulation** model (Odiari et al., 2021): A spatial interaction model will be of the form: $V_{ij} = P_i\, P_j \,/\, d_{ij}{}^{\beta}$, whereby V_{ij} is volume of spatial interaction, P_i is the population at the origin, P_j is the population at the destination, d_{ij} is the distance between locations i and j, and β is the decline in propensity to travel further distances. If the 2001/5 NRTS and 2011 Census are combined for use in such spatial interaction model, a change in volumes of passengers between the periods the datasets were measured would represents a ’concept drift’ (Žliobaitė, Pechenizkiy & Gama, 2016) and a passenger change in behaviour, resulting in a higher propensity to travel longer distances would represent a ’data drift’ (Hofer & Krempl, 2013).

This is the case because the dependant variable which is the volume of passengers would have changed, and the independent variable which is the distance travelled by passengers both would have changed between the use of the NRTS and Census data. As such, the NRTS and Census cannot be objectively combined for development of a conventional spatial interaction model.

On the other hand, a deterministic reweighting spatial microsimulation model will be of the form: $W_{i+1} = W_i(C_{ij} / S_{ij})$, whereby W_{i+1} represents a new weight for individual i , and W_i is the current weight for individual *i*, C_{ij} is element *ij* within the Census target table, S_{ij} is element *ij* within the NRTS seed table. The model is iteratively executed until a prescribed convergence. In this case there would be no concept or data drift, as neither properties of the dependant nor independent variables in the above spatial microsimulation model qualitatively change. This is the case since the range of attribute categories within the NRTS and the Census are the same. The modelling process aims to replicate the joint distribution of a target population subject to an objective function. As the 2001/5 NRTS and the 2011 Census have the same variable categories, replicating the Census population is limited to the choice of variable categories embedded in the NRTS. From this point of view, there is no concept or data drift.

This explanation is pertinent, forming the basis in this paper, for combining the two time-different NRTS and Census datasets within a spatial microsimulation model. The methodology facilitates replication of the NRTS seed to fit the distribution of a target Census data. Volumetric changes in a seed data are independent of the simulated target. Similarly, particular behavioural changes in the seed data do not affect the simulated target; more so as behavioural attributes of the seed do not form the dependant or independent attributes of interest. In summary, two datasets that are acquired at different times can be combined for use in analytics provided the structural difference in the two datasets does not present as a concept or data drift with respect to the model. This means that the use of either of the 2001/5 NRTS or 2011 Census datasets would yield the same result when applied to a microsimulation model. This is the case here as the NRTS and Census both have the same variable categories. However, the two datasets (2001/5 NRST and 2011 Census) would not be objectively combined in a spatial interaction model which has volume of passengers and distance travelled as variables. The two datasets in such a model would present as a concept and data drift. If either the 2001/5 NRTS seed or the 2011 Census target is used to replicate a population of specific joint distribution, the result is in principle the same. This is the case as the values of the features we define to form the micro-similation remain the same (De-Dios-Santos, 2020).

During the first spatial micro-simulation the Census is combined with the NRTS data based on shared variable attributes as illustrated in Fig. 12 (whereby the colour coded lines are used to distinguish sets of linked variables). For analysing mobility on the railways, it is essential that the distribution and identity of individuals in the simulated population be known at all stages of the procedure. The distribution of individuals is essential as origin and destination attributes combine to form a single unique individual attribute. The individual identity information refers to data ID within the NRTS dataset. For example, the first and second tuples of information in the NRTS would likely have ID of say NRTS-1 and NRTS-2 respectively. Such ID information enables the unconstrained attributes of an individual to be read from the original NRTS table and then appropriately attached within the simulated population. By so doing it is possible to compare for different microsimulation strategies, the volumes of individuals simulated and associated with those variables that were not used as constraints in the spatial microsimulation. Knowledge of individual ID is sustained by including ID information in the Census-NRTS spatial micro-simulation for example, but not imposing any constraints on this ID variable. That way identify information is carried through the spatial microsimulation procedure, and this concept is similarly applied in the second spatial micro-simulation that links in the variables of the LENNON ticketing information as shown in Fig. 12.

5.1. Data pre-processing for micro-scale mobility

As intimated earlier, the first data pre-processing stage involves rezoning the Census data from LSOA's boundaries into Postcode Sectors (or Areas) to enable comparison with the NRTS data and LENNON ticketing data which are geocoded as Postcodes. For data pertaining to interactions, the re-zoning requires special consideration in that re-zoning is performed for the origin variables and then repeated for the destination variables. Regular square fishnet mesh created across the region of interest (West Yorkshire) is used. This in-house procedure precludes the use of ONS lookup tables by dividing LSOA zones into minute fishnet squares which are subsequently integrated over the geography of the particular Postcode boundary. The accuracy of the process is as such decided by the granularity of the squares. In practice, the population proportion of an MSOA and Postcode Sector falling within each fishnet square are calculated.

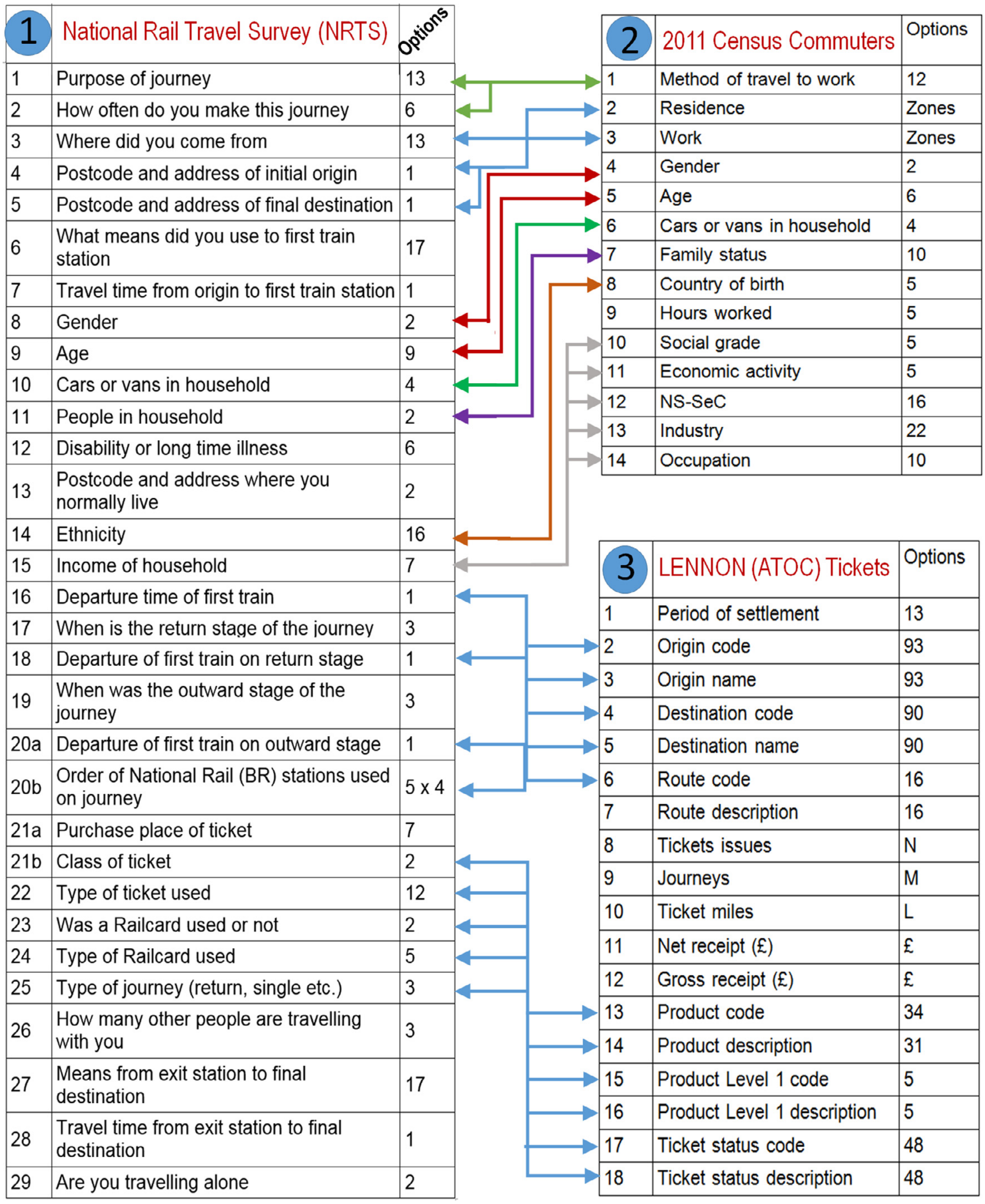
Coarse fishnet zonation is shown in Fig. 13 for the MSOA and Postcode Sectors boundaries associated with West Yorkshire. In practice the fishnets were finer (25 m squares were used in the case study to preclude MAUP[5] phenomena). 25 m was deemed adequate as the unusually small Postcode Sectors are LS155 and LS52 (~8000m^2) with rectangular shapes of about 50 m by 150 m, making a 25 m square fishnet adequate for re-zonation of West Yorkshire Postcode Sectors, bearing in mind also that the average Sector size is 6km^2.

The National Rail Travel Survey (NRTS) and the National Travel Survey (NTS) (Wardman, 2006) show about 60% of rail travel is attributed to commuting. Based on the NRTS Overview Report (DfT-UK, 2010), 77% of all passengers who commute by rail travel 5 or more days a week, whilst an additional 17% of all passengers who commute by rail travel 2–4 days a week. This amounts to a total of 94% of rail commuter travelling at least 2 times a week. The purpose of the journey (commute, business or leisure) is as such decided based on passengers rail travel frequency in addition to their activity at the destination (i.e. normal workplace, going to school/college, etc.), unless that destination is 'home' in which case the purpose is defined by the origin of the trip. These definitions are consistent with the NTS, and form the basis for the analysis of passenger commuting behaviour reported in the National Rail Travel Survey Overview Report (DfT-UK, 2010). These are adopted in this paper despite that the NRTS does not have an explicitly questionnaire on trip purpose.

A further data pre-processing stage concerned establishing a link between journey purposes (and journey frequency) in the NRTS and method of travel to work in the Census (as shown in Fig. 12). We define categories 'RC' and 'NRC' to respectively represent 'rail commuters' and 'not rail commuters' in the NRTS. We then define categories 'TC' and 'NTC' to respectively represent 'train commuters' and 'nom-train commuters' in the Census. We make this distinction in terminology as these categories relate to counts which are derived from different sources. The entire population within the NRTS would fall into either 'RC' or 'NRC', and similarly the entire UK population within the Census would fall into either 'TC' or 'NTC'. Essentially then, 'TC' represents the Census population who indicated that they commute to work by rail, and 'NTC' represents everyone else including those who commute to work by other modes and this does not preclude those who use the rail for other purposes.

As mentioned earlier, within the NRTS, passenger journey purpose and train journey frequency in a week would enable the passenger to be categorised into commuter by train ("RC") if a passenger states so and by definition makes the commute journey up to thrice a week. The only other alternative category would be non-commuters by train ("NRC"),

[5] MAUP is the modified area unit problem is the fallacy whereby the result of data aggregation is dependent on the mapmakers' definition of the geography boundaries. This phenomenon can be alleviated by using point based measures or offsetting by size attributes of the area.



1	National Rail Travel Survey (NRTS)	Options
1	Purpose of journey	13
2	How often do you make this journey	6
3	Where did you come from	13
4	Postcode and address of initial origin	1
5	Postcode and address of final destination	1
6	What means did you use to first train station	17
7	Travel time from origin to first train station	1
8	Gender	2
9	Age	9
10	Cars or vans in household	4
11	People in household	2
12	Disability or long time illness	6
13	Postcode and address where you normally live	2
14	Ethnicity	16
15	Income of household	7
16	Departure time of first train	1
17	When is the return stage of the journey	3
18	Departure of first train on return stage	1
19	When was the outward stage of the journey	3
20a	Departure of first train on outward stage	1
20b	Order of National Rail (BR) stations used on journey	5 x 4
21a	Purchase place of ticket	7
21b	Class of ticket	2
22	Type of ticket used	12
23	Was a Railcard used or not	2
24	Type of Railcard used	5
25	Type of journey (return, single etc.)	3
26	How many other people are travelling with you	3
27	Means from exit station to final destination	17
28	Travel time from exit station to final destination	1
29	Are you travelling alone	2

2	2011 Census Commuters	Options
1	Method of travel to work	12
2	Residence	Zones
3	Work	Zones
4	Gender	2
5	Age	6
6	Cars or vans in household	4
7	Family status	10
8	Country of birth	5
9	Hours worked	5
10	Social grade	5
11	Economic activity	5
12	NS-SeC	16
13	Industry	22
14	Occupation	10

3	LENNON (ATOC) Tickets	Options
1	Period of settlement	13
2	Origin code	93
3	Origin name	93
4	Destination code	90
5	Destination name	90
6	Route code	16
7	Route description	16
8	Tickets issues	N
9	Journeys	M
10	Ticket miles	L
11	Net receipt (£)	£
12	Gross receipt (£)	£
13	Product code	34
14	Product description	31
15	Product Level 1 code	5
16	Product Level 1 description	5
17	Ticket status code	48
18	Ticket status description	48

Fig. 12. | Relational table for the variables from NRTS, Census and LENNON ticket data.

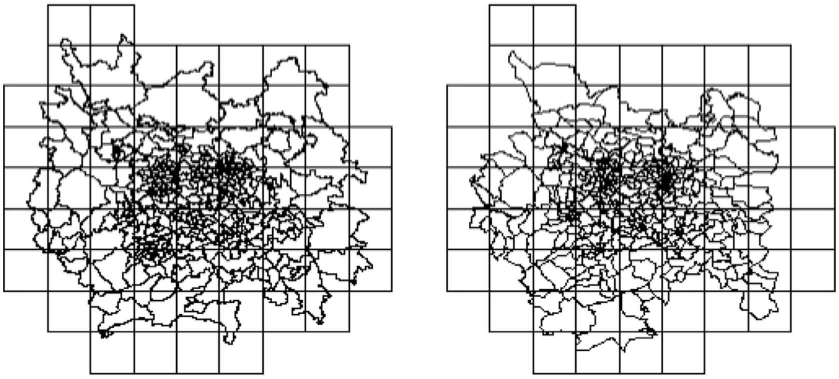


Fig. 13. | Coarse mesh zonation for MSOA's and Postcode Sectors within West Yorkshire UK.

Table 1
Regression results between Office of National statistics (ONS) Small Area Income Estimates against a range of 2011 Census interaction socio-demographic attributes.

Attribute	AIC	Deviance/DoF	R^2 value	Adjusted R^2	p-value
Social grade	8391.9	8.84 $E + 10$	0.8347	0.8324	2.2E-16
Activity	8703.8	25.09 $E + 10$	0.6965	0.6923	2.2E-16
NS-Sec class	8447.3	10.31 $E + 10$	0.8796	0.8737	2.2E-16
Industry	8476.6	11.12 $E + 10$	0.8733	0.8637	2.2E-16
Occupation	8354.1	7.66 $E + 10$	0.8585	0.8541	2.2E-16

Source: Regression parameters from a Generalized Linear Model (GLM) and an LM model.

made up of all the other passengers that do not fall in the "RC" category. The Census in turn has information on method of travel to work, and although there are several modes of travel to work, these can broadly be grouped into commute by train ("TC") and every other mode of travel would then become by definition the non-commute by train ("NTC"). In a proportional sense 'TC' and 'NTC' add up to one, forming the entire UK Census population of which the NRTS is a conditional distribution (i.e. a sub-set). The NRTS as a representative survey and as such can be expanded thus forming a comprehensive (100%) of the UK population who make use of the railways (trains). If that was the case (i.e. 100% expansion), we observe that then counts 'RC' and 'TC' would be exactly the same, and they both represent the same conditional distribution of train commuters. This strategy forms the basis of creating a link between the NRTS and the Census, to complements the more obvious links shown in the relational table in Fig. 13.

In spatial microsimulation, the sample seed is replicated to fit the volume and probability distribution of the target. If the seed has age categories say 11–30yrs, 31–50yrs, 51–70yrs etc., these are essentially multiplied by weights to match the volume and distribution of the target population. The resulting simulated population would also consist of exact same age categories 11–30yrs, 31–50yrs, 51–70yrs etc. A person described as a rail commuter (i.e. 'RC') within the NRTS, could at another time also make a non-commute trip by rail. However, they are categorised as 'RC' within the NRTS and by definition are also 'TC' within the Census. As such, 'RC' and 'TC' are mutually inclusive, and by implication 'NRC' and 'NTC' are similarly mutually inclusive. This forms the basis for relating 'RC' to 'TC' and similarly relating 'NRC' to 'NTC', and in fact adopting the nomenclature 'TC' and 'NTC' within both the NRTS and the Census. During spatial microsimulation the NRTS population within the categories 'RC' and 'NRC' are replicated respectively to yield 'TC' and 'NTC'. For brevity we have simply adopted the use of 'TC' and 'NTC', just like in the example of the age categories given above where we did not distinguish between the 11–30yrs, 31–50yrs, etc. within the NRTS and the categories 11–30yrs, 31–50yrs, etc. within the Census.

5.2. Relationship between income and occupation

Another important pre-processing feature is that we have exploited the relational structure of the dataset attributes in Fig. 12 to derive a disaggregated 'quasi-income' classification for the Census. From the range of variables: 'social grade', 'economic activity', 'NS-Sec', and 'industry' and 'occupation', we check the variable that regresses best with income from the 2011 ONS Small Area Income Estimates (Henretty, 2011/12). It is found that 'occupation' relates best with the NRTS income, such that Census 'occupation' can be used as a substitute disaggregated 'quasi-income' variable. The procedure to achieve the 'quasi-income' variable is that 'social grade', 'economic activity', 'NS-Sec', and 'industry' and 'occupation' variables aggregated from the 2011 Census interaction data (UKDS, 2011) are separately regressed against (ONS, 2016). The results from regression are shown in Table 1 for the variables considered.

Occupation' was found to relate best to the average zonal weekly income, having one of the highest R^2, the lowest AIC (Akaike, 1987), and having the most number of levels within the variable with statistically significant p-values. This formed the basis for combining categories within the occupation variable to inform an estimate of income. The 'occupation' variable also had the added advantage of having only nine variables requiring reconciliation with the seven income categories. The 'NS-Sec' and 'Industry' which have commensurate R^2 values have 15 and 21 variables, making them more difficult to reconcile with the seven income categories. As such, occupation variable was deemed best related, resulting in the classification shown in Fig. 14, yielding nine Census occupation categories (reduced to seven) and commensurate with seven NRTS income categories.

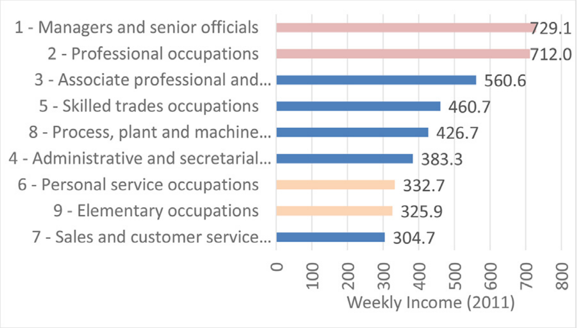


Fig. 14. | Relationship between Income and Occupation categories utilizing the 2011 UK Census.

The blue coloured bars represent a unique occupation bracket. The alternative coloured bars represent groups of occupation categories that are combined into one new occupation category. For example, the personal services and the elementary occupations have been combined into one income bracket, just as the managers and senior officials combined with the professional occupations to form the highest income bracket. As mentioned earlier, this enables the Census occupation categories with seven categories to be related by proxy to the seven NRTS income categories. The map below in Fig. 15 below validates the combinations derived as the income distributions are similar when produced from the small area income estimates and when derived from occupations.

5.3. Microsimulation of passengers in spatial interaction

Another practical consideration is that each of the individuals (say of age25–34 and associated aggregate attribute) are in mobility (by spatially interacting between an origin and destination). As such the individuals age attribute (age25-34 for instance) is associated with a unique origin (out of the N many origins) and a unique destination (out of M destinations). Additional variability is as such inherently introduced to that ag25–34 category due to the association with N origins and M destinations. In essence the single attribute category ag25–34 now has N by M potential variations to it, and so also does all the other attributes and their categories. This multiplicity yields a dataset of high dimensionality, in this case by N by M times. Such increases in granularity in the aggregate dataset requires a commensurate increase in volume of the individual-level seed in order to sustain integrity of the solution (calling for a need for big data). The individual-level seed is in effect reduced in sample ratio by N by M times, commensurately affecting the accuracy and precision of the results due to effectively reduced sample ratios.

A further issue arising from an increased variability due to spatial interactions, is that the number of Census variable categories (~2192) may not be the same as the number of variable categories in the individual-level seed (~1351), indicating that the survey does not have enough data, as the seed sample only represented 1351 variables within the Census. For example, an NRTS variable category LS-NTC-BB which represents a passenger who has the behaviour of not commuting by train (NTC), while travelling between Leeds (LS) and Blackburn (BB). Had the problem not involved spatial interaction and the micro-simulation

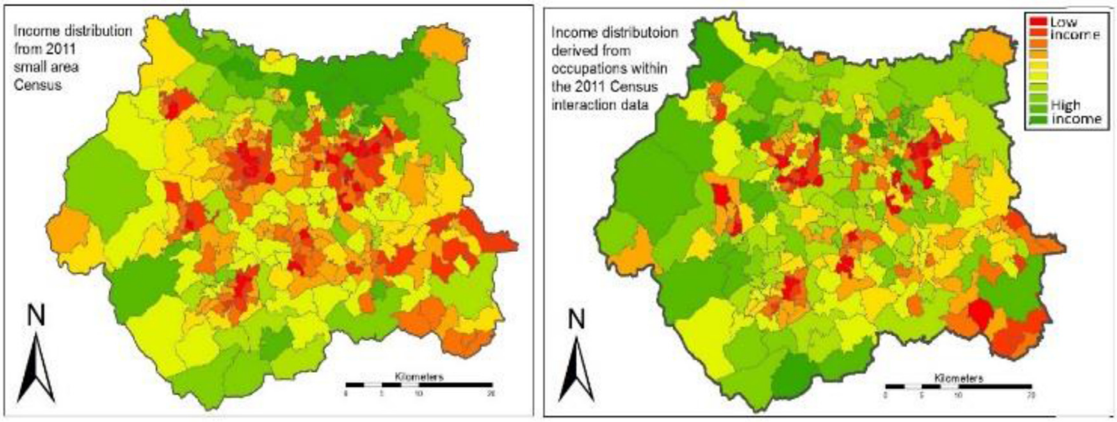

Income distribution from 2011 small area Census
N
Kilometers
0 5 10 20
Income distributoion derived from occupations within the 2011 Census interaction data
Low income
High income
N
Kilometers
0 5 10 20

Fig. 15. | Distribution of quasi-income from Census and income from NRTS for West Yorkshire County.

of resulting origin-destination flows (with the complication of associating each individual attribute to an origin and a destination), then the variable categories would be NTC in both the Census and NRTS, and would as such have been sufficient for identification purposes. Since interaction is involved however, the seed sample would have to include a measure of not just NTC, but also of NTC associated with LS and BB. This subtlety highlights the particular advantage of 'big data' in creating the requisite data volume in spatial micro-simulation of interaction phenomena.

The deterministic spatial micro-simulation strategy copes with insufficient (non-observed) sample seed by synthesizing lower population volumes (associated with when the seed does not have the variable categories to match the aggregate). The stochastic strategy under these circumstances creates a full sub-optimal population. In some instances (as recorded in the 2004 NRTS compared with the 2011 Census), a number of individual flows were not captured in the Census. Under these circumstances it was appropriate to remove such data values from the NRTS, as the Census was the reference aggregate dataset.

5.4. Linking census and NRTS data

The 2011 Census interaction data measures area of usual residence by workplace for a number of separate socio-demographic attributes, see Fig. 12. Linking this Census with the NRTS variables enables the interaction between the separate socio-demographic attributes to be established, creating representative mobility behaviour of a population portrayed at micro-scale with a rich set of attributes. Traditional spatial micro-simulation strategies for linking and combining datasets are typically premised on the seed being representative of the aggregate population, requiring similar distributions between the datasets. The NRTS distribution is a subset of the Census, derived by conditioning on 'travel by rail' variable (i.e. 'NTC' and 'TC'). As such, for validation purposes the resulting simulated populations created by the stochastic and deterministic strategies are assessed to see which one better reflects the distribution of the target Census population.

The R-script for implementation of spatial microsimulation for both the deterministic and stochastic strategies is included in the data file associated with this paper. The results presented in Fig. 16 are crucial, and illustrates that the deterministic m-IPF strategy produces a synthetic population with similar distribution (and median lines) to the target Census. The stochastic strategy on the other hand creates a simulated population distribution (and median line) similar to the NRTS sample seed, indicative that the proposal distribution from a seed that is structurally dissimilar to the target does not evolve to the target Census distribution during the stochastic spatial micro-simulation process. The deterministic strategy is as such preferred for use in spatial micro-simulation when the seed sample is a structurally dissimilar non-representative random sample of the target population.

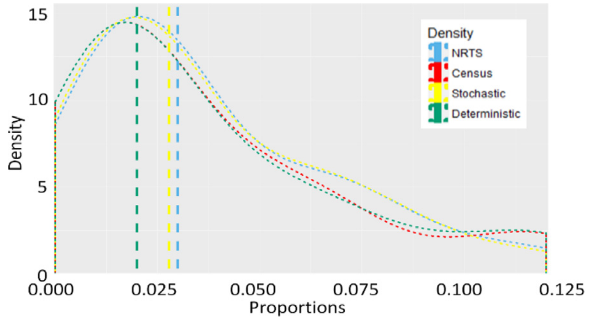


Fig. 16. | Distribution of population attributes for NRTS, Census and populations simulated by stochastic and deterministic methods (the vertical median lines are also indicated).

5.5. Linking-in LENNON ticketing data

Once the Census interaction data and NRTS are combined to create a synthetic population, the challenge then lies in linking-in the LENNON ticketing data. To curtail the computer memory requirements during the first spatial micro-simulation, only the relevant variables in the NRTS data where included (these are variables indicated in the top half of the NRTS table in Fig. 12. To enable a coupling to the LENNON tickets, those NRTS variables in the bottom half of the NRTS table in Fig. 12 which relate to the LENNON data have to be re-attached to that simulated population created by the first spatial micro-simulation. The attachment is achieved by including the identity (Ind-ID) information for each seed sample in the NRTS dataset, during the spatial-microsimulation. Although the identity is included as a seed variable, it is not constrained by the target aggregate. That way the simulated NRTS-Census data includes the distribution of the individual identity (Ind-ID) within the spatial micro-population. (The R script implementation of this is available on request).

The m-IPF deterministic spatial micro-simulation strategy has been adopted as this has been shown in the last section (Fig. 16) to be suitable for combining datasets when the seed is non-representative by not having the same distribution as the target data. The results shown in Fig. 17 show the distribution of the simulated NRTS-Census-LENNON mobility population, for two cases: first when the simulated (NRTS-Census combined with LENNON) population is sampled with probability distribution equal to the spatial micro-simulation weights. As seen from the top two plots, the population created has a distribution of variable cate-

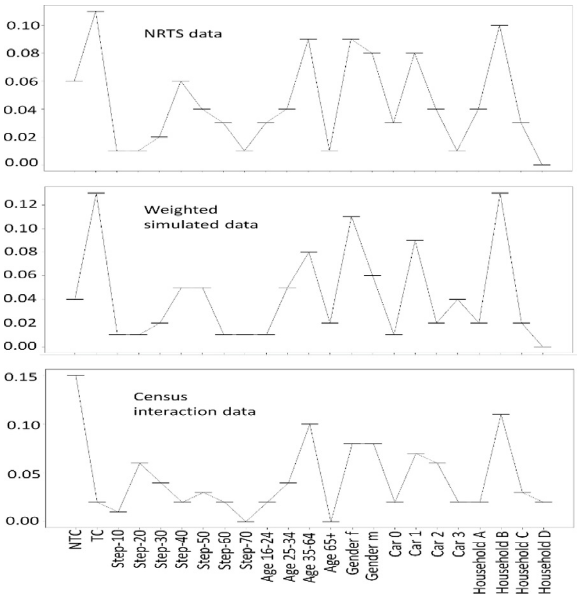


Fig. 17. | Variable categories for NRTS, simulated population and Census showing similarities in distribution of variable categories.

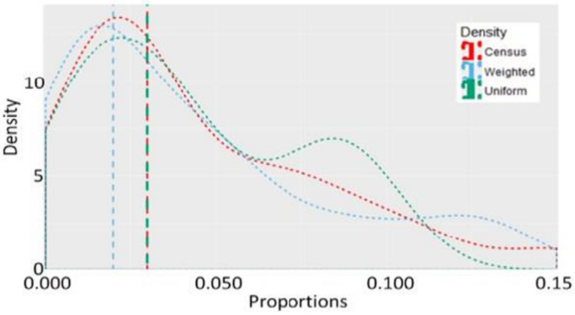


Fig. 18. | Density of variables of simulated population (weighted), uniformly sampled, and Census populations.

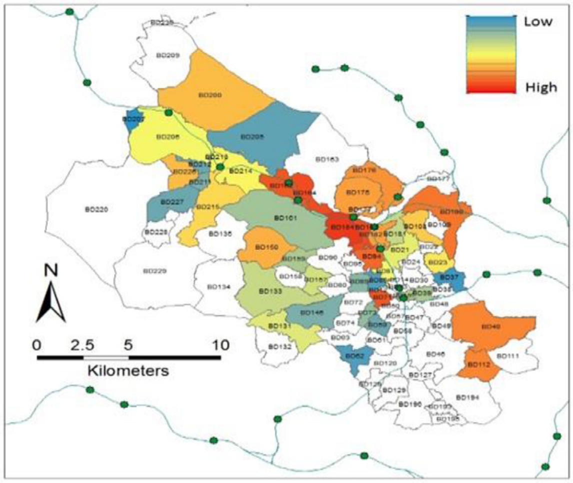


Fig. 19. | Population in the BD Postcode Area who use the train service and are in closer proximity to the train lines as seen in the map.

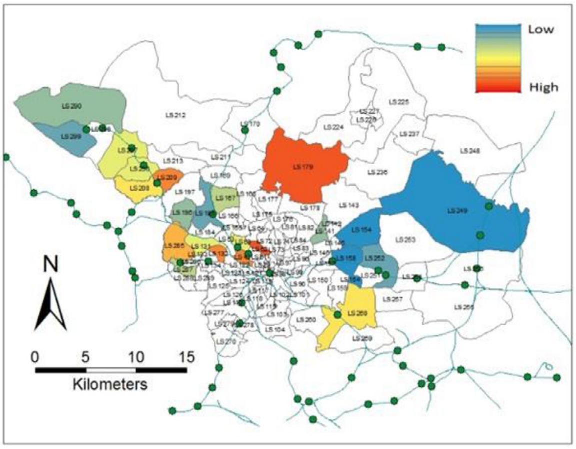


Fig. 20. | Granular query results for Postcode Sectors in the LS Postcode Area, illustrating results from spatial micro-simulation.

gories similar to the NRTS. In particular notice that the NTC is much lower than the TC typical of a population of railway passengers where there are more commuters (TC) than those who do not commute by train (NTC). The lower plot (third from top in Fig. 17) of the distribution of the Census variables shows a noticeable distinction and there are much higher values of NTC than TC, reflective of the wider population where typically only 5% of the population commute by train (TC) whilst the rest commute by alternative modes (NTC).

A uniform (non-weighted[6]) sample of simulated population produces a micro-population with variable distribution akin to that of the wider Census population. This is illustrated in the density plot of Fig. 18 with the median of the uniformly sampled population being similar to that of the Census, but distinct from population derived using the simulation weights. The results produced validates the m-IPF spatial micro-simulation methodology as the simulated population replicates the railways population and then the wider population dependant on whether systematic weighted sampling or uniform sampling is adopted. Representative populations of rail passengers that can be fed through a logistical railways system are created by the weighted systematic sampling, making up a volume equal to the number of LENNON tickets sold.

Validation of cross-tabulated micro-data created

Typical cross-tabulations resulting from the micro-population created are shown in the maps of Fig. 19 and Fig. 20 below. Prior to spatial micro-simulation, only a sub-set i.e. a sample of cross-tabulated data is available as the seed. The aggregate data only reveals a global cross tabulation limited to only three variables from a range of socio-demographic Census attributes. Spatial micro-simulation combines the various attributes in the disparate datasets, and produces a granular cross-tabulation of the variables. The map in Fig. 19 for example is the result of a query on the cross tabulated micro-population, showing the proportion of people residing in Postcode Sectors in Bradford, commute to work by rail, have a Rail card, regularly buy a return ticket, travel within 15 miles of their typical residence, earn between £17.5 – £35k (at 2011 rates) per annum, and live in a household with no car and no children. It is seen that Postcode Sectors in the vicinity of the railways stations (the network), expectedly tend to have a higher number of passengers as they have easier access to the railways network. This is seen in the left hand side Sectors within the Area, and consist of those passengers who are more likely to use the train based on the proximity and access to rail service from the usual residence.

[6] The weights derived from the second spatial micro-simulation represent probabilities, such that a random sample taken from the simulated population using the probability distribution would yield the population of rail passengers (with a distribution akin to the NRTS), whilst a sample taken assuming uniform probability for each simulated population would yield the population of passengers who commute and those who do not commute by rail (with a distribution akin to the Census interaction).

Another result made feasible by the availability of the cross-tabulated micro-population is shown in Fig. 19, showing the proportion of people who reside in Postcode Sectors in Leeds (LS), who do not commute to work by rail (NTC), and live in a household with three or more cars (Car-3+). The results are intuitive and this provides some external validation of the micro-simulated population. These passengers use the train not-for-commute purposes (NTC), and the high volume of passengers simulated for the LS179 Sector near the middle of the map (Fig. 20) are perhaps reflective of an affluent neighbourhood, occupied by households with 3 or more cars (Car-3+), and who do not commute by train (NTC).

6. Synopsis and discussion

This paper contains novel analysis detailing the relative merits of different methodologies for spatial microsimulation. Then by using a case study to simulate a micro-population linking consumer data on rail trip-making with information on socio-demographic characteristics, the robustness of the methodologies developed are tested for application to big-data from a disparate source.

The assessment of the behaviour of deterministic and stochastic strategies under different circumstances enables an informed choice of spatial micro-simulation method for specific applications. The results review previously available research that indicates that the more variables used in constraining a spatial microsimulation procedure, the worse the results. This is the first time the Monte Carlo technique is used to select problem scenarios for the spatial micro-simulation. By so doing, the full range of scenarios are included in the analysis, producing robust and conclusive results that reflect better the *TAE* distribution range of possible solutions. The Monte Carlo selection process is quite distinct from Monte Carlo procedures applied within some spatial microsimulation procedures. Stochastic spatial micro-simulation methods held promise because of the use of MCMC type algorithms. However, the results from the SA and HC indicate that the exploratory and optimization routines used therein need further development. For cases where the seed data is a conditional distribution and non-representative random sample of the target distribution, the proposal densities in stochastic spatial micro-simulation methods have not converged to the target distribution, despite the definition of TAE as the objective function. The internal validation of the procedures is implied by the *TAE* values derived; however, this validation heavily impinges on the data pre-processing stage, whereby the variable categories in the seed and aggregate are required to match.

This paper marks the first case study of the application of spatial micro-simulation to the spatial interaction phenomena within the railways. The particular difficulties of dealing with non-representative railways ticketing consumer datasets (big-data) have been addressed. Such data are a conditional distribution of the wider Census population and as such do not have the distribution of a representative random sample of the population, thereby presenting challenges in use for spatial microsimulation. The high dimensionality and cross-variability in passenger attributes can only be achieved by spatial micro-simulation, thus enabling the Census demographic attributes to be combined with network variables within the NRTS and LENNON databases, showing previously unavailable variability in passenger attributes. External validation is indicated from the intuitive results of queries on the synthetic population, which shows that passengers residing or working in the vicinity of the rail stations inherently have a higher propensity of using the railways.

A limitation of the case study presented in this paper is that passenger flows considered are simply those emanating from and ending in West Yorkshire. As such interregional flows are excluded, implying that about 60% of actual flows have been analysed. This may affect some interesting boundary phenomena like rail-heading, whereby passengers travel further afield to access the rail service across a rail-zone boundary, in order to restrict travel to one zone and thereby benefit from cheaper within-zone fares. The micro-simulation developed in this paper is applied to rail passengers who are in spatial interaction between origin and destination points. Conventional microsimulation creates a synthetic population that is fixed in space by being associated with just one location. Hence conventional microsimulation re-creates households, zone populations, shoppers, etc. In this paper however, we are creating a synthetic population interacting (by mobility) between a journey origin and destination. These result in our use of the term ‘mobility interaction’ as the simulated synthetic population created are in mobility by virtue of spatial interaction between origin and destination points. The consequence of this mobility interaction is to effectively increase the dimensionality of the sample seed which is disaggregation into origin-destination pairs. The implication of this for microsimulation is a requirement for a commensurate increase in sample seed volume to maintain the variability of the simulated population and preclude the simulation of clusters of individuals. This augurs well for the advantages of increasingly available volumes of big data (e. g. mobile phone data on mobility flows) which when used as seeds in interaction analysis; resolve such dimensionality reduction issues and further enables the exploration of time-series cross-section (TSCS) phenomena associated with the nature of the big data phenomena.

There are current limitations on the scalability of current implementation of m-IPF strategies. This mainly results from an increase in RAM requirement as the number of covariates increase in spatial microsimulation. The particular implementation presented in the case study in this manuscript requires two repeats of the m-IPF procedure, the first between the NRTS and Census, and the second to incorporate the LENNON ticketing data. In order to link the results from each step, ID information on each simulated passenger has to be incorporated, thereby increasing the memory requirements for the solution. A further limitation is due to the R-Studio solution platform adopted which limits the size of data tables to 2ˆ31. Parallel computing strategies have the potential to limit some of these constraints; however these have not yet been explored in this research.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by a grant from the Consumer Data Research Centre (CDRC), Economic and Social Research Council (ESRC), Grant Number ES/L011891/1 and by funding from the Association of Train Operating Companies (ATOC).

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317–332.

Asmussen, S., & Glynn, P. W. (2007). *Stochastic simulation: Algorithms and analysis*. Springer Science & Business Media.

Bacharach, M. (1970). Biproportional matrices and input-output change, CUP Archive.

Ballas, D., & Clarke, G. P. (2001). Modelling the local impacts of national social policies: A spatial microsimulation approach. *Environment and Planning C: Government and Policy, 19*(4), 587–606.

Ballas, D., Kingston, R., Stillwell, J., & Jin, J. (2007). Building a spatial microsimulation-based planning support system for local policy making. *Environment and Planning A, 39*(10), 2482–2499.

Barthelemy, J., Suesse,.T., Namazi-Rad,.M., & Barthelemy, M.J. (2016). Package ‘mipfp’.

Bertsekas, D. P. (2014). *Constrained optimization and lagrange multiplier methods*. Academic press.

Birkin, M., & Clarke,.G. (1995). Using microsimulation methods to synthesize census data. Census users’ handbook: 363–387.

Birkin, M., & Clarke, M. (1988). SYNTHESIS—A synthetic spatial information system for urban and regional analysis: Methods and examples. *Environment and planning A, 20*(12), 1645–1671.

Birkin, M., & Clarke, M. (1989). The generation of individual and household incomes at the small area level using synthesis. *Regional Studies, 23*(6), 535–548.

Birkin, M., & Clarke, M. (2011). *Spatial microsimulation models: A review and a glimpse into the future. population dynamics and projection methods* (pp. 193–208). Springer.

Birkin, M., Turner, A., & Wu, B. (2006). A synthetic demographic model of the UK population: Methods, progress and problems. *Regional Science Association International British and Irish Section, 36th Annual Conference*.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Clarke, M., & Birkin, M. (2018). Spatial interaction models: From numerical experiments to commercial applications. *Applied Spatial Analysis and Policy, 11*(4), 713–729.

De Montjoye, Y.-. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports, 3*, 1376.

De-Dios-Santos, J. (2020). Understanding and Handling Data and Concept Drift. *Data Science*.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427–444.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

DfT (2013). National Rail Travel Survey – Overview report. Rail statistics. D. f. Transport. www.gov.uk, Department for Transport.

DfT-UK (2010). National Rail Travel Survey NRTS Overview Report. NRTS Overview Report. D. f. Transport. https://www.gov.uk/government/statistics/, Department for Transport.

Dimitriou, L., Tsekeris, T., & Stathopoulos, A. (2006). Genetic-algorithm-based micro-simulation approach for estimating turning proportions at signalized intersections. *IFAC Proceedings Volumes, 39*(12), 159–164.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological, 58*, 243–263.

Fermat, P.d. (1891). Oeuvres de Fermat, ed. Charles Henry and Paul Tannery 5: 1891-1922.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Fratar, T. J. (1954). Vehicular trip distribution by successive approximations. *Traffic Quarterly, 8*(1).

Furness, K. (1965). Time function iteration. *Traffic Engineering and Control, 7*(7), 458–460.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*(412), 972–985.

Gower, L. (2021). Table 1222 - Passenger journeys by ticket type. April 1986 to December 2021. Office of Rail and Road, UK, Office of Rail and Road, UK.

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 Artifical Societies and Social Simulation, 15*(1), 1–15.

Henretty, N. (2011–2012). *Small area income estimates for middle layer super output areas*. England and Wales, Office for National Statistics.

Hofer, V., & Krempl, G. (2013). Drift mining in data: A framework for addressing drift in classification. *Computational Statistics & Data Analysis, 57*(1), 377–391.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., et al., (2008). Big data: The future of biocuration. *Nature, 455*(7209), 47–50.

Ireland, C. T., & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika, 55*(1), 179–188.

Kavroudakis, D. (2015). sms: Microdata for Geographical Analysis in R. *Journal of Statistical Software, 68*(2), 1–23.

Kavroudakis, D., Ballas, D., & Birkin, M. (2008). *Using Spatial Microsimulation for the analysis of social and spatial inequalities. Studying, Modeling and Sense Making of Planet Earth International Conference. Presented at the Studying, Modeling and Sense making of Planet Earth International Conference*. Department of Geography, University of the Aegean.

Kelley, C.T. (1999). Iterative methods for optimization, SIAM.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), Article 2053951714528481.

Kurban, H., Gallagher, R., Kurban, G. A., & Persky, J. (2011). A beginner's guide to creating small-area cross-tabulations. *Cityscape (Washington, D.C.)*, 225–235.

Lagrange, J. L. (1867). *Oeuvres de lagrange (14 vols)*. Paris: Gauthier-Villars.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Sage Publications.

Le Vine, S., Polak, J., & Humphrey, A. (2017). *Commuting trends in england 1988-2015*. London: Department for Transport.

Lomax, N., & Norman, P. (2016). Estimating Population Attribute Values in a Table:"Get Me Started in" Iterative Proportional Fitting. *The Professional Geographer, 68*(3), 451–461.

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., & Dumont, M. (2016). *Spatial microsimulation with r*. CRC Press.

Lovelace, R., Dumont, M., Ellison, R., & Založnik, M. (2017). *Spatial microsimulation with r*. Chapman and Hall/CRC.

Lynch, C. (2008). Big data: How do your data grow? *Nature, 455*(7209), 28–29.

Manyika, J., Chui,.M., Brown,.B., Bughin,.J., Dobbs,.R., Roxburgh,.C. et al. (2011). Big data: The next frontier for innovation, competition, and productivity.

Markham, F., Young, M., & Doran, B. (2017). Improving spatial microsimulation estimates of health outcomes by including geographic indicators of health behaviour: The example of problem gambling. *Health & Place, 46*, 29–36.

Müller, K., & Axhausen, K. W. (2010). *Population synthesis for microsimulation: State of the art* (p. 638). Vancouver: Arbeitsberichte Verkehrs-und Raumplanung.

Namazi-Rad, M.-. R., Mokhtarian, P., & Perez, P. (2014). Generating a dynamic synthetic population–using an age-structured two-sex model for household dynamics. *PloS one, 9*(4), E94761.

Odiari, E. A. (2018). *A framework for big data in urban mobility and movement patterns analysis*. University of Leeds.

Odiari, E., Birkin, M., Grant-Muller, S., & Malleson, N. (2021). *Spatial microsimulation models for rail travel: A west yorkshire case study. big data applications in geography and planning*. Edward Elgar Publishing.

ONS (2016). Small area income estimates for middle layer super output areas, England and Wales. Earnings and working hours. www.ons.gov.uk/, Office for National Statistics.

ONS-UK (2013). 2001 to 2011 Census method of travel to work, Office for National Statistics (ONS). UK Open Government Licence (OGL v3).

ORR. (2016). Passenger Rail Usage: Quality and Methodology Report, Passenger Rail Usage Retrieved 4th April 2017, from www.orr.gov.uk/.

Rees, P. (1994). Estimating and projecting the populations of urban communities. *Environment & planning A, 26*(11), 671–697 1.

Rees, P., Martin, D., & Williamson, P. (2002). Census data resources in the United Kingdom.

Smith, D. M., Clarke, G. P., & Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A, 41*(5), 1251–1268.

Stillwell, J., & Duke-Williams, O. (2003). A new web-based interface to British census of population origin–destination statistics. *Environment and Planning A, 35*(1), 113–132.

Sun, W., & Yuan, Y.-X. (2006). *Optimization theory and methods: Nonlinear programming*. Springer Science & Business Media.

Tanton, R. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation, 7*(1), 4–25.

Tanton, R., & Edwards, K. L. (2012). *Limits of static spatial microsimulation models. spatial microsimulation: A reference guide for users* (pp. 161–168). Springer.

Tanton, R., & Vidyattama, Y. (2010). Pushing it to the edge: Extending generalised regression as a spatial microsimulation method. *International Journal of Microsimulation, 3*(2), 23–33.

Team, R. C. (2016). R: A language and environment for statistical computing. *Vienna: R Foundation for Statistical Computing;, 2014*.

UKDS (2011). Office for National Statistics, 2011 Census. Flow data. O. f. N. Statistics. https://census.ukdataservice.ac.uk/use-data/citing-data/2011.

Upton, G.J. (1985). Modelling cross-tabulated regional data. Nijkamp P, LeitnerH, Wrigley, N Martinus, Measuring the Unmeasurable: 197-218.

Wardman, M. (2006). Demand for rail travel and the effects of external factors. *Transportation Research Part E: Logistics and Transportation Review, 42*(3), 129–148.

Weber, G. M., Mandl, K. D., & Kohane, I. S. (2014). Finding the missing link for big biomedical data. *JAMA, 311*(24), 2479–2480.

Whitworth, A., Carter, E., Ballas, D., & Moon, G. (2017). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. *Computers, Environment and Urban Systems, 63*, 50–57.

Williamson, P., Birkin, M., & Rees, P. (1993). *The simulation of whole populations using data from small area statistics, samples of anonymised records and national surveys. research on 1991 census conference'*. University of Newcastle September.

Wong, D. W. (1992). The Reliability of Using the Iterative Proportional Fitting Procedure*. *The Professional Geographer, 44*(3), 340–348.

Zhu, Y., & Ferreira Jr, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record, 2429*(1), 168–177.

Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. *Big data analysis: New algorithms for a new society*, 91–114.