



Spatial microsimulation of osteoarthritis prevalence at the small area level in England – Constraint selection for a 2-stage microsimulation process

Onosi Sylvia Ifesemen^{1,2*}, Thomas Bestwick-Stevenson^{1,2*},
Kimberley L. Edwards^{1,2*}

¹Centre for Sports, Exercise and Osteoarthritis, Versus Arthritis, Nottingham, United Kingdom; ²Academic Orthopaedics, Trauma and Sports medicine, School of Medicine, Queens Medical Centre, University of Nottingham, Nottingham, United Kingdom

Abstract The presence of identical benchmark/constraint variables in both geographic and survey datasets is a principal requirement for static spatial microsimulation models, particularly in the field of medicine and health sciences. This is also a key limitation of static spatial models because geographical datasets rarely contain all variables required to realistically simulate an outcome. We believe this challenge can be overcome by a multilevel approach to spatial microsimulation using a case study of estimating the small area level prevalence of knee osteoarthritis in England. In the paper, we describe constraint selection and demonstrate a novel two-stage spatial microsimulation procedure using SimObesity, a static deterministic combinatorial spatial microsimulation model. We also present the validation parameters of our synthetic data, important areas for consideration and avenues for future research. Our findings demonstrate that important benchmark variables absent from the geographical dataset can be incorporated into spatial microsimulation models without compromising model robustness. **JEL classification:** I100

DOI: <https://doi.org/10.34196/ijm.00201>

***For correspondence:** Onosi. ifesemen@nottingham.ac.uk; thomas.bestwick-stevenson1@nottingham.ac.uk; Kimberley.edwards@nottingham.ac.uk

Competing interest: See page 13

©<http://creativecommons.org/licenses/by/4.0/>This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Author Keywords: Spatial Microsimulation, Deterministic, Knee Osteoarthritis
© 2019, Ifesemen et al;
DOI: <https://doi.org/10.34196/ijm.00200>

1. Background

Spatial microsimulation methodologies represent a powerful tool used to create and analyse spatially disaggregated data without expending the vast resources required to obtain these data through primary data collection. Spatial microsimulation models could be static (looking at just one point in time), or dynamic (including projections overtime), and could be deterministic (does not incorporate random variability) or stochastic (incorporating random variability). There are several published reviews on the description, strengths and weakness of various spatial microsimulation methodologies (O'donoghue et al., 2014; Tanton, 2014; Birkin and Clarke, 2011; Rahman et al., 2010; Harland et al., 2012). This paper focuses on static deterministic spatial microsimulation models.

Static deterministic spatial models have the added advantage of the ability to conduct powerful counterfactual scenario modelling (Tanton, 2014). This is a modelling technique where the effect of input parameter manipulation can be assessed in the derived synthetic dataset. For example how would a 20% increase in the population of people in the lowest deprivation quintile affect the prevalence of obesity in a given geographical area? This can provide policy makers and researchers an insight into the potential effect of an intervention or policy across various geographical areas before the intervention/policy is implemented. This process could be described as mimicking the effect of a

randomised control trial, without the attendant time and resources required to conduct one in real-time (*Prakash et al., 2017*). With such a powerful tool at our disposal and with the improvement in computational power making it easier to carry out complex calculations quickly and more effectively, it is no surprise that the number of published literature on applications of spatial microsimulation/microsimulation have increased exponentially in the past few years. These models have found use in various disciplines ranging from transport, economics, fiscal policy and even healthcare and demography (*Smith et al., 2011; Ballas et al., 2006; van Leeuwen and Dekkers, 2013; Lovelace et al., 2014; Lymer et al., 2008; Rephann and Holm, 2004*).

That being said, penetration of spatial microsimulation modelling into core medical and epidemiological research has been rather slow. We recognise that microsimulation methodologies have been used extensively in infectious disease epidemiology however these models typically dynamic and conducted at much coarser geographical levels and therefore lacking the small areal level spatial component (*Habbema et al., 1996; Morris and Kretzschmar, 2000*). Also, available published literature on spatial models used to simulate health conditions like smoking, obesity and osteoarthritis, these studies have found little use outside the microsimulation community (*Edwards and Clarke, 2009; Smith et al., 2011; Cataife, 2014*). We believe that this may be due to an inherent weakness in static, deterministic spatial microsimulation methodologies – Data limitations (*Tanton and Edwards, 2012*). This could either be differences in variable definitions in the geographical and population dataset or complete absence of certain predictor variables in the geographic and/or population dataset. The latter is of particular importance as it may affect the accuracy of the simulated dataset and constrain further analysis of the derived microdata (*Cassells et al., 2013*). This is explained as follows – in order to generate microdata at the required geographical scale typically two datasets are required, the population dataset (which contains the outcome of interest) and the geographical dataset (which contains the geographical identifier). Here, observations (in this case people) are assigned to geographical areas based on matched attributes shared between the two datasets for example the geographical dataset (usually the census) and the population dataset can be matched based on sex, age and ethnicity (*Ballas et al., 2005*). This means that these attributes selected should be associated with the outcome of interest and must be present in both datasets (*Edwards and Clarke, 2009*).

Demographic variables such as age, sex, ethnicity and social class may be sufficient to simulate certain outcomes however simulation of medical conditions generally require more information and this information that may not be collected in one single survey. For example, it may not be accurate to simulate the spatial prevalence of lung cancer based on age and sex alone, without collecting smoking history because smoking is a well-documented risk factor for lung cancer (*Doll and Hill, 1950; Doll and Hill, 1956; Lee et al., 2012*). However, smoking data may not be available in the geographical dataset as only a few countries collect small area level smoking data. This means that it may not be possible to conduct an accurate spatial microsimulation model of lung cancer or convince policy makers about the validity of the resulting dataset.

Several methods have been employed to overcome this challenge ranging from imputation (*Cassells et al., 2013*) in the case of missing benchmark variables in the survey dataset, to merging geographically disaggregated data from another source with census data (where important predictor variables are missing from the census data) (*Edwards and Clarke, 2009*). Although these methods were appropriate in the situations in which they were used, and simulated outputs were found to be robust, these methods may not be applicable in all scenarios for example in the case of lung cancer mentioned above where smoking history data is unavailable in the census dataset and not accessible from other sources.

1.1. Study aim/objectives

The aim of this paper is to introduce a 2-stage approach to spatial microsimulation for situations where a key predictor variable is absent from the geographical dataset, using a case study of knee osteoarthritis (OA) prevalence in England. The objectives are to select constraints, generate and validate small area level knee OA prevalence data for England using this two-stage spatial microsimulation procedure.

The rest of the document includes a brief introduction to the medical condition (knee OA), an account of the methods used to generate and validate the synthetic data, and finally discussions around the strengths and limitations of this technique and areas of further research.

2. Case study

2.1. Introduction

Osteoarthritis is a debilitating joint condition affecting over 8 million people aged 45 years and over in the United Kingdom. It presents with symptoms such as pain, soreness, stiffness, and joint swelling. It commonly affects the knee and hip but can occur in any joint. The aetiology is thought to be multifactorial including both genetic and environmental factors. Disease prevention through risk reduction is a key management strategy (*Buckwalter et al., 2004; Allen and Golightly, 2015*). Data on the prevalence of knee OA are not available at a small area level in England hence the need for synthetic microdata.

2.2. Data sources

Typically, two data sources are required to build static spatial microsimulation models. The first is a representative sample survey, usually called the population dataset. This provides detailed information on the outcome of interest but lacks geographic information at the small area level. The second, called the geographic dataset, may contain limited data items but is disaggregated at the required spatial level (*Cassells et al., 2013*). In our project, the English Longitudinal Survey of Aging (ELSA) and Health Survey for England (HSE) were our population datasets, and the 2011 census, the geographical dataset.

ELSA has been extensively described elsewhere but in brief, is a multidisciplinary cohort study of a representative sample of the English population aged 50 years and above. ELSA participants are followed up every two years (waves). Each wave involves questionnaires, anthropometric measurements and tissue samples (in every other wave) and consists of about 11,000 participants (*Mindell et al., 2012*). This study utilised ELSA wave 6, collected in 2012 and 2013. Wave 6 was chosen because of its proximity to the 2011 census and because it also included weight and height measurements which are only collected in every second wave. Our study population consisted of core sample members of the ELSA dataset. Individuals less than 50 years and non-core sample members of the ELSA (regardless of age) were excluded from the analysis.

Health Survey for England (HSE) is a series of annual cross sectional surveys about the health of people living in England. The survey started in 1991 and is an authoritative source of health statistics used to plan the nation's health policy. Each year consists of about 10,000 participants of all ages. (*Mindell et al., 2012*) Our study amalgamated HSE survey data from 2012 to 2014 to obtain a larger sample population. This was done to obtain HSE closest to the 2011 national census data and also proximity to ELSA wave 6. Only individuals aged 50 years and above were included in our analysis.

The 2011 United Kingdom (UK) census which is the most recent was used as our geographical dataset. These data were obtained from the office of National Statistics (*Longhurst et al., 2007*). The geographical units of measurement used in this study were Lower Super Output Area (LSOA) [1] and 2011 electoral ward. LSOAs were chosen because relative homogeneity of the population and for ease of geo-referencing, ward microdata was used for sensitivity analysis. According to the 2011 census, there are 32,844 LSOAs and 7,689 electoral wards in England (*Nomis*).

2.3. Definition of outcomes

Knee OA - from ELSA was defined as a combination of self-reported doctor-diagnosed knee OA, knee pain, as well as knee replacement due to Arthritis. This was coded as a binary variable.

Body Mass Index (BMI) – was defined as weight (in kilograms) ÷ height (in meters)², and this was obtained from the HSE using the variable name "BMI valid". This variable was categorised as follows in kg/m² – <19(underweight), 19–25 (normal weight), 25–30 (overweight), >30(obese).

2.4. Constraint selection

The following selection describes data management and constraint selection.

It is very important that constraints are highly correlated with the outcome of interest (*Edwards and Clarke, 2009*). This was a major consideration in our study therefore we conducted an extensive literature search to identify risk factors associated with knee OA. Sociodemographic variables present in all above mentioned data sets were also identified as potential constraint variables. Following

Table 1. Names and categories of covariates used in our analyses

Variable	Categories
Age	50–59 years, 60–69 years, 70–79 years, 80–89 years, ≥90 years
Sex	Male, Female
*Ethnicity	White, Non-White
Health	Good, Fair, Poor
BMI	Underweight, Overweight & Obese
*NSSEC	Higher managerial & professional, Lower-professional, Intermediate Occupations, Lower supervisory & technical, Semi routine, routine & other.
Marital Status	Single, Married, Separated, Divorced, Widowed
Level of Education	NVQ4&5, Higher Education below Degree, NVQ3, NVQ2, NVQ1
Smoking history	Never smoker, Ex-smoker, Current smoker
Alcohol	Never, once or twice a year, Every couple of Months, Once or twice a month, Once or twice a week, Every other day, Everyday
Fruit & Vegetable Consumption/week	Less than 5 portions, Greater than 5 portions

*Categorised to match ELSA.

identification, these variables were re-categorised to ensure consistent definitions across datasets (*Cassells et al., 2013*). The unit of analysis was the individual and as such household level variables and variables that could not be consistently defined/re-categorised were excluded at this stage. Categorization of potential constraint variables are presented in **Table 1** below.

3. ELSA analysis

Univariate and Multivariate logistic regression were used to identify associations between knee osteoarthritis and other covariates. All significant predictors of knee OA identified during the univariate analysis (adjusted Wald's test, $p \leq 0.05$) were included in a multivariate backward elimination regression analysis. Only predictors (except age and sex, which were considered as priori predictors and retained in the model regardless of statistical significance) with a p value of ≤ 0.05 were left in the final model. We reinserted excluded predictors to the final model one at a time to further check whether they became statistically significant.

An optimal number of constraint variables is necessary to obtain accurate synthetic data however more constraint variables are not necessarily associated with accurate simulations (*Tanton and Vidyattama, 2010*). As an additional step, calibration was done by testing different combinations of variables in the final model using the Archer-Lemeshow goodness of fit test (*Archer et al., 2007; Williams, 2015*). Variables in the model with the highest p value were selected as constraint variables.

4. HSE analysis

HSE 2012–2014 were combined to increase the size of the survey population, potentially increasing the pool of individuals to choose from during spatial microsimulation modelling. Multinomial logistic regression, with BMI as the outcome variable was used to identify covariates associated with BMI. Variables found to be significantly associated with BMI using multivariate multinomial regression were retained. We were unable to statistically compare nested models to obtain the most parsimonious model because maximum likelihood assumptions that underpin these tests are violated by complex sampling designs used in HSE (*Hahs-Vaughn et al., 2011; Williams, 2015*). Therefore, we employed the following approach to final constraint selection. All retained covariates were chosen as constraints and used to simulate BMI data in different combinations, dropping and adding each variables in turn except age and sex which were included in all simulations. The resultant synthetic datasets were internally validated against predetermined criteria (described below), discarding configurations that did

not meet our criteria. Covariates in the dataset with the best validation scores was selected as final constraint variables (Smith et al., 2007).

5. Missing data and collinearity

A single stochastic imputation using a chained equation approach based on candidates with complete data (age and sex) was used to replace missing data (Cassells et al., 2013). Data was assumed to be missing at random (Sterne et al., 2009; Marston et al., 2010). Variance Inflation Factor (VIF) test was used to detect multiple collinearity among variables in the regression models. We did not find any value above 10.

5.1. Spatial microsimulation

Our project used SimObesity, a static spatial microsimulation model developed within the School of Geography, University of Leeds. It utilizes a deterministic combinatorial optimisation procedure. Models like SimObesity have been used extensively in healthcare modelling and have been shown to produce robust estimates (Burden and Steel, 2016; Kosar and Tomintz, 2014). In addition, these models are simple and relatively quick to execute. A detailed description of the exact algorithm and applications of SimObesity can be found elsewhere (Edwards and Clarke, 2009; Timmins and Edwards, 2016). Briefly, SimObesity works in two steps. The first step determines the combination of individuals in a geographical area using the reweighting algorithm. The second step involves intergerisation so only 'whole' individuals and not fractions of people are assigned to each geographical area (Ballas et al., 2005). SimObesity algorithm is displayed in the equation below.

For P_{ij}

$$x_{ij} = w_{ij} \times c_{ij}/s_{ij} \tag{1}$$

$$y_{ij} = x_{ij} \times \sum c_j / \sum x_j \tag{2}$$

Where,

P_{ij} represents each person in the population dataset.

x_{ij} is the reweight value for person i in area j .

w_{ij} is the person's original weight in the population table for the first constraint variable and is the resulting weight (y_{ij}) from the previous constraint for all subsequent constraint variables.

c_{ij} is element ij of the corresponding constraint table.

s_{ij} is element ij of the corresponding summary table.

$\sum c_j$ is the sum of the relevant area column for the constraint variable.

$\sum x_j$ is the sum of the relevant area column for the reweight value calculated in the previous step.

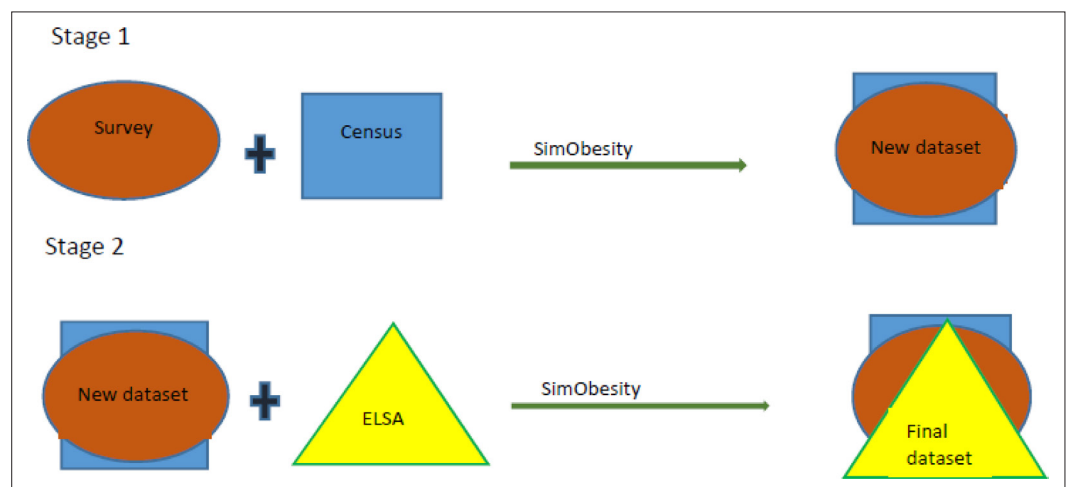


Figure 1. 2-stage Spatial Microsimulation

In our study, spatial microsimulation was conducted in a two-step sequence. First, constraints selected from the HSE analysis (described above) were combined with the 2011 Census (geographical dataset) using SimObesity to create a dataset that we called the hybrid dataset. This hybrid dataset contained information on both sociodemographic variables and BMI. Step 2 of the sequence involved combining the hybrid dataset with ELSA, also using SimObesity (based on constraint variables identified during the ELSA regression analysis) to obtain the final dataset. This dataset contained both knee OA data as well as BMI data. This processes illustrated in **Figure 1** below. Initial (default) weights were set at '1' for all simulations (**Smith et al., 2009**).

6. Model validation/calibration

This is an integral part of spatial microsimulation and there are broadly two approaches - internal and external. Internal validation examines how well the simulated data represents source data (constraint variables), while external validation tries to establish how closely the simulated estimates represent the actual spatial distribution of the outcome of interest. This is very challenging (hence the need for spatial microsimulation in the first place) (**Edwards and Tanton, 2012**).

Although several statistical techniques exist to internally validate spatial models, there is no consensus on which method is most appropriate. A review of the pros and cons of common internal validation techniques can be found in published literature (**Timmins and Edwards, 2016; Rahman et al., 2013**). In general, multiple techniques are used for internal validation. Our approach to validation was based on the simplicity of interpretation, preferring methods with an objective statistical measure over methods with a subjective interpretation.

Our simulated datasets were validated against the original census data ie, the hybrid and final dataset were each validated against the census data. This way we could access how close or removed our final dataset was from the census data. This was done with the following measures: R^2 , Standard Absolute Error (SAE), and Standard Error about Identity (SEI) (**Tanton and Vidyattama, 2010**). Scatterplots were also used to visualize the relationship between simulated and census data. R^2 and SEI values ≥ 0.98 , and SAE ≤ 0.05 were considered to be a good fit to generate the hybrid data, and R^2 and SEI values ≥ 0.94 and SAE values ≤ 0.10 to generate final knee OA dataset (**Smith et al., 2009**). R^2 and SEI cut off points were chosen pragmatically to account for a slight decay in the values at the second stage of simulation which may be caused by the intergerisation phase of SimObesity.

Prior to the validation methods mentioned above, we broadly compared constraint variables across all datasets i.e., we compared proportions of subgroups within all constraint variables in the hybrid dataset with those of the census and final datasets respectively. In addition, we aggregated the estimated data to a higher spatial scale, Government Office Region (GOR) and compared the proportion of people with knee osteoarthritis in ELSA (our final population dataset) with that of our synthetic dataset.

Data cleaning, constraint selection and model validation were conducted using STATA 14 SE (**Statacorp, 2015**) and Spatial microsimulation was done using SimObesity. Ethical approval was obtained from the University of Nottingham Ethics Committee SDA23062015.

7. Results

There were a total of 9,169 core ELSA sample members, 12,521 observations from a combination of HSE 2012–2014, and about 18 million individuals in the 2011 census aged 50 years and above.

Table 2 below shows that the distribution of variables in all the 3 datasets is broadly similar with only slight differences in the frequency distribution of 'separated' individuals in ELSA compared with the Census and HSE. The frequency distribution of Education qualification also differs slightly between census, and HSE and ELSA datasets. However overall, HSE and ELSA are broadly representative of the census data making them ideal for spatial microsimulation.

Tables 3 and 4 below display the output of ELSA univariate and multivariate logistic regression analysis and GOF test respectively. Age, Sex, NSSEC, Health status and BMI were found to be significantly associated with Knee OA in the Univariate analysis.

Although Age was not significantly associated with Knee OA in multivariate analysis, it was considered to be an important predictor and thus included in the final model. **Table 3** shows the results of the archer-Lemeshow Goodness of Fit (GOF) statistic. Comparing different combinations of variables in the final model – Age, Sex, NSSEC and BMI were shown to be the best predictors of knee OA,

Table 2. Frequency distribution of variables in geographical (census) and survey (HSE and ELSA) datasets

Variables/categories	ELSA, n (%)	HSE, n (%)	Census LSOA, n (%)	
Age	50–59 years	2,062 (22.49)	4,901 (32.67)	6,397,087 (35.09)
	60–69 years	3,451 (37.64)	4,018 (32.09)	5,680,431 (31.16)
	70–79 years	2,444 (26.66)	2,837(22.66)	3,713,474 (20.37)
	80–89 years	1,005 (10.96)	1,383 (11.05)	2,032,084 (11.16)
	≥90 years	207 (2.26)	192 (1.53)	403,817 (2.22)
Sex	Female	5,096 (55.58)	6,756 (53.96)	9,658,440 (52.98)
	Male	4,073 (44.42)	5,765 (46.04)	8,571,453 (47.02)
*Marital	Single, Cohabitees	603 (6.58)	1,425 (11.39)	1,464,682 (8.03)
	Married, Civil partnerships	5,872 (64.06)	7,726 (61.73)	11,167,413 (61.26)
	Separated	25 (0.27)	271 (2.17)	420,140 (2.31)
	Divorced	1,070 (11.67)	1,268 (10.13)	2,306,990 (12.65)
	Widowed	1,597 (17.42)	1,826 (14.59)	2,870,668 (15.75)
*Ethnicity	White	8,845 (96.49)	11,792 (94.54)	17,027,207 (93.40)
	Non-white	322 (3.51)	681 (5.46)	1,202,686 (6.60)
*NSSEC	Higher Managerial & Professional	3,167 (34.74)	3,982 (32.36)	5,475,646 (30.04)
	Lower Managerial	1,249 (13.70)	1,922 (15.57)	2,624,312 (14.40)
	Intermediate Occupations	1,059 (11.62)	1,298 (10.52)	2,080,200 (11.41)
	Lower supervisory & Technical	846 (9.28)	937 (7.59)	1,457,493(7.99)
	Semi routine, routine & others	2,794 (30.65)	4,204 (34.06)	6,592,242(36.16)
	NVQ 4 & 5	1,551 (17.06)	2,191 (17.50)	6,965,766 (38.21)
	Higher Education, below Degree	1,268 (13.94)	1,459 (11.65)	1,773,428 (9.73)
*Education	NVQ3	752 (8.27)	1,173 (9.37)	1,930,079 (10.59)
	NVQ2	1,703 (18.73)	2,336 (18.66)	991,560 (5.44)
	NVQ1	366 (4.03)	514 (4.11)	1,254,681 (6.88)
	Foreign/Other	1,043 (11.47)	353 (2.82)	4,180,862 (22.93)
	No Qualification	2,410 (26.50)	4,444 (35.49)	1,133,517 (6.22)
	Excellent, good	6,330 (72.10)	8,010 (63.97)	11,213,005 (61.51)
	Fair	1,729 (19.69)	3,079 (24.59)	4,856,203 (26.64)
Health	Poor	720 (8.20)	1,425 (11.38)	2,160,685 (11.85)
	Never smoker	3,275 (36.73)	5,777 (46.14)	–
*Smoke	Ex-smoker	4,790 (52.25)	4,910 (39.21)	–
	Current Smoker	1,102 (12.02)	1,802 (14.39)	–
	Underweight	69 (0.93)	85 (0.82)	–
*BMI	Normal	1,958 (26.52)	2,761 (26.58)	–
	Overweight	3,029 (41.03)	4,380 (42.16)	–
	Obese	2,327 (31.52)	3,163 (30.45)	–
Knee Osteoarthritis	Yes	1,728 (18.85)	–	–
	No	7,441 (81.15)	–	–
TOTAL	9,169 (100)	12,521 (100)	18,229,893 (100)	

*Variables with missing data in HSE & ELSA.

Table 3. ELSA Univariate and Multivariate Logistic Regression Analysis – F test and P values

Variables	Univariate		Final model	
	F - test	P value	F -test	P value
*Age	3.01	0.0172	0.43	0.9038
Sex	22.74	<0.0001	20.67	<0.0001
NSSEC	19.54	<0.0001	6.15	0.0001
Health	174.26	<0.0001	54.06	<0.0001
BMI	47.93	<0.0001	24.81	<0.0001
[†] Illness	234	<0.0001	76.7	<0.0001
Marital	1.98	0.0953		
Ethnicity	1.85	0.1741		
Education	5,847	<0.001		
Smoke	5.85	0.0029		
Alcohol	14.14	<0.0001		
Fruit & Vegetable consumption	3.94	0.0196		

*Not significant but included in the final model.

[†]Found to be defined differently in the census and was subsequently dropped from analysis.

Sex, NSSEC, Health, Marital status, Ethnicity, Highest Qualification, and Smoking status were identified as constraint variables. The census does not collect smoking status data, so smoking was not used as a constraint variable.

8. Validation - HSE Vs Census

Age, Sex, NSSEC, Health, Marital Status, Ethnicity, and Education were used in different combinations as constraint variables to predict BMI. The Age, Sex and NSSEC simulation model provided the best fitted dataset based on our validation criteria and is the only result presented here. See following tables.

Table 4. ELSA Goodness of fit test

Model parameters	Archer-Lemeshow GOF values
Age, Sex, NSSEC, Illness, Health, BMI	0.3374
Age, Sex, NSSEC, Health, BMI	0.4861
Age, Sex, Health, BMI	0.1147
Age, Sex, BMI	0.6283
Age, Sex, NSSEC	0.6708
Age, Sex, Health	0.7564
Age, NSSEC	0.3453
Age, Sex, NSSEC, BMI	0.9742
Age, Sex	0.5680

Table 5. HSE Univariate and Multivariate Multinomial Regression Analysis – F test and P values

Variables	Univariate		Final model	
	F - test	P value	F -test	P value
Age	2.88	<0.0001	8.87	<0.0001
Sex	41.49	<0.0001	35.73	<0.0001
NSSEC	7.22	<0.0001	1.98	0.0139
Health	35.24	<0.0001	28.75	<0.0001
Marital	13.87	<0.0001	28.89	<0.0001
Ethnicity	3046.85	<0.0001	1236.81	<0.0001
Education	5.7	<0.0001	2.29	0.0016
Smoke	12.54	<0.0001	14.63	<0.0001
Fruit & Vegetable consumption	1.06	0.3843		

with the highest GOF p value of 0.9742. This is consistent with current available literature on the predictors of knee OA (*Hart et al., 1999; Reyes et al., 2015; Blagojevic et al., 2010*) therefore these variables were selected as constraint variables for knee OA spatial simulation modelling.

Table 5 above shows the results of multivariate multinomial regression analysis of HSE data. Age,

From **Table 6** above, we can see that both the hybrid and final datasets have similar proportions of variables as the original census and HSE, and ELSA datasets (for BMI and Knee OA respectively)

Table 7 above displays internal validation measures at both simulation stages. It can be seen that in general, there was a slight reduction in SEI and R² values, and an increase in SAE values from the hybrid (BMI simulations) to the final dataset (OA simulations). All BMI simulations met our validation criteria of an SAE of ≤0.05 and R² value of ≥0.98, however age-group ≥90 years and males just failed to meet our set SEI criteria of ≥0.98. In the final simulations, age group ≥90 years also failed to meet our SAE and SEI criteria; and all sex sub-categories did not meet our set criteria as well. However, it can be observed that most

Table 6. Frequency distribution of variables in Original and synthetic datasets

Variables/Categories		Census (%)	Hybrid, n (%)	Final, n (%)	ELSA, n (%)	HSE, n (%)
Age	50–59 years	35.09	35.02	34.59	22.49	32.67
	60–69 years	31.16	31.16	31.19	37.64	32.09
	70–79 years	20.37	20.48	20.75	26.66	22.66
	80–89 years	11.16	11.15	11.24	10.96	11.05
	≥90 years	2.22	2.19	2.21	2.26	1.53
Sex	Female	52.98	52.84	54.07	55.58	53.96
	Male	47.02	47.16	45.92	44.42	46.04
NSSEC	Higher Managerial & Professional	30.04	30.16	31.39	34.74	32.36
	Lower Managerial	14.4	14.29	13.84	13.70	15.57
	Intermediate Occupations	11.41	11.49	11.39	11.62	10.52
	Lower supervisory & Technical	7.99	8.08	8.17	9.28	7.59
	Semi routine, routine & others	36.16	35.98	35.20	30.65	34.06
BMI	<18.5 (Underweight)	–	0.98	0.98	0.93	0.82
	18.5–<25 (Normal)	–	26.20	26.39	26.52	26.58
	≥24–<30 (Overweight)	–	42.20	41.51	41.03	42.16
BMI	≥30 (Obese)	–	30.62	31.10	31.52	30.45
Knee Osteoarthritis	Yes	–	–	18.69	18.85	–
	No	–	–	81.31	81.15	–

sub categories of constraint variables met our validation criteria at both stages of the microsimulation sequence.

The results of the scatter plots above show a good correlation between the original census dataset and the final knee OA simulated dataset howbeit with a lower R^2 value for all constraint variable than obtained when the census was compared with the hybrid dataset (**Figures 2 and 3**). This effect is more pronounced in the sex category and in the age group 90 years and above. However the scatter plots

Table 7. Validation parameters (SAE, R^2 and SEI) for BMI and OA spatial microsimulations

Variables/Categories		BMI simulation			OA Simulation		
		SAE	R^2	SEI	SAE	R^2	SEI
Age	50–59 years	0.0186	0.9933	0.9918	0.0433	0.9681	0.9569
	60–69 years	0.0132	0.9983	0.9980	0.0200	0.9959	0.9951
	70–79 years	0.0295	0.9933	0.9931	0.0560	0.9768	0.9753
	80–89 years	0.0304	0.9949	0.9949	0.0617	0.9807	0.9800
	≥90 years	0.0916	0.9801	0.9797	0.1741	0.9416	0.9289
Sex	Male	0.0344	0.9847	0.9792	0.0680	0.9370	0.9327
	Female	0.0304	0.9851	0.9847	0.0777	0.9367	0.9064
NSSEC	Higher Managerial & Professional	0.0199	0.9978	0.9977	0.0583	0.9908	0.9800
	Lower Managerial	0.0374	0.9888	0.9872	0.0818	0.9550	0.9450
	Intermediate Occupations	0.0300	0.9956	0.9954	0.0542	0.9867	0.9850
	Lower supervisory & Technical	0.0271	0.9940	0.9929	0.0556	0.9788	0.9710
	Semi routine, routine & others	0.0124	0.9987	0.9984	0.0353	0.9938	0.9881

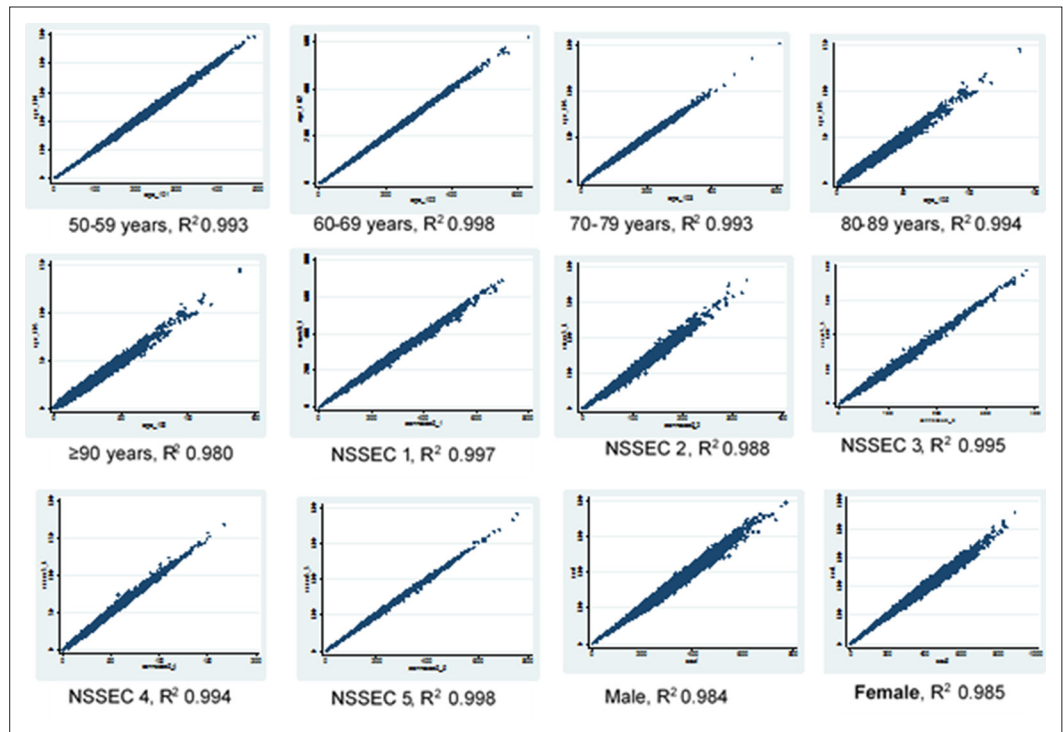


Figure 2. Scatter plots for age, sex and NSSEC categories, simulated counts from the hybrid dataset versus census totals

show similar patterns across the different variables in both datasets which shows that the census data is correlated with the final OA simulation data.

Figure 4 above compares the proportion of individuals with knee osteoarthritis from ELSA with that generated through spatial microsimulation. The error bars represent 95% confidence intervals. It can be seen here that synthetic data point estimates of the prevalence of knee osteoarthritis are

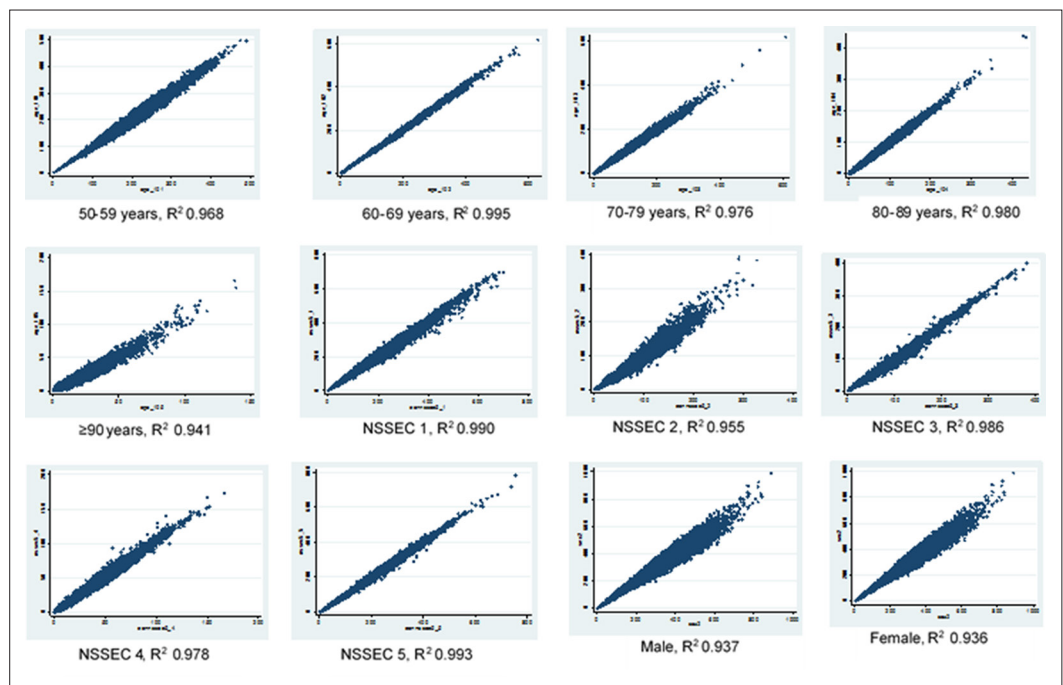


Figure 3. Scatter plots for age, sex and NSSEC categories, simulated counts from the final dataset versus census

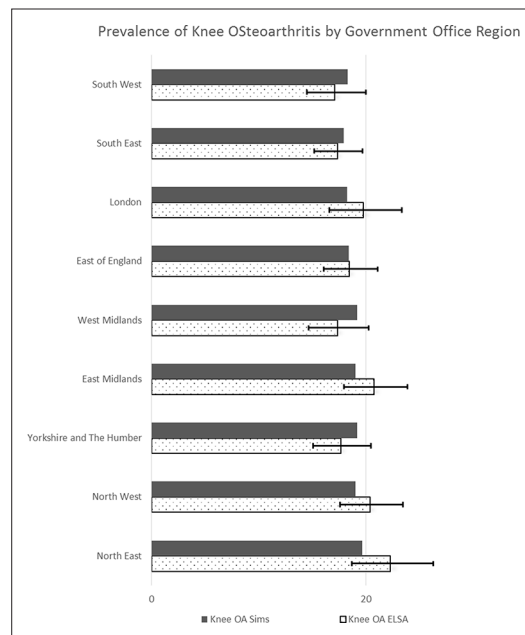


Figure 4. Comparing regional aggregates of our simulated synthetic microdata data with ELSA data (grouped by regions)

within 95% confidence limits across all GORs. The highest difference of 2.7% points was seen in the North-East of England, with almost a perfect match in the East of England. ELSA proportions also appear to show more variability than the simulated data.

9. Discussion

The case study described above provided a detailed account of constraint selection and internal validation of a novel approach to spatial microsimulation. It demonstrated our approach to overcoming a key challenge encountered when generating small area level microdata. To this end, BMI, a variable found to be strongly associated with knee OA (from a literature search and our regression analyses) (Blagojevic *et al.*, 2010) but not collected by our geographical dataset was successfully incorporated into the simulation by conducting simulations in 2 stages. This served a dual purpose because combining data from two survey datasets inadvertently increased the pool of individuals to choose from, as well as providing data on hitherto unavailable information. Although there seemed to be some 'decay' in validation scores in the final output, as shown

by the slight increase in SAE and reduction in the R^2 values in the final dataset, the association between constraint variables in the census and final datasets is still very strong.

In addition, the proportions of individuals in various categories (of both constraint and outcome variables) were very similar across both survey datasets, hybrid and final datasets. It is also interesting to note that categories with only few observations (e.g. individuals aged 90 years and above) showed a poorer fit than categories with a large number of observations (50–59 years) in both the hybrid and final simulations. This also attests to the strong correlation between the original census and final datasets. Furthermore, the prevalence of our outcome of interest derived from our synthetic data approximated that derived from ELSA, with the largest point estimate difference of less than three percentage points. Another strength of this approach is that the similarity of source datasets such that individuals in the HSE and ELSA are alike in terms of variable categorization and sampling framework probably making it easier to clone individuals based on selected attributes however HSE contained slightly higher absolute numbers of participants in each category.

Bearing this in mind, it is also important to highlight some important considerations – first, the case study described above was conducted using a static deterministic reweighing and combinatorial spatial microsimulation method. It may not be applicable to other reweighing spatial simulation methods e.g. methods utilizing iterative proportional fitting, and other dynamic spatial models.

Secondly, choice of constraint is a very important determinant of the simulated output. Any other selection of constraints would produce a different output however our choice of constraint variables were based on correlates of the outcome of interest using robust calibration techniques. Our final choice of simulation output was based on predetermined internal validation criteria. In addition, some variables such as presence of long term illness and smoking history which were significantly associated with BMI in our regression analysis were not used as constraint variables to obtain the hybrid dataset. This is because the UK census does not collect data on smoking prevalence and also a systematic review of published literature has shown an inconclusive relationship between smoking and Osteoarthritis, our main outcome of interest (Hui *et al.*, 2011; Hart *et al.*, 1999). The presence of long term illness was not defined consistently across datasets and therefore could not be used in the simulations. We are not certain of the effect of these exclusions on the final output.

Another important consideration is the effect of the single imputation of missing BMI data in both survey datasets (HSE and ELSA) prior to simulation. This had the advantage of providing a larger pool of individuals for the simulations, but it is not quite clear what other biases this could have been introduced into the model. However, simulation models we conducted using complete-case analysis did not have better validation scores.

It is arguable that this two stage simulation could be conducted using ELSA data only, given that ELSA and HSE contain similar variables, but interestingly our two-stage simulations using only ELSA did not yield consistent validation results (please see supplementary material). The slight decay in some validation criteria noticed with HSE/ELSA Knee Osteoarthritis simulations was amplified with ELSA/ELSA simulations. We believe this may be due to error augmentation, which is the additive effect of errors generated in the first stage and those generated during the second stage of simulations. This concept was not explored further in this paper as the introduction of HSE provided consistent validation results both at ward and LSOA levels.

Furthermore, we need to consider the effect of time on this two stage simulation approach. All input data were collected at slightly different timeframes with the census data collected in 2011, HSE 2012–2014 and ELSA from 2012 to 2013. We expect that the final simulation results represent LSOA prevalence of knee OA in 2011 and considering Knee OA is a chronic condition and the UK population growth rate is fairly stable therefore we do not expect the time overlap to introduce major biases to our estimates. Also the prevalence of sociodemographic variables was similar across all included datasets.

Similarly, though we have successfully incorporated a key variable into our simulation models using our two-staged approach, knee OA is affected by a host of other factors that are not present in all of our input datasets (e.g. joint injury) (Coggon *et al.*, 2000). Inclusion of these variables may affect the simulation output. Further work needs to be done to explore other ways to include more variables into these spatial models however the trade-off between number of constraints and internal validity simulation outputs is worth considering.

Finally, this paper did not discuss the external validity of our simulation estimates. We recognise that external validation of this model would pose additional challenges due to the difficulty encountered in the definition of knee osteoarthritis.

10. Conclusion

To our knowledge, this is a first attempt using a two-staged approach to conduct static deterministic combinatorial spatial microsimulation. The overall objective was to incorporate BMI, which is an important determinant of knee osteoarthritis into the spatial model in order to produce more robust estimates. This was achieved successfully and consistently with good validation results.

Ethical approval

This study was approved by the University of Nottingham Ethics Committee. Ethics Reference Number SDA23062015.

ORCID iD

Onosi Sylvia Ifesemen  <https://orcid.org/0000-0002-4234-7905>

Acknowledgements

The Authors would like to thank Dr Sudhir Venkatesan, Dr Boliang Guo and Dr Gwen Sasha Fernandez for contribution to various aspects of this study, the peer reviewers for their constructive feedback and finally, our funders, Versus Arthritis.

Funding

This work was supported by Versus Arthritis (formerly known as Arthritis Research UK) Centre for Sports Exercise and Osteoarthritis Research, grant number 20194.

Data availability

2011 census data are publicly available via Nomis, provided by the Office for National Statistics (ONS) and can be assessed at <https://www.nomisweb.co.uk/>. Health Survey for England 2012 – 2014 were

provided by NatCen Social Research, sponsored by the Information Centre for Health and Social care. Data are available courtesy the UK Data Service for scientific research upon registration. The English Longitudinal Study of Ageing was developed by a team of researchers based at the University College London, NatCen Social Research, and the Institute for Fiscal Studies. The data were collected by NatCen Social Research. The funding is currently provided by the National Institute of Aging (R01AG017644), and a consortium of UK government departments coordinated by the National Institute for Health Research. Data are available courtesy the UK Data Service for scientific research upon registration.

Code availability

SimObesity was developed within the School of Geography, University of Leeds.

Competing interests

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript:

Onosi Sylvia Ifesemen

Thomas Bestwick –Stevenson

Kimberley L Edwards

References

- Allen KD, Golightly YM. 2015. Epidemiology of osteoarthritis: state of the evidence. *Current Opinion in Rheumatology* **27**:276–283.
- Archer KJ, Lemeshow S, Hosmer DW. 2007. Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis* **51**:4450–4464. DOI: <https://doi.org/10.1016/j.csda.2006.07.006>
- Ballas D, Clarke GP, Wiemers E. 2006. Spatial microsimulation for rural policy analysis in Ireland: the implications of CAP reforms for the National spatial strategy. *Journal of Rural Studies* **22**:367–378. DOI: <https://doi.org/10.1016/j.jrurstud.2006.01.002>
- Birkin M, Clarke M. 2011. Spatial Microsimulation Models: A Review and a Glimpse into the Future. In: Stillwell, J, Clarke, M (editors). *Population Dynamics and Projection Methods*. Dordrecht: Springer Netherlands.
- Ballas D, Rossiter D, Thomas B, Clarke G, Dorling D. 2005. *Geography matters: Simulating the local impacts of national social policies*. York: University of Leeds Joseph Rowntree Foundation.
- Blagojevic M, Jinks C, Jeffery A, Jordan KP. 2010. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. *Osteoarthritis and Cartilage* **18**:24–33. DOI: <https://doi.org/10.1016/j.joca.2009.08.010>
- Buckwalter JA, Saltzman C, Brown T. 2004. The impact of osteoarthritis: implications for research. *Clinical Orthopaedics and Related Research* **427**.
- Burden S, Steel D. 2016. Constraint choice for spatial Microsimulation. *Population, Space and Place* **22**:568–583. DOI: <https://doi.org/10.1002/psp.1942>
- Cassells R, Miranti R, Harding A. 2013. Building a Static Spatial Microsimulation Model: Data Preparation. In: Tanton, R, Edwards, K (editors). *Spatial Microsimulation: A Reference Guide for Users*. Dordrecht. Springer Netherlands:
- Cataife G. 2014. Small area estimation of obesity prevalence and dietary patterns: a model applied to Rio de Janeiro city, Brazil. *Health & Place* **26**:47–52. DOI: <https://doi.org/10.1016/j.healthplace.2013.12.004>
- Coggon D, Croft P, Kellingray S, Barrett D, McLaren M, Cooper C. 2000. Occupational physical activities and osteoarthritis of the knee. *Arthritis & Rheumatism* **43**:1443–1449. DOI: [https://doi.org/10.1002/1529-0131\(200007\)43:7<1443::AID-ANR5>3.0.CO;2-1](https://doi.org/10.1002/1529-0131(200007)43:7<1443::AID-ANR5>3.0.CO;2-1)
- Doll R, Hill AB. 1950. Smoking and carcinoma of the lung. *BMJ* **2**:739–748. DOI: <https://doi.org/10.1136/bmj.2.4682.739>
- Doll R, Hill AB. 1956. Lung cancer and other causes of death in relation to smoking. *BMJ* **2**:1071–1081. DOI: <https://doi.org/10.1136/bmj.2.5001.1071>
- Edwards KL, Clarke GP. 2009. The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity. *Social Science & Medicine* **69**:1127–1134. DOI: <https://doi.org/10.1016/j.socscimed.2009.07.037>
- Edwards KL, Tanton R. 2012. Validation of spatial microsimulation models. In: *Spatial microsimulation: A reference guide for users*. Springer.

- Habbema JDF**, Van Oortmarssen OJ, De Vlas SJ, Plaisier AP. 1996. The microsimulation approach to epidemiologic modeling of helminthic infections, with special reference to schistosomiasis. *The American Journal of Tropical Medicine and Hygiene* **55**:165–169. DOI: <https://doi.org/10.4269/ajtmh.1996.55.165>
- Hahs-Vaughn DL**, Mcwayne CM, Bulotsky-Shearer RJ, Wen X, Faria A-M. 2011. Methodological considerations in using complex survey data: an applied example with the head start family and child experiences survey. *Evaluation Review* **35**:269–303.
- Harland K**, Heppenstall A, Smith D, Birkin M. 2012. Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation* **15**:1. DOI: <https://doi.org/10.18564/jasss.1909>
- Hart DJ**, Doyle DV, Spector TD, David D. 1999. Incidence and risk factors for radiographic knee osteoarthritis in middle-aged women: the Chingford study. *Arthritis & Rheumatism* **42**:17–24. DOI: [https://doi.org/10.1002/1529-0131\(199901\)42:1<17::AID-ANR2>3.0.CO;2-E](https://doi.org/10.1002/1529-0131(199901)42:1<17::AID-ANR2>3.0.CO;2-E)
- Hui M**, Doherty M, Zhang W. 2011. Does smoking protect against osteoarthritis? meta-analysis of observational studies. *Annals of the Rheumatic Diseases* **70**:1231–1237. DOI: <https://doi.org/10.1136/ard.2010.142323>
- Kosar B**, Tomintz M. 2014. simSalud: a web-based spatial microsimulation to model the health status for small areas using the example of smokers in Austria..
- Lee PN**, Forey BA, Coombs KJ. 2012. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer* **12**:385. DOI: <https://doi.org/10.1186/1471-2407-12-385>
- Longhurst J**, Tromans N, Young C, Miller C. 2007. Statistical disclosure control for the 2011 UK census. Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester;17–19.
- Lovelace R**, Ballas D, Watson M. 2014. A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. *Journal of Transport Geography* **34**:282–296. DOI: <https://doi.org/10.1016/j.jtrangeo.2013.07.008>
- Lymer S**, Brown L, Yap M, Harding A. 2008. 2001 regional disability estimates for new South Wales, Australia, using spatial Microsimulation. *Applied Spatial Analysis and Policy* **1**:99–116. DOI: <https://doi.org/10.1007/s12061-008-9006-4>
- Marston L**, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. 2010. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiology and Drug Safety* **19**:618–626. DOI: <https://doi.org/10.1002/pds.1934>
- Mindell J**, Biddulph JP, Hirani V, Stamatakis E, Craig R, Nunn S, Shelton N. 2012. Cohort profile: the health survey for England. *International Journal of Epidemiology* **41**:1585–1593. DOI: <https://doi.org/10.1093/ije/dyr199>
- Morris M**, Kretzschmar M. 2000. A microsimulation study of the effect of concurrent partnerships on the spread of HIV in Uganda. *Mathematical Population Studies* **8**:109–133. DOI: <https://doi.org/10.1080/08898480009525478>
- Nomis**. Official Labour Market Statistics [Online]. <https://www.nomisweb.co.uk/> [Accessed 3rd July 2018].
- O'donoghue C**, Morrissey K, Lennon J. 2014. Spatial microsimulation modelling: a review of applications and methodological choices.
- Prakash MK**, Lang B, Heinrich H, Valli PV, Bauerfeind P, Sonnenberg A, Beerenwinkel N, Misselwitz B. 2017. CMOST: an open-source framework for the microsimulation of colorectal cancer screening strategies. *BMC Medical Informatics and Decision Making* **17**:80. DOI: <https://doi.org/10.1186/s12911-017-0458-9>
- Rahman A**, Harding A, Tanton R, Liu S. 2010. Methodological issues in spatial Microsimulation modelling for small area estimation. *International Journal of Microsimulation* **3**:3–22.
- Rahman A**, Harding A, Tanton R, Liu S. 2013. Simulating the characteristics of populations at the small area level: new validation techniques for a spatial microsimulation model in Australia. *Computational Statistics & Data Analysis* **57**:149–165. DOI: <https://doi.org/10.1016/j.csda.2012.06.018>
- Rephann TJ**, Holm E. 2004. Economic-demographic effects of immigration: results from a dynamic spatial microsimulation model. *International Regional Science Review* **27**:379–410. DOI: <https://doi.org/10.1177/0160017604267628>
- Reyes C**, Garcia-Gil M, Elorza JM, Mendez-Boo L, Hermosilla E, Javaid MK, Cooper C, Diez-Perez A, Arden NK, Bolibar B, Ramos R, Prieto-Alhambra D. 2015. Socio-economic status and the risk of developing hand, hip or knee osteoarthritis: a region-wide ecological study. *Osteoarthritis and Cartilage* **23**:1323–1329. DOI: <https://doi.org/10.1016/j.joca.2015.03.020>
- Smith D**, Clarke G, Harland K. 2007. SimHealth: estimating small area populations using deterministic spatial Microsimulation in Leeds and Bradford.
- Smith DM**, Clarke GP, Harland K. 2009. Improving the synthetic data generation process in spatial Microsimulation models. *Environment and Planning A: Economy and Space* **41**:1251–1268. DOI: <https://doi.org/10.1068/a4147>
- Smith DM**, Pearce JR, Harland K. 2011. Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? an example of smoking prevalence in New Zealand. *Health & Place* **17**:618–624. DOI: <https://doi.org/10.1016/j.healthplace.2011.01.001>
- Statacorp L**. 2015. Stata statistical software: release 14. [computer program]. *StataCorp LP*.
- Sterne JAC**, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**:b2393. DOI: <https://doi.org/10.1136/bmj.b2393>
- Tanton R**, Edwards KL. 2012. Limits of Static Spatial Microsimulation Models. In: *Spatial Microsimulation: A Reference Guide for Users*. Springer.
- Tanton R**. 2014. A review of spatial microsimulation methods. *International Journal of Microsimulation* **7**:4–25.

- Tanton R**, Vidyattama Y. 2010. Pushing it to the edge: extending generalised regression as a spatial microsimulation method. *International Journal of Microsimulation* **3**:23–33.
- Timmins KA**, Edwards KL. 2016. Validation of spatial microsimulation models: a proposal to adopt the Bland-Altman method. *International Journal of Microsimulation* **9**:106–122.
- van Leeuwen E**, Dekkers J. 2013. Determinants of off-farm income and its local patterns: a spatial microsimulation of Dutch farmers. *Journal of Rural Studies* **31**:55–66. DOI: <https://doi.org/10.1016/j.jrurstud.2013.02.002>
- Williams R**. 2015. Analyzing Complex Survey Data: Some key issues to be aware of [Online]. <https://www3.nd.edu/~rwilliam/stats2/SvyCautions.pdf> [Accessed 22 June 2018].